What you should do (and what you really do) when you don't know what to do: Information seeking during value based decision making

by

Jeffrey Cockburn

B.Sc., University of Victoria; Victoria, BC, 2004

M.Sc., University of Victoria; Victoria, BC, 2009

M.Sc., Brown University; Providence, RI, 2010

A dissertation submitted in partial fulfillment of the

requirements for the degree of Doctor of Philosophy

in Cognitive, Linguistic, and Psychological Sciences at Brown University

PROVIDENCE, RHODE ISLAND

Jan 2015

This dissertation by Jeffrey Cockburn is accepted in its present form by Cognitive, Linguistic, and Psychological Sciences as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date_____ _____

Michael J. Frank, Ph.D., Advisor

Recommended to the Graduate Council

Date_____ _____

David Badre, Ph.D., Reader

Date_____ _____

Michael Littman, Ph.D., Reader

Date_____ _____

Fiery Cushman, Ph.D., Reader

Approved by the Graduate Council

Date_____ _____

Peter M. Weber, Dean of the Graduate School

Vitae

Education

- Doctoral Candidate - Cognitive, Linguistic and Psychological Sciences
  Brown University | Providence R.I., U.S.A
- Master of Science - Cognitive Science, 2010
  Brown University | Providence R.I., U.S.A
- Master of Science - Cognitive Science and Computer Science, 2009
  University of Victoria | Victoria B.C., Canada
- Bachelor of Science - Computer Science, 2004
  University of Victoria | Victoria B.C., Canada

Publications

- Cockburn, J., Collins, A. G. E., & Frank, M. J. (2014). A reinforcement learning mechanism responsible for the valuation of free choice. Neuron, 83(3), 5517.
- Doll, B. B., Waltz, J. A., Cockburn, J., Brown, J. K., Frank, M. J., & Gold, J. M. (2014). Reduced susceptibility to confirmation bias in schizophrenia.Cognitive, Affective, & Behavioral Neuroscience, 1-14.
- Tanaka, J. W., Wolf, J. M., Klaiman, C., Koenig, K., Cockburn, J., Herlihy, L., et al. (2012) The perception and identification of facial emotions

in individuals with Autism Spectrum Disorders using the Let's Face It! Emotion Skills Battery. Journal of Child Psychology and Psychiatry, 53, 1259-1267.

- Cockburn, J., & Frank, M. J. (2011). Reinforcement learning, conflict monitoring, and cognitive control: An integrative model of cingulate-striatal interactions and the ERN. In R. B. Mars, J. Sallet, M. F. S. Rushworth, & N. Yeung (Eds.), Neural basis of motivational and cognitive control (Vol. 9, pp. 311-331). MIT Press.
- Cockburn, J., & Holroyd, C. (2010). Focus on the positive: Computational simulations implicate asymmetrical reward prediction error signals in childhood Attention-Deficit/Hyperactivity Disorder. Brain research, 1365, 18-34.
- Tanaka, J. W., Wolf, J. M., Klaiman, C., Koenig, K., Cockburn, J., Herlihy, L., Brown, C., et al. (2010). Using Computerized Games to Teach Face Recognition Skills to Children with Autism Spectrum Disorder: The" Let. Journal of Child Psychology and Psychiatry, 51(8), 944-952.
- Wolf, J. M., Tanaka, J. W., Klaiman, C., Cockburn, J., Herlihy, L., Brown, C., South, M., et al. (2008). Specific impairment of face-processing abilities in children with autism spectrum disorder using the Lets Face It! skills battery. Autism Research, 1(6), 329-340.
- Cockburn, J., Bartlett, M., Tanaka, J., Movellan, J., Pierce, M., & Schultz, R. (2008). SmileMaze: A Tutoring System in Real-Time Facial Expression Perception and Production in Children with Autism Spectrum Disorder. Proceedings from the IEEE International Conference on Automatic Face & Gesture Recognition, 978 (Vol. 986).

- Jahnke, J. H., McNair, A., Cockburn, J., Souza, P., Furber, R.A., & Lavender, M. (2005). Component-based engineering of distributed embedded control software. Component-Based Software Development for Embedded Systems (pp. 296-319). Springer.

Acknowledgements

To all the people that helped to create this, thank you. I am deeply thankfu and will always be indebted to my advisor Michael J. Frank. I came to work with Michael feeling fairly certain I would benefit from the experience as a scientist and as a person. My expectations were doubled. His mentorship, both in terms of how to think about science and how it should be done will benefit me for the rest of my career. I am also thankful for the helpful guidance of my committee members, David Badre, Fiery Cushman, and Michael Littman. I would also like to thank all the members of the Laboratory for Neural Computation and Cognition, both past and present. You helped to provide an amazing environment where ideas were encouraged, challenged, and pursued with enthusiasm. Finally, I would like to thank my family. Without your support I never would have entertained the thought of pursuing a doctoral degree, but I'm happy to know that none of you will ever feel comfortable referring to me as 'Doctor'.

Abstract of " What you should do (and what you really do) when you don't know what to do: Information seeking during value based decision making " by Jeffrey Cockburn, Ph.D., Brown University, Jan 2015

Learning and action selection have been well studied within the confines of a stable environment. These efforts have led to significant advances in our understanding of these processes from behavioural, cognitive and neuroscientific perspectives. Yet, the world is volatile, and we have only recently begun to probe how instability influences learning and action selection processes. Here, I outline an examination of theoretical and empirical findings related to learning and action selection in volatile settings, focusing primarily on the role of uncertainty; that is, when the state of the environment is not fully known. I propose a heuristic approximation of the optimal solution to action selection in a volatile environment, where expected reward and expected uncertainty reduction blend to shape action utility. Here, three experiments probing human strategies of coping with uncertainty are compared to model-based predictions, revealing human response patterns are better predicted by the heuristic model than the optimal solution. Finally, ongoing scalp electroencephalography (EEG) was analyzed to reveal separable signals associated with action utility and uncertainty reduction. Together, the results discussed here suggest that the brain may deploy a computationally attractive heuristic strategy that blends reward and information gain when coping with action selection during periods of uncertainty.

Contents

List of Tables

List of Figures

Chapter One

Introduction

Contemporary research has generally pursued a value-based approach to understanding learning and decision making in the brain. This approach, which is born of the school of thought that one of the brain's primary functions is to make predictions, has focused on how the brain comes to endow actions or events with value, and how the brain weighs those values to guide action. An organism's capacity to predict the outcomes of its actions will have a strong impact on its fitness: its ability to find food, to mate, and otherwise avoid an untimely end Orr (2009). From this perspective, learning is a process through which predictions are shaped and assigned value according to experience, and action selection is a process of comparing predicted outcomes.

Theoretical and experimental work investigating the brain's prediction learning system have generally focused on the learning signals themselves as an entry point into understanding the brain's plasticity. In order to avoid unwanted variance and confounding/complicating factors, simplistic environments are typically investigated. However, this approach can't tell us the whole story given that the brain has evolved in a complex and volatile world. As a result, issues related to acting and learning when the state of the world is uncertain owing to its volatility have gone largely unaddressed. In this paper, I will attempt to collate and examine ideas related to operating in a shifting world, focusing primarily on the influence of latent state ambiguity; that is, when observable information cannot falsify all but a single hypothesis regarding the state of the world.

As a brief example, we imagine ourselves suddenly transported behind the wheel of a car exiting a parking lot. Inching our way forward while checking for oncoming traffic from the left, a car suddenly whizzes by us from the right. Startled, we reason that the world is not behaving as expected. Either we've tried to enter an unmarked one-way street, we're somewhere where they drive on the left side of the road, or

someone was driving like a maniac. Clearly it's in our best interest to ascertain which of these hypotheses is correct if we hope to get home alive. Having calmed down, we decide to gather a little more information when we observe a car traveling in the opposite direction in the left lane. Unless we've just observed two incidences of maniacal driving, it would seem that we're in a part of the world where they drive on the left side of the road. Accepting this hypothesis, we adapt our driving strategy and cautiously embark on a left-lane driving adventure.

In this example, we took advantage of observable information to disambiguate the state of the world and adopt the appropriate actions selection policy. To better understand the this process, I begin with an outline of various formal frameworks used to describe learning and action selection. With the relevant formal frameworks established, I will move on to investigate the relationships between those models and the brain. I will begin by investigating the neural processes thought to be involved in exploiting known reward contingencies, uncertainty and exploration, and environmental volatility. I will then outline an experimental design aimed at exposing various factors associated with information probing. We will first explore this design computationally, the results of which will inform two behavioural experiments. Finally, we will build on the computational and behavioural work to identify brain signals associated with information seeking behaviour using scalp electroencephalogram (EEG).

1.1 Formal models of learning and decision making

Early mathematical models aimed at understanding behaviour and brain function focused primarily on capturing the static representational characteristics and rules of neural processing. Examples like the work of McCulloch & Pitts (1943) generally avoided issues related to learning, the thinking at the time being that the dynamics involved were unnecessarily complicated. This left a considerable gap in our understanding of brain function. Ironically, as exemplified by the early works of Bush & Mosteller (1955), acknowledging the dynamics of learning actually simplify matters by defining the target to which learning was moving toward.

Formal models of learning and action selection have been developed using both a forward engineering approach in the domain of artificial intelligence (AI) and reverse engineering in the domain of brain sciences. AI typically tackles its challenges by by asking 'What is the optimal way to solve problem X?', whereas the brain sciences generally ask 'How does the brain solve problem X?'. Here, I outline some key computational strategies, beginning with a formalization of the learning and action selection problem, followed by methods of solving it.

1.1.1 Defining the learning problem: The Markov decision process

A Markov decision process (MDP) is a formal framework that specifies the interaction between an agent and its environment (Sutton & Barto, 1998). Typically, an MDP defines the environment as a finite set of discrete states, though continuous definitions

have also been developed (Doya, 2000). In sum, the agent selects and delivers an action to the environment, to which the environment responds by delivering a reward and a state signal indicating the new state of the environment.

An MDP makes two critical assumptions. First, the Markov property states that the next state and reward depend only on the current state and action. All information is summarized in the current state, meaning any additional knowledge related to preceding states and/or actions is superfluous, greatly simplifying what needs to be considered when making choices. Second, an MDP assumes that the state signal is fully informative of the environment's current state, which implies that the agent has full knowledge of the world during action selection. When this is not the case, the problem is referred to as a partially observable Markov decision process (POMDP), which will be discussed later. An MDP is specified by the tuple $\{S, A, T, R\}$ where:

- S: A set of states that define the environment, referred to as the state-space
- A: A set of actions the agent can perform
- $T(s', a, s) = Pr(s'|a, s)$: A state transition function that defines a probability of arriving in state $s' \in S$ given that the agent started in state $s \in S$ and took action $a \in A$.
- $R(s, a)$: A reward function that defines the reward, $r = R(s, a)$, delivered after taking action $a \in A$ while in state $s \in S$.

This specification fully defines the agent/environment interaction, and the one-step Markovian dynamics ensure that the next reward can be predicted (once learned) knowing only the current state and action. By iteratively sampling the environment, the agent can learn to predict future states and rewards.

Value functions and action policies

An MDP specification defines the problem the agent must solve; that is, given the current environment dynamics defined by the MDP, what is the best action to take for a given state? This, in turn, can be thought of as a search problem, where the goal is to find the optimal action policy, π, that defines the probability of taking each action at each state, $\pi(s, a)$. A flexible and intuitive approach is to operationalize optimality in terms of discounted future reward. Formally, the goal is to maximize:

$$E\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] \tag{1.1}$$

where E is the expectation, r_t is the reward delivered at time-step t, and $0 \leq \gamma \leq 1$ is a discount factor weighting reward value as a function of time-steps in the future. Geometrically discounting future rewards supports flexible policy development. As $\gamma \to 0$, future reward will have very little weight, giving preference to policies that pursue immediate rewards regardless of their long term consequences. As $\gamma \to 1$, distant rewards will have a greater impact on current expectations, effectively giving preference to policies that consider both immediate and future outcomes.

This definition of optimality depends solely on reward, making it highly flexible and generalizable. Unfortunately, this definition doesn't tell us how to actually find the optimal policy itself. Of particular concern is the fact that we need to somehow determine the expected value of discounted reward starting from the current time-step carrying on forever. One approach to this problem is to employ a value function that quantifies how good it is to be in a particular state, and iteratively improve our estimate of this value function by sampling the world. The value function takes the

form:

$$V(s) = \sum_{s'} T(s', a, s)(R(s, a) + \gamma V(s')) \tag{1.2}$$

The value function can be used to iteratively improve upon the action policy:

$$\pi(s) = \arg\max_{a} \left\{ \sum_{s'} T(s', a, s)(R(s, a) + \gamma V(s')) \right\} \tag{1.3}$$

where $\arg\max_a$ is the best action for the current state.

Various algorithms have been proposed for solving this optimization problem. Exact solutions are difficult to come by, but approximations have been found to work relatively well. One such approach is temporal difference reinforcement learning.

Model-free temporal difference reinforcement learning

The method of temporal difference reinforcement learning (TDRL) attempts to find the optimal policy by iteratively propagating reward values to credit the states and/or actions most predictive of their delivery. This is a solution to the temporal credit assignment problem, and has been proven to converge toward the optimal policy in the limit (Dayan & Sejnowski, 1994). Standard model-free RL is a particularly appealing means of finding the optimal policy for a given MDP due to its efficiency and simplicity, which comes from making no attempt to learn the transition function $T(s', a, s)$. Instead, the agent learns the value of states and/or actions by tracking expected future rewards regardless of exactly how those rewards were arrived at. Doing away with the computational complexities associated with learning an MDP's transition function facilitates the use of an extremely efficient representational structure of the environment, requiring only a cached value look-up table.

TDRL is a class of algorithms that employ the same basic policy learning strategy previously outlined. At its core is the reward prediction error which quantifies the discrepancy between the expected and actual outcomes. This learning signal may then be used in various ways in pursuit of shaping the action policy. The actor/critic framework was one of the earliest TDRL approaches to solving an MDP, which explicitly learns both a value function and an action policy (Sutton & Barto, 1998). The critic is responsible for maintaining the value function estimate, V, which is used to critique the action policy, π, maintained by the actor. The core of the actor/critic learning algorithm is a reward prediction error based on the critic's estimate of the value function:

$$\delta = r + \gamma V(s') - V(s) \tag{1.4}$$

where V is the current value function estimated by the critic, r is the reward delivered after transitioning out of state s and into state s'. This reward prediction error can then be used to improve the critic's value estimate for s:

$$V(s) \leftarrow V(s) + \alpha_c \delta \tag{1.5}$$

where α_c is the critic's learning rate. The same error signal is also used to update the action policy by adjusting the action weights that define the policy:

$$Q(s,a) \leftarrow Q(s,a) + \alpha_a \delta \tag{1.6}$$

where a was the action taken in state s, and α_a is the actor's learning rate. The actor/critic provides a flexible learning architecture that supports a modular decomposition of action and state value learning, which has been suggested to be implemented by the brain (O'Doherty et al., 2004; Holroyd & Coles, 2002).

Since action weights are decoupled from the value function, stochasticity in the

environment can lead to unbounded weight growth. Weight normalization or learning rate annealing techniques can be applied to remedy the problem; however, a simpler solution is to collapse the value function and action policy into a single structure. One approach, referred to as Q-learning, estimates the expected reward of state/action pairs, $Q(s, a)$ (Watkins & Dayan, 1992). Here, the core learning signal is based on action values directly:

$$\delta = r + \gamma \arg\max_{a}[Q(s', a)] - Q(s, a) \tag{1.7}$$

where $r = R(s, a)$ is the reward delivered by taking action a while in state s, γ is the discount factor, and $\arg\max_a[Q(s', a)]$ is the current estimate for the best known action in the subsequent state s'. The reward prediction error quantifies the discrepancy between the agent's current prediction, $Q(s, a)$, and actual outcome, which is a combination of the immediate reward, r, and a discounted estimate of reward value to be expected at the subsequent state, $\gamma \arg\max_a[Q(s', a)]$.

Note that the Q-learning prediction error is constructed according to the action with the highest estimated value at the subsequent state regardless of the action actually taken. This is referred to as off-policy learning since the learning signal is independent of actions selected according to the policy. This is in contrast to on-policy learning where the prediction error is computed using the value of the action actually taken at the subsequent state. SARSA (state-action-reward-state-action) is an example of on-policy learning, where the prediction error is computed as:

$$\delta = r + \gamma[Q(s', a)] - Q(s, a) \tag{1.8}$$

Regardless of whether on- or off-policy learning is used, the learning signal is

used to improve the agent's policy:

$$Q(s,a) \leftarrow Q(s,a) + \alpha\delta \tag{1.9}$$

where α is a learning rate. Higher learning rates allow estimates to reflect more recent outcomes and rapidly adapt to to recently observed outcomes. Conversely, lower learning rates allow Q-values to reflect longer outcome histories, which can lead to more stable and accurate reward estimates in stochastic and/or unstable environments.

The efficiency of model-free learning comes with a cost. Learning must bind reward value with preceding actions and states, possibly across large temporal gaps with multiple intervening events. In volatile and/or complex environments the value propagation process may be impossibly slow. Furthermore, the action policy can only be updated when outcomes are experienced, meaning the learning agent must sample the environment to learn. In short, model-free TDRL lacks the adaptability many organisms seem to exhibit.

Model-based temporal difference reinforcement learning

Standard model-free RL makes no attempt to learn state transition probabilities: values are represented but an explicit model of the environment is lacking. That is, the agent has no way of distinguishing between actions that lead to different but equally valued outcomes. This may be problematic if the agent's goals are subject to change. Model-based approaches have been developed to provide a means of exploiting additional information in the state transition function, $T = (s', a, s)$ (Sutton & Barto, 1998). One such approach attempts to maintain and improve

estimates of $T = (s', a, s)$ by calculating a state prediction error, δ_s at each state transition:

$$\delta_s = 1 - T(s', a, s) \tag{1.10}$$

where $T(s', a, s)$ is the probability of transitioning into state s' after taking action a while in state s. The state prediction error can then be used to update the transition probabilities:

$$T(s', a, s) \leftarrow T(s', a, s) + \alpha_s \delta_s \tag{1.11}$$

where α_s is a state transition learning rate. Since $T = (s', a, s)$ defines a probability distribution, the transition probabilities for all other states, s'' must be normalized:

$$T(s'', a, s) \leftarrow T(s'', a, s)(1 - \alpha_s) \tag{1.12}$$

Like model-free RL, model-based RL computes Q-values to define the action policy; however, model-based Q-values can incorporate additional state transition information. One method of doing so is to update Q-values by recursively defining values at the current state by those at successive states, weighted by the probability of transitioning into those states:

$$Q(s, a) = \sum_{s'} T(s', a, s) \times (R(s, a) + \arg\max_a [Q(s', a)]) \tag{1.13}$$

By explicitly representations state transition probabilities, a model-based approach is capable of considering not only expected rewards but resulting states as well. This can be particularly useful when flexible, goal-directed behaviour is required. However, these benefits come at a computational cost. In order for Q-values to accurately reflect transition probabilities, then transitions into all possible states

must be considered, not just the subsequent state actually encountered as in the model-free approach. In sum, a model-based approach provides a powerful structure for learning and decision making, but it becomes computationally intractable in a large state-space.

Action selection

TDRL algorithms quantify the expected value of an action given a particular state. The most straightforward approach to action selection is to simple choose the action with the highest expected value for the current state. This is referred to as adopting a greedy policy. This may work well in special circumstances, but unfortunately a greedy policy can be victimized by its simplicity. Exclusively exploiting what is currently estimated to be the best action at each state prohibits sampling alternative actions that may be more rewarding. A greedy policy's exploitive dedication may prevent it from fully exploring the policy space, making it susceptible to local maxima.

This can be avoided by relaxing the requirement that action selection choose greedily all the time. An ϵ-greedy approach selects actions with the highest expected reward on a proportion of choices specified by $0 \leq \epsilon \leq 1$, but selects an action at random the rest of the time. The ϵ parameter can be adjusted to make action selection depend more, $\epsilon \to 1$, or less, $\epsilon \to 0$, on the current policy. Although this approach can escape local maxima, random exploration implies that disastrously bad actions may be selected even when the agent knows better.

Exploration would more sensible if the probability of taking an action depended on that action's expected value. A Softmax solution does just this, where the prob-

ability of selecting action a in state s is given by the sigmoid function:

$$P(a|s) \propto e^{\beta Q(s,a)} \tag{1.14}$$

where β controls the function's slope (Sutton & Barto, 1998). Based on this form of the equation, β can be thought of as a weight that determines the degree to which expected value will govern action selection. As $\beta \to \infty$, more emphasis is given to the values themselves, making action selection more exploitive. Conversely, as $\beta \to 0$ Q-values are less influential, making action selection more explorative and random. Indeed, recent investigations into exploratory action selection suggests that people may adopt something like a Softmax action selection strategy (Daw et al., 2006).

1.1.2 Learning in large & complex environments

TDRL, particularly model-free variants, are simple, efficient, and relatively flexible. However, the single-state value propagation process can make learning prohibitively slow, which may be exposed in environments where sequences of actions are required to reach a goal. Take the environment depicted in Figure 1.1 as an example. The agent's goal is to reach the state denoted with a star. To do so the agent only needs to learn which of two possible actions should be selected at state S_1, with all other states offering only a single action forward. Assuming the agent makes the correct choice upon first encountering S_1, it will unavoidably receive a reward. But, due to the one-state value propagation approach employed by standard TDRL, the agent will learn nothing about it's choice at S_1 on this first pass through the environment. Action selection will be random when the agent returns to state S_1.

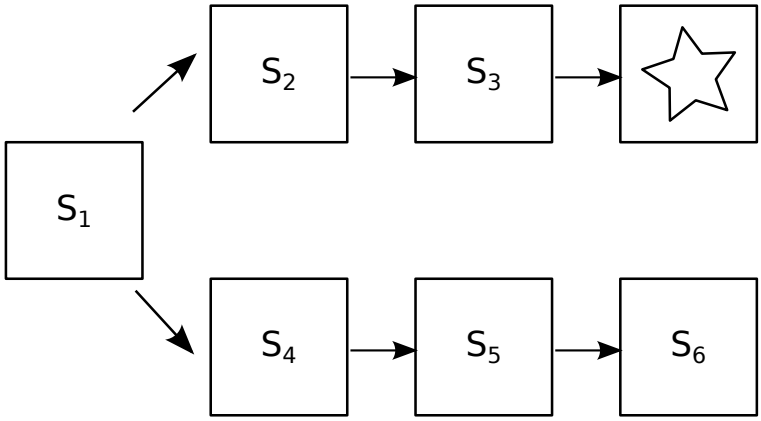

Figure 1.1: Single choice value propagation

Eligibility traces

One way to improve value propagation is for the learning signals to influence multiple state/action values simultaneously. Eligibility traces do just this. An eligibility trace acts much like a short-term memory trace of states and/or actions encountered by the agent. The magnitude of these traces, which decay over time, determines how eligible values are for learning. Assume that at time t, the agent takes action a_t while in state s_t. The eligibility trace for all state/action pairs is updated according to:

$$e_t(s,a) = \begin{cases} 1 & \text{if } s = s_t \text{ and } a = a_t \\ \gamma\lambda e_{t-1}(s,a) & \text{otherwise} \end{cases} \tag{1.15}$$

where $0 \leq \lambda \leq 1$ is the eligibility decay rate. The eligibility trace can be incorporated into a learning algorithm such as Q-learning by updating all state-action pairs after each transition according to the trace weight:

$$Q(s,a) \leftarrow Q(s,a) + \alpha\delta e(s,a) \tag{1.16}$$

where α is the learning rate and δ is a standard reward prediction error. In this way, learning signals stretch back in time, updating the value of distal actions to

facilitating rapid learning.

Hierarchical TDRL

Eligibility traces can improve learning when the task structure is comprised of many sparsely connected states. Sparsity allows learning signals to quickly and effectively carve out rewarding paths through the environment. However, the benefits of using eligibility traces decline as task structure complexity increases, where states afford multiple actions and can transition into multiple states. Here, the spatial credit assignment problem may come to overwhelm the benefits provided by traces.

Recent interest has come to focus on hierarchical TDRL (HRL) as a potential solution to learning and action selection in large and complex environments. Consider, for example, leaving your house for work in the morning. This simple task could be decomposed into several sub-tasks, such as getting your coffee, packing your bag, finding your keys and wallet, etc... HRL attempts to extract and encapsulate sub-tasks, referred to as options, embedded within the larger task domain for reuse. Here, options are closed-loop action policies, and as such, they serve to encapsulate temporally extended sequences of actions into a single unit (Sutton et al., 1999). Importantly, options can be arbitrarily abstract, but are acted upon as though they were singular concrete actions. That is, an option may be a single low level action (e.g. a muscle twitch), or consist of a nested hierarchy of options (e.g. leaving the house for work). Once an option has been selected, that option defines the action policy until it terminates. As such, developing an efficient option decomposition can reduce the choice space considerably, improving learning and simplifying action selection.

HRL can be carried out by augmenting a standard TDRL algorithm with the additional structure and signals required to learn and operate upon options. A neurologically inspired implementation described by Botvinick et al. (2009) is rooted in an actor/critic architecture. The framework includes a standard actor and critic that learn a 'root' action policy and value function respectively. Option structures are also included, with each option maintaining its own action policy and value function, but also defining their own termination function. HRL differs from standard TDRL in that attaining a sub-goal, as defined by an option's termination function, delivers an intrinsic reward referred to as a pseudo-reward. This pseudo-reward is critical in that it shapes an option's policy toward actions that lead to that option's sub-goal, and it does so independently of external rewards. As such, the learning agent is able to learn useful options, but it won't unlearn options when they are inappropriately selected; rather, it will learn not to select that option in the future.

Note that within this framework primitive actions are considered to be single step options with a pseudo-reward of zero. Thus, at least two prediction error will be generated upon completion of a composite option. A pseudo-reward prediction error is spurred by the termination of the composite option's last primitive action (using the pseudo-reward) and used to drive learning in that option's action policy and value function. A standard reward prediction error is also generated by the composite option's termination, which is used to drive learning in the parent option (possibly the root policy). Recent work by Ribas-Fernandes et al. (2011) identified neural signals consistent with pseudo-rewards, suggesting that the brain may employ a computational strategy akin to the HRL framework.

HRL can of deliver a significant performance boost when learned options are well suited to the environment (Sutton et al., 1999). However, in a volatile environment, mutiple policies may need reshaping to accomodate the environment's new structure,

which compounds the learning problem, and can severely compromise the agent's performance (Botvinick et al., 2009).

A larger theoretical challenge lay in determining how pseudo-rewards and termination functions should be integrated into an HRL framework. In standard TDRL, the external environment determines the reward function. However, HRL pseudo-rewards are intrinsically generated, and as such, defining a set of sub-goals is entirely in the modeller's hands. This might be acceptable in an engineering context, but it requires inference of the very processes under investigation when using a reverse engineering approach as required in cognitive neuroscience. This doesn't argue against the fact that the brain may indeed be using something akin to pseudo-rewards to guide learning and action selection, and indeed recent work by Ribas-Fernandes et al. (2011) identified neural signals consistent with pseudo-rewards, but does imply that investigation into the neural correlates of HRL must be done with caution.

1.1.3 The partially observable Markov decision process

We return now the driving example outlined in the introduction. We have ample driving knowledge and skill, but this knowledge can only be applied if the state of the environment is known. Indeed, if we were given the pertinent information, namely that we were somewhere in western Ireland, then the learning and action selection methods just outlined would provide a good solution to the problem at hand. Unfortunately, those strategies don't offer much in the absence of state information owing to the fact that both the action policy and the value function are defined in terms of having a completely observable state signal.

We could adopt an 'ignorance is bliss' approach by simply picking the most likely

state and behaving as though that was the true state of the world. This approach may suffice when the cost of guessing wrong is low, but this seems unnecessarily risky in our case. We might offer a better solution if we acknowledge our ignorance of the true state of the world and recognize the fact that state signals delivered by the environment can be incomplete.

A partially observable Markov decision process (POMDP) assumes that the environment's dynamics are determined by an MDP, but the agent does not have direct access to state information (Kaelbling et al., 1998). Since state information may not be directly observable, a POMDP is defined in terms of observations that are available to the agent. As such, the agent must maintain and operate on a belief-state, b, which is a probability distribution over all possible states. A POMDP is defined by the tuple $\{S, A, O, T, \Omega, R\}$ where:

- S is a set of states
- A is a set of actions
- O is a set of observations
- $T(s', a, s) = Pr(s'|s, a)$ is a set of conditional transition probabilities that define the probability of landing in state s' after taking action a in state s.
- $\Omega(o, s', a) = Pr(o|s', a)$ is a set of conditional observation probabilities that define the probability of making observation o after taking action a and landing in state s'.
- $R(s, a)$ is a reward function defining the immediate reward received after taking action a while in state s.

At each time step the agent delivers some action, $a \in A$, to the environment in

state $s \in S$. This causes the environment to transition according to the underlying MDP, captured by $T(s', a, s)$, delivering reward $r = R(s, a)$ and observation $o \in O$ with probability $\Omega(o, s', a)$. Critically, a POMDP adds a layer of abstraction in the form of observations, the hope being that this additional information can be leveraged to disambiguate states of the environment.

Solving a POMDP

A POMDP is solved by finding a policy that optimizes a performance objective. As we did for MDPs, we assume that the objective is to maximize the expected discounted rewards over an infinite horizon. Unfortunately, a POMDP presents an additional challenge since the environment's state cannot be observed directly. However, the probability that that the environment is in state s can be calculated. This transforms the optimization issue from one of maximizing reward of a finite set of discrete states into one of maximizing reward over a finite set of probability distributions. Each distribution, referred to as a belief state, represents the agent's belief that the world is in a particular state.

Let b be a belief state, and $b(s)$ be the probability that the the environment is in state s. If action a is followed by observation o, the successor belief state is defined by $b' = \tau(b, a, o)$, where $b'(s')$ denotes belief that the environment is in state s' according to:

$$b'(s') = \frac{\Omega(o|s', a) \sum_{s \in S} T(s', a, s) b(s)}{Pr(o|a, b)} \tag{1.17}$$

where $Pr(o|a, b)$ is a normalizing factor ensuring the belief state is a proper distribution. In short, the new belief of being in a particular state s' is defined in terms of the probability of transitioning into that state, $T(s', a, s)$, weighted by the previ-

ous belief state, all weighted by the probability of making observation o. Critically, including the previous belief state in the update function allows information from previous actions and observations to factor into the current belief state. This acts a form of memory, and if done properly, ensures that the process over belief states is Markovian. We can now define a belief state transition function as:

$$\tau(b', a, b) = P(b'|a, b) = \sum_{o \in \Omega} P(b'|a, b, o)P(o|a, b) \tag{1.18}$$

In line with our approach to finding an optimal solution for a discrete MDP, we can specify an optimal POMDP policy by taking the best known action for the current belief state and acting optimally thereafter:

$$V(b) = \arg\max_a [r + \gamma \sum_{o \in O} \Omega(o, s', a)V(\tau(b', a, b))] \tag{1.19}$$

Here, future rewards are the discounted value of all possible subsequent belief states for a given observation, $V(\tau(b', a, b))$ weighted by the probability of those observations (Kaelbling et al., 1998).

An exact solution demands that the optimal action for every possible belief state be known. This presents an extremely challenging, and intractable problem for all but the most trivial environments. Various approaches have been applied to POMDP policy search which may be grouped into three classes: model-free memoryless, model-free with finite memory, and model-based. Model-free memoryless approaches treat POMDP observations as though they were states and naively apply a standard model-free MDP solution, but can perform surprisingly well in some circumstances (Loch & Singh, 1998). Model-free finite memory solutions employ external memory that can be used to augment and disambiguate observational in-

formation, effectively transforming the POMDP into an MDP(Peshkin et al., 2001). This approach has been successfully applied to human behaviour and neural signals in the brain (Todd et al., 2009; Rougier et al., 2005; Braver & Cohen, 2000; Hazy et al., 2006). Model-based solutions take a more ambitious approach and attempt to learn the underlying POMDP structure.

Model-based policy search

Various model-based approaches have been developed to search for the optimal (Kaelbling et al., 1998) or a near optimal (Hansen, 1998) POMDP policy. Unfortunately, model-based approaches come at a high computational cost due to the complexities involved in learning $\tau(b', a, b)$ and $R(s, a)$ simultaneously across a continuous belief state space. For example, the *witness* algorithm proposed by Kaelbling et al. (1998) finds the optimal policy for a POMDP by iteratively testing policies to see if they offer any improvement in the expected value. This requires computing the expected value of a given policy across all possible belief states, and including that policy if there exists a belief state for which that policy offers a higher expected value than any existing policy. Computing the expected value is non-trivial, as it requires sampling the expected value of all possible observations in all possible belief states, which in turn depends on the discounted expected value of all possible observations at all possible subsequent belief states. Sufficed to say, computing these values for anything but a relatively small and stable environment quickly becomes intractable. Further complicating matters is the fact that finding the optimal policy depends upon complete knowledge of both the structure and transition dynamics of the belief state space, which is rarely if ever available to an organism interacting with the world.

A simplified model-based framework that does not attempt to learn $\Omega(o, s', a)$, and thus is not forced to learn a policy over the entire belief-space has been proposed by Doya et al. (2002). The approach is based on a model-based RL framework augmented with the capacity to learn and operate according to multiple policies. Competition and cooperation among policies leads to a POMDP decomposition that can be tackled by standard RL algorithms such as Q-learning.

The framework is rooted in modules comprised of a reinforcement learning and observation prediction controllers. Each module's observation prediction controller learns the transition function $\Omega(o, s', a)$, while it's RL controller learns an action policy π. At each time step each module generates a candidate action, one of which is selected for execution. Given the selected action and the current belief state, the observation controller predicts the subsequent observation. Once the ensuing observation is delivered by the environment, a responsibility signal quantifying the discrepancy between each module's prediction and the actual observation is calculated.

This responsibility signal enforces a competitive drive among modules, which in turn, promotes task decomposition. The responsibility signal is used to weight learning signals for both the RL and prediction controllers, emphasizing learning in modules that are doing a better job of predicting observations. This weighting may appear counterintuitive. One might suggest that it would be better to emphasize learning in modules that are not yet capable of predicting the task dynamics with hopes of improving their performance. However, emphasizing learning signals in modules according to prediction accuracy ensures that each module will focus on learning a unique aspect of the task, which helps to develop an advantageous task decomposition.

POMDP task dynamics are typically ambiguous and unpredictable when the task is considered as a whole and only current observations are considered. Standard RL approaches that attempt to find a single optimal policy are victimized by opposing pulls from the underlying but unobservable MDP. However, regularities in the environment may be exploited by learning task sub-structures, effectively carving the task at points of ambiguity driven conflict. Although it would greatly simplify learning and action selection if we could describe the world as an MDP, the fact is that all organisms depend on exhaustible and error prone sensory systems and interact with an environment full of hidden forces. A POMDP includes state uncertainty as an explicit component of the task description, making it a more complex but realistic description of the environment.

We will now move on to discuss how the brain might employ strategies linked to the formal models outlined here, beginning with the neural circuitry thought to facilitate exploitive action selection. I will then discuss issues pertaining to exploration strategies, followed by an analysis of how the brain might monitor ongoing performance in order to adaptively adjust according according to goals and environmental dynamics.

1.1.4 Quantifying uncertainty

Uncertainty is a heterogeneous concept that can be decomposed into at least four sub-categories (Payzan-LeNestour & Bossaerts, 2011). At the highest level, when developing a model of the environment, we encounter structural uncertainty. This reflects the degree of knowledge regarding the relevant elements in the environment and the relationships among them. Once relevant model structures have been more or less identified, we can begin to ask what is known about the model's constituent

parts. The degree of knowledge for a given model element is referred to as its estimation uncertainty, essentially capturing the reliability of our understanding of that element and what is left to be learned about it. This is a broad concept that can be subdivided further by addressing the source of uncertainty as expected or unexpected (Yu & Dayan, 2005).

Nearly all natural processes have some level of inherent and irreducible stochasticity: flipping a coin for example. One can have perfect knowledge of the underlying process and be able to predict its longterm summary statistics with great precision; however, the outcome of a single sample may nonetheless remain highly unpredictable. To know the underlying process requires knowledge of this irreducible stochasticity. This is captured as expected uncertainty, also referred to as risk. Conversely, there may be forces acting within the environment that are unknowable and therefore unpredictable. Unbeknownst to us, samples are suddenly generated by a loaded coin, resulting in unexpectedly long runs of heads or tails. Unexpected uncertainty, also referred to as volatility, speaks to inherently unobservable changes in the environment that influence observations. This alludes to an intimate and antagonistic relationship between expected and unexpected uncertainty. Environmental changes can be identified reasonably easily when outcomes are relatively deterministic (low expected uncertainty). However, matters are muddied as outcomes become increasingly stochastic. In the absence of any additional information, one could never know if a series of outcomes came from a single fair coin, or a random selection between two biased coins.

Returning now to the domain of partially observability within the context of our driving example outlined in the introduction. The challenges presented to us as a driver are many fold. First, after observing a car coming from an unexpected direction, we must determine whether the current level of uncertainty is cause for concern.

Working within a POMDP framework, state uncertainty is tantamount to the variance of the belief state distribution. In accordance with the decomposition outlined here, belief state variance can be cast as estimation uncertainty over states. We can subdivide this further as a combination of expected and unexpected uncertainty, quantifying irreducible ambiguity in the observations delivered by the environment, and the environment's volatility. Second, should uncertainty be deemed to be unacceptably high, we must determine which action(s) are likely to effectively reduce the current degree of uncertainty before choosing actions in pursuit of the true goal. Finally, newly acquired information must be integrated to inform future decisions. These present significant, and in general, unsolvable problems. However, attempts have been made to find optimal solutions when possible, or good approximations when a generalizable optimal solution appears beyond our reach.

Quantifying the value of exploration

A little exploration is usually a good idea when the best path forward is unclear. But just how uncertain should you be and how can you quantify the possible benefits that may come from exploration? In our previous discussions of optimal performance we assumed operation under an infinite horizon in which the agent has an unlimited number of opportunities to both sample and exploit the environment's reward structure. This assumption greatly simplifies reasoning about optimal performance; however, an infinite horizon implies that the tradeoff between exploration and exploitation can be ignored. Yet, organisms are often forced to operate in environments that are better thought of as having a finite horizon. Indeed, the organism's horizon may become undesirably short if they don't balance the exploration/exploitation tradeoff well. However, we can leverage the generality of an infinite horizon by viewing the discount factor $0 \leq \gamma < 1$ as the probability that the task will continue after

the current trial. This effectively translates the infinite horizon problem into one concerning a probabilistic termination point and finite expected number of trials.

Optimal performance within a finite number of trials demands a balance between acquiring information and exploiting that information to maximize reward. Actions that appear sub-optimal with respect to their immediate outcomes may in fact be highly informative regarding the environment's structure. This knowledge could be exploited in the future, suggesting an intimate relationship between the value of exploration and the horizon being considered. However, weighing the relative benefits of exploring the environment or exploiting current knowledge requires some means of quantifying each in comparable units. But how are we to quantify the unknown?

The value of information can be quantified by asking how much one is willing to pay for it. Decision theory (and common sense) suggests that we should never pay more than the profit we stand to gain by attaining new information. We begin with the thought experiment, as outlined by (Howard, 1966), in which an agent has some degree of uncertainty regarding a competitor's bid on a lucrative contract. If the agent's bid is too high they will loose the contract and profit nothing. If the agent bid is too low they may win the contact but loose money due to the cost of contractual obligations. Thus, it's in the agent's best interest to make every reasonable effort to reduce its uncertainty regarding the competitor's bid. Assume that the agent can gain access to a clairvoyant with knowledge of the variable of interest (the competitor's bid) for a price. What should the agent be willing to pay o have some or all uncertainty eliminated? Conceptually, the value of information is the difference between the expected value of an action given additional information or not:

$$E[I_x] = E[a|I_x] - E[a] \tag{1.20}$$

where I_x is denotes information regarding the variable of interest.

Relating this idea back to the issue of acting under state uncertainty, the agent should be willing to pay for information that will eliminate some degree of state uncertainty as long as the cost of information is less than the expected value of the subsequent state. Unfortunately, analytically computing these values presents an extraordinarily difficult task and is seldom possible. However, Gittins & Jones (1979) outlined a strategy and demonstrated its optimality for a special class of problem known as the n-arm bandit task. In its simplest form, a stationary n-armed bandit problem presents a learning agent with n possible actions, each delivering a reward of 1 or 0 with constant but unknown probability, p_i. The agent's goal is to maximize the number of rewards within N trials. The agent must sample each action sufficiently such that the best option can be identified, but cannot sample so much that they miss rewards they could have received before the task ends.

An n-armed bandit task can be framed as one of estimating the unknown parameters p_i of the Bernoulli process delivering rewards for each action. Given a horizon of N trials, the Gittins index quantifies the difference between exploring and exploiting by recursively computing the value of all possible outcomes that could be observed for each action. Assuming a 2-armed bandit problem for simplicity, the Gittins index quantifies the difference between the expected value of the best known action (exploitation) and the expected value of each action after observing all possible outcomes followed by the optimal policy thereafter.This implies that the expected value depends upon both observed outcomes and the number of samples; and as such, the explore/exploit tradeoff is balanced by incorporating reward certainty into expected value.

Conceptually, the Gittins index facilitates a comparison between the best known

action weighted by the certainty of its value, and the value of alternative actions weighted by the certainty of their values. In general, the Gittins index favours an exploratory strategy early in learning before the expected value for any one action has had a chance to shake its uncertainty. An exploratory strategy will also be adopted longer if expected values are closer together. Both human and non-primate choice behaviour during an n-armed bandit task has been shown to approximate the optimal (Acuña & Schrater, 2010) or an approximation of the Gittins index (Krebs et al., 1978). This suggests that the brain may balance the explore/exploit tradeoff according to some approximation of the Gittins index; however, it's unlikely that the brain computes the true index value at each choice point.

First and foremost, calculating the Gittins index is computationally intractable in general. Solving the n-armed bandit problem involves solving a system of $N^{a_i * o_i}$ non-linear equations at each time-step, where N is the horizon, a_i is the number of possible actions, and o_i is the number of possible outcomes. This presents a number of seemingly insurmountable challenges, one of which being that n-armed bandit tasks may return continuously valued rewards, requiring calculation of an infinite number of possible outcomes for each action. Even if such a computation were possible, it would seem to provide minimal benefit in light of the resources (time and calories) required for its computation. Additionally, the Gittins index depends upon knowledge of the termination point, N, which is seldom known in the wild. Finally, the Gittins index only provides the optimal solution to tasks that can be reduced to an n-armed bandit. Although it's possible that organisms have developed specialized solutions to various commonly encountered problems, it seems more likely that a generalizable solution would be favoured.

In sum, the Gittins index illustrates that introducing a tension between reward and uncertainty can provide an optimal solution in a confined and well understood

problem space. Although it remains to be seen whether similar approaches can be found for other problems, it highlights the strength of targeting exploration toward the reduction of uncertainty. Although it is unlikely to be the preferred strategy of living organisms, it does define what optimal performance should be.

Actively navigating uncertainty

A more general framework for considering strategies of exploratory sampling is referred to as active learning. Active sampling adopts a strategy somewhere between optimal and random sampling. An active learning agent aims to achieve greater accuracy with fewer data samples by choosing the data samples according to some measure of informativeness (Settles, 2009). Given a set of hypotheses with associated prior probabilities, and a set of possible actions with associated costs, the aim is to choose a series of actions that minimizes the expected cost of eliminating all but one of the hypotheses (Bryant et al., 2001). A hypothesis may take many forms. It might define a category boundary between visual stimuli, or it may be a set of rules that define a logical relationship between entities and categories. In general, the problem can be translated into one of finding a binary decision tree with minimum expected cost, where actions are nodes and hypotheses are leaves. Unfortunately, finding the optimal solution to a binary decision tree search is known to be NP-hard (Fedorov, 1972).

However, problem specific structure can often be leveraged to provide a tractable and satisfactory solution that may not be optimal but is nonetheless significantly better than passive learning approaches. Many solutions have been proposed and applied to category learning problems such as word identification for speech recognition (Dagan & Engelson, 1995), data mining in large data sets (Settles et al., 2008),

and automated scientific hypothesis testing (King et al., 2004). These examples are by no means generalizable learning solutions, as they take advantage of domain specific features and hypothesis generation that depend upon expert knowledge encoded in a logical framework. Nonetheless, they do demonstrates that actively sampling actions according to a measure of entropy can significantly improve learning efficiency.

1.2 Learning and decision making: Behaviour and its neural correlates

Over the past two decades an overwhelming mass of data, probing everything from social behaviour to ion channel dynamics, has advanced our understanding of how the brain learns about the environment and exploits this information to maximize reward. The challenge it seems is not producing data, but integrating that data into a unified framework. Formal models such as those just discussed have encouraged such integration, shifting our endeavour from information collection into a theoretically driven science. I begin our discussion by focusing on the midbrain dopamine system, which is through to encode the brain's reward prediction error. We will then move on to review major targets of the dopamine system, the basal ganglia in particular, in hope of gaining some appreciation for how the brain learns and exploit reward information.

1.2.1 Dopamine's many roles

The midbrain dopamine (DA) system and its influence on target systems has become one of the most intensely studied topics in the brain sciences. This is due in

part to the emerging story of DA's impact on range of brain functions in addition to it's role in a wide range of atypical behaviour and psychiatric disorders. DA is produced by dopamine neurons in the substantia nigra par compacta (SNc) and the ventral tegmental (VTA) areas (Björklund & Dunnett, 2007). Unlike most neurotransmitters, DAs impact on post-synaptic targets cannot be adequately described in terms of excitation or inhibition; rather, its influence is better described as a neuromodulator, the effect of which depends upon the target system and its current state (Hernández-López et al., 1997; Schultz, 2002).

DA has been proposed to play a critical role in gating information among neural sub-systems by modulating the activation function of target neurons (Goto & Grace, 2008; Servan-Schreiber et al., 1990). DA is also thought to facilitate long term potentiation (LTP) (Pedarzani & Storm, 1995) and long term depression (LTD) (Sajikumar & Frey, 2004) in both frontal cortex (Goldman-Rakic et al., 1989) and the basal ganglia (Smith & Bolam, 1990) via synaptic triads. At these sites, DA terminals make contact with a synapses to implement a three-factor learning rule such that connections are only strengthened when the pre-synaptic, post-synaptic and dopamine neurons are simultaneously active (Wickens et al., 1996). This can be thought of as implementing a modified Hebbian learning rule where neurons that fire together wire together, but only when there's enough dopamine around. Midbrain DA neurons exhibit two distinct modes of activity: tonic and phasic.

Phasic dopamine activity and its theorized significance

Afferent drive can push the midbrain DA system into a phasic mode of activity (Floresco et al., 2003). This phasic activity, characterized by a rapid and punctate increase or decrease in firing rate, results in a correspondingly rapid increase

or decrease in the concentration of synaptic DA. The temporal dynamics of the DA systems phasic component has been extensively studied with respect to learning and action selection. In what has become seminal work, phasic bursts are initially observed following unexpected reward delivery (Schultz et al., 1997), transiently increasing synaptic DA levels. As reward contingencies continue to be experienced, phasic burst activity shifts from the time of reward delivery to the time of stimuli predicting the reward. Should an expected reward be withheld, a precisely timed phasic pause is observed during which synaptic DA concentrations fall below tonic baseline levels (Schultz et al., 1997). These results demonstrate that dopamine neurons are not sensitive to reward per se; rather, they appear to be involved in a signalling reward predictability.

A transient increase or decrease in the firing rate of DA neurons has been hypothesized to encode a reward prediction error (Montague et al., 1996; Schultz et al., 1997). Within this framework, a phasic increase in the firing rate of midbrain DA neurons indicates that an event was better than predicted; and conversely, a phasic decrease in firing rate indicates that an event was worse than expected. A considerable body of evidence has shown that midbrain dopamine system activity is compatible with a TDRL model in which the DA system encodes a reward prediction error signal that can be used by target systems for reinforcement learning (Montague et al., 1996; Schultz et al., 1997; Frank et al., 2004; Rutledge et al., 2010). Probing the DA system deeper has revealed that reward prediction errors encoded by the DA system closely match behavioural preferences based on reward magnitude (Tobler et al., 2005), probability (Satoh et al., 2003), and delay (Kobayashi & Schultz, 2008), demonstrating the the midbrain DA system is sensitive to high order statistics of the reward distribution in accordance with TDRL.

Establishing a link between DA and a formal learning model has brought consid-

erable attention and resources to bear on the issue of exactly what the DA system is doing. Recent research has begun to uncover more complex and heterogeneous functions within the midbrain DA system, suggesting that its function goes beyond encoding reward prediction errors as initially proposed. Perhaps the most consistent and long standing complication comes from results showing the DA neurons respond to stimuli that are novel but are not associated with reward (Schultz, 1998). Attempts have been made to integrate these signals as novelty bonuses intended to induce exploration (Kakade & Dayan, 2002), but little evidence has come to support this hypothesis.

More recent studies have found that DA neurons express a preference for informative cues over uninformative cues (Bromberg-Martin & Hikosaka, 2009). In this study, monkeys learned reward contingencies associated with various cues, some of which were informative of the coming reward while others were not. DA cells came to elicit phasic activity in accordance with cue predicted reward. Monkeys were then shown an additional cue informing them as to whether they would be seeing an informative cue (good or bad) or an uninformative cue. Surprisingly, DA neurons responded to information predictive cues as though they had reward value, despite the fact that reward expectation was equal for the information predictive and no-information predictive cues. Furthermore, monkeys displayed a strong preference for informative over uninformative cues, coinciding with observed phasic activity. This is taken to suggest that DA neurons motivate actions to acquire rewards, as well as actions that deliver information allowing accurate predictions to be made. What is not clear is whether the system views information as inherently rewarding.

Finally, phasic DA signals have also been proposed to play a role in the action selection process itself by modulating incentive salience (McClure et al., 2003b). DA receptor antagonists have been shown to have a strong and immediate impact on

behaviour, suggesting DA is not only involved in the valuation of rewards themselves but the initiation of actions necessary to acquire them (Berridge & Robinson, 1998). Seeking to understand DA within both reward learning and incentive salience frameworks, (McClure et al., 2003b) proposed a model in which prediction error signals facilitated value learning and modulated the probability of executing a selected action. This model can account for learning effects traditionally associated with phasic DA signals, but it can also predict the immediate behavioural effects associated with DA receptor antagonism, typically observed as a reduction in reward seeking behaviour.

To summarize, the phasic component of DA signalling has been intensely studied, and chiefly thought to convey a reward prediction error signal used to drive learning in target systems. Despite the dominance of the prediction error hypothesis, Caplin & Dean (2008) correctly point out that it suffers from a shaky empirical foundation that could be remedied by carful experimentation aimed at testing an axiomatic grounding. However, more recent and detailed electrophysiological work suggests that midbrain DA neurons exhibit more heterogeneous activity and facilitate a wider set of functions than previously thought (Bromberg-Martin et al., 2010).

Tonic dopamine and its theorized significance

Tonic activity is characterized by moderate, consistent, and self-sustained firing rates (2-10Hz) (Grace & Bunney, 1983) regulated by afferent stimulation from the ventral pallidum (Floresco et al., 2003) and by inhibitory DA auto-receptors that respond to extracellular DA concentrations (Grace, 2000). Despite the seemingly endless resources and effort aimed at understanding dopamine's role in the brain, relatively little attention has been given to the dopamine system's tonic phase of activity.

Biophysical models incorporating spatial and temporal cellular constraints have demonstrated that tonic, but not phasic, activity can influence levels of extra-cellular DA (Cragg & Rice, 2004). This has been hypothesized to regulate the activation function of neurons in the sphere of influence, effectively controlling the flow of information through the system (Goto & Grace, 2008). Extra-cellular DA can also influence the midbrain DA system itself via inhibitory auto-receptors, pointing to a complex feedback loop whereby tonic activity has an indirect effect on both phasic and tonic modes of activity (Grace, 2000). Tonic DA levels have been shown to play a critical role in providing a baseline from which phasic signals exert their influence on target systems whereby atypically low or high tonic activity will bias learning driven by phasic DA signals (Frank et al., 2004)

More recently, it has been suggested that tonic DA encodes the average reward rate which modulates response vigour in free-operant behavioural tasks, where response vigour is the rate at which actions are made (Niv et al., 2007). It is commonly observed that animals respond at a rate that corresponds to the rate of reinforcement (Herrnstein, 1974). This implies that animals are capable of estimating the rate at which rewards are available and can time their responses to coincide with reward delivery. One method of developing a suitable response timing schedule is to learn about both actions and the latencies with which to perform them. This provides a simple mechanism that balances the cost of responding against the estimated benefits of reward. Niv et al. (2007) demonstrated that a TDRL model augmented with the capacity to estimate the average rate of reward matches free-operant behaviour well. Furthermore, this model could predict the behaviour of DA depleted animals by parametrically reducing the model's average reward rate. Unfortunately, little evidence has come to bear on this idea, leaving the computational role of tonic dopamine largely a mystery.

The DA system's tonic component has not received the same attention from either an experimental or theoretical perspective. Tonic DA has a clear impact on engagement and response vigour, but the precise mechanisms through which this occurs are not well understood. Complicating matters is the fact that tonic and phasic components likely interact, meaning that one cannot be fully understood in isolation of the other. Despite these uncertainties, DA is one of the better understood neurotransmitters that plays an indisputable role in learning. I turn now to a discussion focused on the basal ganglia, one of the best studied and primary targets of the DA system, with an eye toward the nature of the information encoded there.

1.2.2 The basal ganglia: the root of exploitive action selection

The basal ganglia (BG) are comprised of an anatomically and functionally linked group of subcortical nuclei located at the base of the forebrain (Mink, 1996). The BG are thought to facilitate a wide range of faculties, including motor, motivational, and cognitive processes. Such breadth could be interpreted as an indication that multiple disparate roles are subsumed by the BG; however, a more unified and integrative functional role has been suggested by recent model-based investigations of the BG.

Broadly speaking, the BGs functional role can been conceptualized as a system that dynamically and adaptively gates the flow of information among cortical regions via cortico-basal ganglia-thalamo-cortical loops. Anatomically, the BG is well suited for this role, serving as a central way-station where projections from numerous cortical structures converge, including prefrontal cortex, sensory cortex, hippocampus and the amygdala. Hence, the BG is positioned to integrate information from mul-

tiple systems, including candidate motor or cognitive actions represented in frontal cortex, and to gate the most adaptive of these actions to be carried out while suppressing the execution of competing actions.

But how do the BG 'know which actions should be facilitated and which should be suppressed? Any plausible model of such functionality should avoid appealing to what amounts to a homunculus in the BG, demanding a priori knowledge and pulling the levers as necessary. In the following I provide a high level overview of a biologically constrained model of the BG. This model has accounted for a variety of seemingly disparate behaviours, and has led to several novel predictions that have been confirmed via pharmacological, diseases, neuroimaging, and genetic studies. By encapsulating the BGs core structure and activation dynamics, the model explicitly specifies the mechanisms though which the BG learns to integrate and gate information.

The basal ganglia's circuitry

The core computational processes ascribed to the BG have been extended to model cognitive functions such as working memory (Hazy et al., 2006), the role of instructions (Doll et al., 2011), the influence of choice on valuation (Cockburn et al., 2014), and the complementary roles of orbitofrontal representations of reward value and their influence on the BG (Frank & Claus, 2006).

As illustrated in Figure 1.2, the BG model consists of several layers which comprise the core structures associated with the BG: the striatum; globus pallidus (internal and external segments, GPi and GPe), substantia nigra pars compacta (SNc); thalamus; and subthalamic nucleus (STN). When a stimulus is represented in sensory

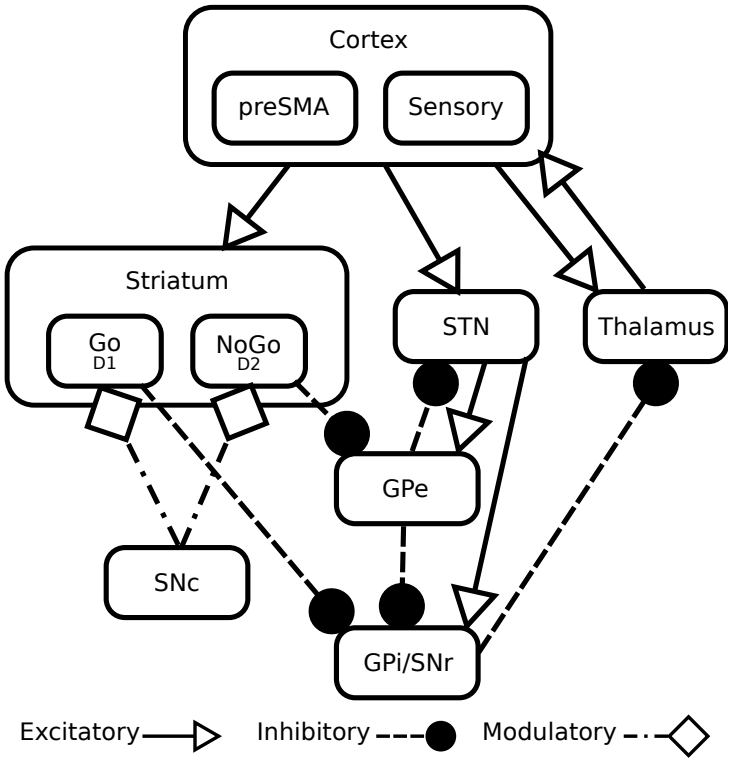

Figure 1.2: Functional anatomy of the BG circuit

cortex, candidate actions are generated in premotor cortex (preSMA), and output from both of these regions converge at the striatum. Columns of striatal units then integrate and encode information related to candidate actions in the context of the sensory stimulus. Information is transformed through the BG's circuitry such that thalamic output will bias action representations in the preSMA toward the candidate action that has delivered the best outcomes in the past. To simplify exposition, the model depicted here includes cortico-BG loops involved in motor action selection; however, the same basic mechanisms can also be applied to frontocortico-BG loops involved in biasing cognitive action selection to adaptively gate information into working memory (Hazy et al., 2006; Doll et al., 2009).

The BG's core gating function emerges as activity from 'direct and 'indirect pathways converges upon the thalamus, resulting in either the facilitation or suppression of cortical activations due to recurrent connectivity between thalamus and

cortex. The direct pathway originates in striatal D1-receptor expressing units that provide focused inhibitory input to the GPi. The GPi is tonically active in the absence of synaptic input and sends inhibitory projections to the thalamus, preventing the thalamus from facilitating any cortical response representations. When striatal units in the direct pathway fire, they inhibit the GPi, and remove tonic inhibition of the thalamus. This disinhibition process allows representation of the candidate action in question to excite corresponding column in the thalamus, which reciprocally amplifies the cortical activity via recurrent thalamo-cortical activity. Once a given action is amplified by striatal disinhibition of the thalamus, the competing candidate actions are immediately suppressed due to lateral inhibition in cortex. Thus we refer to activity in the direct pathway as 'Go signals which facilitate the selection of particular cortical actions.

The 'indirect pathway, so labelled due to its additional synapse passing through the GPe, provides an oppositional force to the direct pathway. Originating in striatal D2-receptor expressing units, the indirect pathway provides focused inhibitory input to the GPe, which in the absence of input is also tonically active and sends focused inhibitory input to the GPi. We refer to activity in this pathway as 'NoGo signals since these striatal columns inhibit GPe, which transitively disinhibit their respective columns in the GPi further, ultimately preventing the flow of thalamo-cortical activity. Together, the balance between activity in the Go and NoGo pathways for each action determines the probability that the action in question is gated.

Learning in the the basal ganglia

The relationship between DA and learning is perhaps best characterized in the BG. Evidence from patient studies and pharmacological manipulations that primarily af-

fect and target the BG (Pavese et al., 2006; Sawamoto et al., 2008) suggest that dopaminergic processes in the BG are involved in reinforcement learning and action selection (Frank et al., 2004; Cools et al., 2006; Palminteri et al., 2009). These human studies are complemented by optogenetic manipulations, voltammetry, synaptic plasticity, and single cell recordings in non-human animal studies (Reynolds et al., 2001; Roesch et al., 2007; Samejima et al., 2005; Shen et al., 2008; Tsai et al., 2009) demonstrating that synaptic plasticity in the striatum depends on phasic DA signals.

DA plays a critical role in the BG model, allowing it to learn which cortical representation should be facilitated and which should be suppressed. DA's impact depends on both the receptor type and the state of the target unit. For units in the Go pathway, DA has a D1 receptor-mediated contrast-enhancing effect by further increasing activity in highly active units while simultaneously decreasing activity in less active units. Thus only those Go units which have the strongest activity due to strong weights associated with the stimulus-action conjunction in question are amplified. For units in the NoGo pathway, DA has a D2 receptor-mediated inhibitory effects, such that greater levels of dopamine inhibit NoGo units, whereas these units are more excitable when DA levels are low. The net result is that increased DA levels facilitate Go activity while suppressing NoGo activity. Conversely, by reducing the inhibition of D2-expressing units, decreased DA levels increase the activity of NoGo units while simultaneously reducing activity in Go units. Hence, variation in DA levels has a differential effect on Go and NoGo signals projected by striatum. In the context of this model, dopaminergic effects play a critical role in both modulating the overall propensity for responding (by modulating Go relative to NoGo activity), and, by modulating activity-dependent plasticity, also allowing the model to learn which representations should be facilitated and which should be suppressed.

In the model, phasic DA bursts are simulated by the SNc layer to encode re-

ward prediction errors. Following a phasic SNc burst, activity increases in Go units associated with the action selected in the current stimulus context, while activity in other Go units and NoGo units declines. Driven by activity-dependent Hebbian learning principles, the weights between representative Go units and active cortical units (sensory and motor) are strengthened, while all other weights are weakened. Thus, Go learning to choose a particular action in the context of a given stimuli is supported by phasic dopamine bursts.

Conversely, dopamine dips are simulated by the SNc layer to encode negative reward prediction errors. D2-expressing striatal units in the NoGo pathway, which are typically inhibited by baseline dopamine activity, are disinhibited following a phasic SNc dip, thereby increasing NoGo unit activity driven by excitatory cortical input. This increased activity strengthens the weights between active cortical and NoGo units. Hence, phasic DA dips support learning to avoid certain actions given a particular stimulus.

In summary, the model presented here encapsulates several of the BGs key structures and the activation dynamics among them. Direct Go and indirect NoGo pathways, differentially modulated by dopamine, allow the model to learn not only what it should do, but what it should not do given a particular stimulus context. Alongside a wealth of neurological and behavioural evidence, and supported by several empirically confirmed model-based predictions, the model presented here links processes within the BG to action selection, reinforcement learning, and the development of direct inter-cortical connections.

The basal ganglia drive exploitive action selection

Several lines of evidence suggest that the BG forms the hub of a network aimed at maximize reward. Early investigations into this model's veridicality demonstrated that dopaminergic modulation in Parkinson's (PD) patients predicted their ability to exploit reward information (Frank et al., 2004). Unmedicated PD patients suffer low tonic levels of DA. The model predicts that this will result in a bias to learn from phasic DA dips more effectively than phasic bursts due to tonic disinhibition of D2-expressing cells in the NoGo pathway. Conversely, DA medications commonly used to treat PD patients can overdose regions of the BG. The model predicted that this will result in a bias to learn from phasic DA bursts more effectively than phasic dips due to tonic excitation of D1-expressing cells in the Go pathway. These predictions were confirmed behaviourally by testing PD patients on a task aimed at quantifying a subject's ability exploit reward information to choose rewarding actions and avoid un-rewarding actions (Frank et al., 2004). Medicated PD patients exhibited an increased capacity to select rewarding actions, but impaired ability to avoid non-rewarding actions. Conversely, those same patients while off medication displayed an increased capacity to avoid un-rewarding actions, but an impaired capacity to choose rewarding actions.

Subsequent studies investigating genetic contributions to reinforcement learning in the BG further support the hypothesis that the BG guides exploitive action selection. A functional polymorphism of the Darpp-32 gene, which is thought to influence D1-driven plasticity in the Go pathway (Meyer-Lindenberg et al., 2007; Stipanovich et al., 2008), has repeatedly been shown to predict individual differences in pursuit of rewards, but not exploration. Additionally, the DRD2 gene, which is thought to influence D2-dependent plasticity in the NoGo pathway (Hirvonen et al., 2009),

has repeatedly been show to predict individual differences in avoiding undesirable outcomes, but not exploration (Doll et al., 2011; Frank et al., 2007, 2009; Cockburn et al., 2014). In conjunction with these finding, numerous imaging studies have found that striatal activity is involved exploitive learning and action selection (Delgado & Nystrom, 2000; McClure et al., 2003a; O'Doherty et al., 2004; Daw et al., 2006). Noting the correlative relationship between BOLD signal and behaviour, (Jocham et al., 2011) showed that striatal prediction error related BOLD signal change was modulated by dopaminergic drug (low-dose D2-receptor antagonist), and was predictive of learning from positive reward feedback.

Considerable evidence suggests that the BG act as a selective information gate that bias action selection mechanisms toward more rewarding choices. This can be framed in a RL framework whereby the activity patterns in striatum reflect Q-values. An elegant study reported by Samejima et al. (2005) showed that striatal neurons encode the expected value of actions available to the monkey, suggesting that striatal cells are indeed encoding something like Q-values. The globus pallidus then provides a mechanism akin to an expected value action selection mechanism, where candidate Q-values compete amongst themselves for control over output via the thalamus. Of course, BG output is only one among many possible influences guiding action selection. Evidence suggests that the BG is particularly engaged in exploitive action selection, but that other sub-structures are more involved in guiding exploratory choices (Daw et al., 2006).

1.2.3 Exploratory behaviour and its neural correlates

The formal models of learning and action selection outlined earlier provide a framework for understanding action selection in terms of reward maximization. This has

been used with great success to investigate exploitive behaviour and its underlying neural correlates. Unfortunately, due to a number of challenges a generalized formal framework has yet to be developed to the same standard for exploratory behaviour. One such challenge stems from the various levels of uncertainty outlined above. Each flavour of uncertainty may favour different uncertainty reduction sampling strategies, complicating analysis both in terms of predicting behaviour, and how outcome information will be integrated to affect future behaviour. In short, exploratory behaviour is guided by a complex feedback loop where observations are integrated according to the current sampling strategy, and the current sampling strategy depends upon how previous observations have been integrated.

We are only beginning to understand the relationship between the state signal, learning, and sampling strategies. Gureckis & Love (2009) have demonstrated that human decisions are strongly impacted by the characteristic of the environment's state signal. Human participants took part in a variant of the rising optimum task, where optimal performance required resisting the temptation of large immediate rewards in favour of small rewards that lead to larger delayed rewards. Following a choice, participants were shown one of three state signals that varied in their correspondence to the underlying state of the task. The details of the state cues can be ignored for our purposes here, sufficed to say, participants shown information consistent with the underlying task structure performed best by adopting a strategy that preferred large delayed rewards, whereas participants that were shown no state information adopted an impulsive sub-optimal reward seeking strategy. These results demonstrate that information encoded in the state signal plays a significant role in learning and action selection independent of the underlying environment structure.

In line with these results, Acuña & Schrater (2010) have shown that human performance may well be sub-optimal with respect to an ideal actor with full task

knowledge, but this is due to the fact that the problem faced by task participants is much more difficult since they must cope with both structural and estimation uncertainty. This suggests that the nature of the information encoded in the state signal, specifically the degree to which it reduces various forms of uncertainty, will play a significant role in action selection. To understand this better, I will now outline evidence suggesting that exploratory human behaviour does indeed conform to uncertainty reduction models of action selection.

Evidence of uncertainty sampling behaviour

The approach outlined by King et al. (2004) falls within a class of active learning techniques called uncertainty sampling where, at a coarse level, the approach is to quantify and exploit some measure of uncertainty to guide action selection. The obvious question raised by the success of uncertainty sampling in the engineering domain is whether the brain uses a similar strategy. Regrettably, this question has gone relatively unstudied. However, improved techniques for modelling uncertainty and exploration have begun to provide much needed traction on the topic. One such investigation focused on active category learning. Participants were asked to learn the category structure of visual stimuli varying across two dimensions using either passive or active sampling (Markant & Gureckis, 2014). The active group could freely choose the samples they learned about, whereas the passive learning group was shown samples yoked to the active group. Behavioural results demonstrated that the active group learned more effectively. Furthermore, it was found that the active learning group sampled closer to the true category boundary as they learned the task. Sampling behaviour was investigated in more detail using a stochastic boundary learning model. The model was able to explain several aspects of the behavioural discrepancies between active and passive learning groups through an

uncertainty sampling mechanism whereby active learners selected samples close to their current hypothesized category boundary, leading to more efficient hypothesis updating.

More closely related to the issues at hand here, Payzan-LeNestour & Bossaerts (2011) have reported a detailed analysis of human choice behaviour during a complex task aimed investigating the effects of expected, unexpected and estimation uncertainty. Participants were presented with a volatile 6-armed bandit task. Arms were colour coded, three red and three blue, with red arms delivering rewards of larger magnitude (-2, 0, 2) than blue arms (-1, 0, 1). The outcome probabilities among arms was independent; however, outcome probabilities within a colour group changed simultaneously, with red arms changing more frequently (higher unexpected uncertainty) than blue arms. Behaviour was found to to be most accurately predicted by a Bayesian model that incorporated unexpected, expected and estimation uncertainty, suggesting that human participants were sensitive to these forms of uncertainty as well. However, counter to much of the discussion here, they found that participants exhibited an aversion to uncertainty.

Unfortunately, the authors offer no explanation as to why uncertainty aversion might arise. I would argue that gap between the task's underlying structure and the state signals available to the participant play a significant role in determining the direction of uncertainty's influence. Regrettably, the optimal task strategy is not discussed, leaving the possibility that uncertainty aversion might actually be the optimal strategy. To clarify, red arms are highly volatile, and as outlined by the authors, this makes it more difficult to develop a good estimate of their underlying reward contingencies relative to blue arms. Should participants come to learn that blue arms are more stable, and data shows they do learn this, it's possible that the best option might be to rapidly learn and exploit the blue arms despite the fact

that they deliver small rewards (and smaller penalties). Continually sampling the red arms could be disadvantageous because reward contingencies may switch before learned values can be exploited. This could be captured more formally by computing the Gittins index using different horizons for each arm colour (i.e. $N_b > N_r$ for blue and red arms respectively). Accordingly, I suggests that humans may seek or avoid uncertainty depending on the task's structure and dynamics. This, I think, could provide an interesting avenue of investigation that would dovetail nicely with the question of how we choose actions and integrate outcome information under uncertainty.

Neural correlates of uncertainty sampling

A recent series of studies have examined the impact uncertainty has on behaviour and the neural mechanisms involved (Frank et al., 2007, 2009; Cavanagh et al., 2011; Badre et al., 2012). These studies have all demonstrated that exploratory behaviour can be predicted in part by the relative uncertainty among available choices. Interestingly, the extent to which individuals employ an uncertainty sampling approach was shown to vary considerably among study participants. These differences were found to be predicted by a polymorphism in the COMT gene Frank et al. (2007, 2009), an enzyme known to to breakdown DA in prefrontal cortex (PFC) (Meyer-Lindenberg et al., 2005). Previous work has argued that higher concentrations of DA in the PFC support more robust maintenance of activation patterns (Seamans & Yang, 2004). Thus, individuals with higher DA should do a better job tracking uncertainty associated with various actions, and are therefore better able to direct exploration toward relatively uncertain actions. Conversely, individuals with low DA have relatively unstable representations of uncertainty, resulting in less uncertainty driven exploration. However, poor uncertainty sampling could emerge as the result

of inaccurate representation leading uncertainty sampling astray, or because individuals adopt a sampling strategy more suited to their strengths. Exactly which is unclear.

Subsequently, an fMRI study using the same task found that activity in frontopolar cortex (FPC) correlated with the relative uncertainty among candidate actions (Badre et al., 2012). Furthermore, this correlation was strongest in individuals that exhibited more uncertainty-driven exploratory behaviour. An EEG version of the same task corroborates these results. Trial-to-trial variation in lateral mid-frontal theta-band activity, which has been implicated in RL and adaptive control, were linearly related to the relative uncertainty among actions (Cavanagh et al., 2011). These results provide converging evidence to suggest that the brain is indeed tracking action uncertainty as a means of guiding exploratory behaviour, and that these processes rely upon DA-dependent plasticity in the PFC.

Previous studies have also found FPC-related activity to be associated with a more general notion of exploration. A restless 4-armed bandit task was employed to investigate the cortical substrates involved in exploration (Daw et al., 2006). Here, exploration was defined according to an RL model's estimate of expected reward value. Exploitive actions were deemed to be actions in pursuit of the highest estimated reward, and all other actions were considered to be exploratory. Their results reveal that FPC activity corresponded with exploration, showing preferential activation during exploratory decisions. Conversely, they found regions within the basal ganglia and vmPFC that exhibited activity in accordance with exploitive action selection. This use of estimated value to define exploration is slightly different than the uncertainty approach discussed previously; however, the two methods would likely share similar predictions in a restless 4-armed bandit task. The relative uncertainty associated with all other arms will increase relative to the currently

sampled arm given that the value of all arms were drifting on each trial independent of choice. Thus, the choice of any arm other than the one with the highest current estimate, as Daw et al. (2006) quantify exploration, would also be considered exploration according to uncertainty measures. Although the results reported by Daw et al. (2006) demonstrate the FPC is involved in exploratory behaviour, the task employed does not definitively dissociate between various variables that might be guiding exploratory behaviour.

More recently, results reported by Boorman et al. (2009) suggest that FPC might be tracking evidence in support of alternative actions. Here, participants were asked to make choices in a volatile 2-armed bandit task. One arm would consistently deliver rewards more reliably than the other, but without warning, the reward reliability would flip. Critically, the task was designed carefully control and de-correlate the expected value and the probability of reward at each arm such that these two parameters could be investigated independently. They found that FPC continually accumulated evidence in favour of switching to the alternative action, whereas vmPFC encoded the relative value of the current decision. (Bunge & Wendelken, 2009) offered a slight reinterpretation of these results to suggest that FPC may be encoding the alternative options themselves, not just evidence in their favour. Together, these results suggest that frontal structures such as FPC are involved in maintaining and operating upon alternative hypotheses regarding the optimal course of action.

Other studies have also found FPC-related activity to be associated with alternative strategies. One such example found that FPC activity tracked state entropy in a virtual maze navigation task (Yoshida & Ishii, 2006). After some initial familiarization with a maze structure, participants were placed at random position within the maze, from which they were to find their way to a goal position. The participant

had to orient themselves within the virtual maze by exploring their surroundings, and leverage this information to move toward the goal position. A Bayesian model was applied to estimate two hidden cognitive states: the first being the participant's belief about their current position in the maze (state belief), and the second reflected the participant's belief as to whether they should proceed with the current plan or back-track and start over (operant belief). In their model, state belief entropy was a function of the number of positions consistent with what had been observed, and was found to correlate with BOLD signal change in FPC. The authors hypothesize that FPC maintains beliefs about possible positions in the maze (states), and that increased activity reflects internal conflict among possible states. Conversely, they suggest that the mPFC, the ACC in particular, is involved in monitoring ongoing events to guide the current action selection strategy (Yoshida & Ishii, 2006).

More generally, the FPC has been argued to be the apex of a hierarchical organization in frontal cortex supporting executive control (Koechlin et al., 2003; Badre & D'Esposito, 2007; Badre, 2008). It has been hypothesized that as one moves along a caudal/rostral axis, regions of PFC are involved in increasingly abstract levels of information processing and behavioural control, across either temporal contingencies (Koechlin et al., 2003), or representational structure (Badre & D'Esposito, 2007). This has been summarized nicely using concepts from information theory, where the demand for executive control can be quantified as the amount of information required for action selection (Koechlin, 2007). The amount of information $H(a)$ needed to select action a given stimulus s is the the sum of the information conveyed by s involved in choosing action a, (i.e. the mutual information between s and a, $I(s, a)$), and the remaining information needed to choose a independent of s (i.e. the conditional information, $Q(a|s)$). Conditional information may be decomposed further to incorporate increasingly abstract information if necessary.

For example, $Q(a|s) = I(c, a|s) + Q(a|s, c)$ decomposes $Q(a|s)$ into the mutual information shared between the action a and a contextual signal c given the current stimulus ($I(c, a|s)$), and any remaining conditional information ($Q(a|s, c)$). As the informational demands required for action selection increase, so does the demand for executive control. Although there is some debate regarding the veridicality and representational nature of this hierarchy (Badre, 2008), an influential hypothesis suggests that the FPC facilitates maintenance and operation upon the highest level of abstraction, tracking multiple concurrent task sets relevant to the current control of behaviour (Koechlin, 2007). Should this be the case, FPC could be involved in guiding exploration based on information pertaining to alternative strategies, whether this be with respect to the relative uncertainty among candidate actions (Badre et al., 2012), evidence in support of an alternative strategy (Boorman et al., 2009), or belief state entropy (Yoshida & Ishii, 2006). As such, the FPC will play a critical role directing action selection strategies under state uncertainty, though it's precise computational role is yet to be defined.

1.3 Balancing exploration and exploitation

As discussed previously, optimal performance on a n-armed bandit task depends upon actions that emphasize reward and actions that stress uncertainty reduction (Gittins & Jones, 1979). Our discussion leading up to this point has outlined evidence suggesting that we might, in a broad sense, consider the BG to be the nucleus of exploitive learning and action selection (Daw et al., 2006), while the FPC is involved in tracking uncertainty (Badre et al., 2012). Given that these neural systems normally operate in parallel, their outputs cannot be expressed simultaneously should they prefer different actions. There are two possible means through which the brain may

arbitrate between the two systems. First, dynamic competition among systems may govern behaviour whereby the relative activation of each system determines which system will guide action execution. A solution similar to this has been hypothesized to determine preference for immediate versus delayed rewards (McClure et al., 2004). However, a competitive solution such as this does not easily support rapid goal directed strategy adjustments on its own. A second solution would be to have a third party decide which system will be given control by either biasing or blending the output of the exploitive and exploratory systems. In line with this hypothesis, I will now outline evidence suggesting that the locus coeruleus (LC) norepinephrine system (NE) and the anterior cingulate cortex (ACC) play critical roles in adaptively mediating between exploratory and exploitive systems.

1.3.1 The locus coeruleus norepinephrine system

The locus coeruleus (LC) norepinephrine (NE) system, has received somewhat less attention than the midbrain dopamine system. It has nonetheless been recognized as a critical component in the neural circuitry involved in learning and action selection. The LC is a small bundle of norepinephrine (NE)-containing neurons located deep within the brain in the pontine region of the brainstem (Berridge & Waterhouse, 2003). Although the LC consists of a relatively small number of neurons (10,000 to 15,000 per hemisphere), it projects to virtually the entire brain and provides the primary source of NE to target structures, with the notable exception of the basal ganglia which is nearly devoid of NE input (Berridge & Waterhouse, 2003).

Like DA, NE is best described as a neuromodulator thought to play a role in modulating the activation function of target neurons (Servan-Schreiber et al., 1990). For example, simultaneous application of NE and glutamate to Purkinje neurons had

a strong excitatory effect relative to glutamate applied alone. Similarly, simultaneous application of NE and GABA had a strong inhibitory effect relative to GABA applied in isolation. This enhancing effect has been noted to be similar across cortical regions (Foote & Morrison, 1987).

Early theories posited NE's primary role as driving arousal levels. Experiments demonstrated that a reduction in NE lead to drowsy and non-alert behaviour, resulting in poor task performance due to task disengagement. Conversely, an increase in NE produced highly distractible behaviour, resulting in poor task performance due to an inability to stay on task (Aston-Jones et al., 1999). These data points were taken as evidence toward the hypothesis that, at a coarse level, NE appears to play a critical role in engagement and attentional processes (Aston-Jones et al., 1999; Berridge & Waterhouse, 2003). However, the arousal theory of NE did not address the fine-grained dynamics of the NE-LC system, nor did it provide a mechanism through which NE levels may be adaptively adjusted.

To address these issues, recent investigations into NE's functional role have focused on evidence suggesting that, like the midbrain DA system, the LC exhibits tonic and phasic modes of activity. Tonic activity consists of firing rates between 0 and 5 Hz, whereas tonic activity is characterized by bursts of rapid firing, up to 20 hz, followed by sudden period of suppressed firing lasting from 200-500 ms due to self-inhibitory auto-receptors (Berridge & Waterhouse, 2003). Typically, phasic activity is observed 100-150 ms following onset of task relevant but not distractor stimuli. Critically, phasic activity is only observed when tonic activity is relatively low and the animal is engaged in the task.

The adaptive gain and optimal performance theory of LC-NE activity posits that phasic and tonic modes of activity play a role in balancing the explore/exploit trade-

off (Aston-Jones & Cohen, 2005). This theory build upon work suggesting that the phasic component of LC activity is a temporal attention filter that improves signal detection throughout cortex when motivationally salient events are detected (Usher, 1999). Phasic signals selectively boost relatively active neurons while dampening others, increasing the signal-to-noise ratio in the system (Servan-Schreiber et al., 1990). From a network dynamics perspective, this encourages a high-energy network to settle into a stable attractor basin, which is is a particularly desirable effect in large multilayered networks as we assume the brain to be. This provides a mechanism to adaptively balance a tradeoff between flexibly sluggish computation and rigidly efficient processing (Usher, 1999).

As long as performance is good and outcome expectations are being met, precisely timed phasic LC activity will capture task relevant events and selectively weight that information for processing while concomitant low tonic activity prevents irrelevant information from influencing the system. Together, this relationship between tonic and phasic LC-NE components ensures that pertinent events can be acted upon as efficiently as possible without wasting computational resources on irrelevant phenomena. However, if performance starts to decline and task utility begins to wane, it may be in the agent's best interest to find a new strategy. To facilitate a more exploratory strategy, phasic LC activity is reduced to minimized the emphasis placed upon information that no longer seems to have much utility. Simultaneously, tonic activity is increased to broaden the spread of attention such that new information may be captured and exploited to improve performance. Thus, the adaptive gain and optimal performance theory postulates that NE provides a mechanism through which information can be weighted according to its utility. An exploratory strategy is emphasized when the phasic component is relatively absent but the tonic component is engaged. Conversely, an exploitive strategy is stressed when the tonic

component is relatively inactive but the phasic component is active (Aston-Jones & Cohen, 2005).

Recent attempts to understand the functional role of LC activity have addressed the issue within a normative framework aimed at optimizing the tradeoff between exploration and exploitation in a volatile environment. As discussed previously, adaptive performance in non-stationary environments demands consideration of both expected and unexpected uncertainty. Yu & Dayan (2005) offer evidence suggesting that tonic LC activity may be tracking unexpected uncertainty, with tonic activity increasing in accordance with the agent's belief that the environment has undergone an unpredicted change and a new policy should be adopted. This fits well with the idea that increased tonic LC activity favours an exploratory strategy, and provides an important link between tonic LC activity and normative statistics. This work has been extended and applied to the phasic LC component as well, which is argued to respond to rare events within a stable environments (Dayan & Yu, 2006)[1]. The basic idea being that transitions within a stable environment can nonetheless be unpredictable. An efficient strategy in a relatively predictable world is to take advantage of environmental regularities and prepare actions in anticipation of the most likely circumstance. Adopting a default action policy akin to habitual responding avoids unnecessary cognitive effort, which is not only more efficient with respect to resource usage but can also be faster and more accurate than are deliberative strategy. Dayan & Yu (2006) suggest that a phasic NE burst signals a rare event, which interrupts on going default mode processing and allows additional resource to be brought online to make necessary adjustments.

In summary, there appears to be a general consensus that the LC-NE system's

[1]Dayan & Yu (2006) use the term unexpected uncertainty in reference to volatility both within and between tasks. I will refer to unexpected uncertainty within a task as a 'rare' event for clarity.

role in adaptive control is to balance efficiency and flexibility in support of the explore/exploit tradeoff (Aston-Jones & Cohen, 2005), with some support showing that it may do so according to estimates of unexpected uncertainty (Yu & Dayan, 2005). Higher levels of NE increase the system's signal-to-noise ratio, facilitating rapid but restricted information processing while events are unfolding as expected. This is supported at a coarse level by tonic LC activity, which mediates an adaptive balance between exploration and exploitation at the task level in the minute-to-minute timescale. This same basic machinery is at work on a finer timescale within a task's dynamics. Phasic NE signals operating at the millisecond timescale support an adaptive tradeoff between exploiting prepared action plans, or, when events don't unfold as expected, abandoning this plan to find a more appropriate alternative. Noting the apparent lack of NE input to the BG (Berridge & Waterhouse, 2003), NE levels could possibly play a role by modulating the processing efficacy of the exploratory system relative to the exploitive system (i.e. the BG) as a means of controlling the action selection strategy. This suggests that the LC-NE system responds to increasing uncertainty by biasing action selection strategies toward uncertainty reduction.

1.3.2 The anterior cingulate cortex

The anterior cingulate cortex (ACC), and more generally the medial prefrontal cortex (mPFC), is thought to play a critical role in executive control. ACC activation has been noted in tasks involving learning and memory, language, perception, and motor control, suggesting that the ACC serves a general cognitive function. A unifying hypothesis positions the ACC as a performance monitor that signals when events are not going as well or as smoothly as expected such that additional adaptive control processes can be brought on line (Ridderinkhof et al., 2004). Consequently, action

selection under state uncertainty will likely engage the ACC.

Anatomically, the ACC positioned to integrate information pertaining to rewards and action selection, forming reciprocal loops with numerous cortical, sensorimotor, and subcortical structures. This includes the orbitofrontal cortex, which is thought encode reward value (Kringelbach, 2005), the amygdala, which is thought to play a role in emotion (Ledoux, 2007), the basal ganglia, which guides action selection (Mink, 1996), and the dorsolateral prefrontal cortex (dlPFC), which is though to play a role in working memory (MacDonald, 2000). The ACC also receives input from the midbrain DA system, and is reciprocally connected to the LC-NE system (Aston-Jones & Cohen, 2005).

Despite the attention garnered by the ACC as the hub of executive control, the nature of its function remains in dispute. The ACC has been theorized to be responsive to response conflict (Botvinick et al., 1999), reward prediction errors (Holroyd & Coles, 2002), error likelihood (Brown & Braver, 2005), environmental volatility (Behrens & Woolrich, 2007), the cost of effort (Hillman & Bilkey, 2010), and action outcome violations (Alexander & Brown, 2011). In the following I will describe a few of the more influential theories of ACC function.

The ACC as a conflict monitor

ACC activity increases when prepotent responses must be overridden, when the appropriate response is underdetermined, and when errors are made. Botvinick et al. (2001) encircled these seemingly disparate phenomena and reinterpreted them as response conflict. For example, during the Stroop task participants are asked to either read colour names or name the text's colour. ACC activity has repeatedly and

reliably been shown to be maximal when participants must override the prepotent response (word reading) and name the text colour when text colour and word are inconsistent (Carter et al., 2000). The same region has also been shown to be sensitive to competition among possible responses during flanker tasks, again, suggesting that competition among possible responses selectively engages the ACC (Botvinick et al., 1999).

The conflict monitoring theory proposed that the ACC is a response conflict monitor, a system that tracks the co-activation of mutually exclusive responses, as an indication that additional cognitive resources may be required to help resolve the increasing conflict and improve performance on the task at hand. A variety of methodologies have demonstrated that ACC activity conforms to the predicted behaviours of a conflict monitor. Imaging studies have repeatedly shown increased ACC activation in accordance with increased conflict (see (Ridderinkhof et al., 2004) for review). The error-related negativity (ERN), a component of the event-related potential (ERP) localized to the ACC (Debener et al., 2005), has also been shown to be modulated in agreement with conflict monitoring proximal to the time of response (Yeung et al., 2004), and feedback (Cockburn & Frank, 2011).

The ACC as a volatility monitor

More recent work investigating how the brain copes with environmental instability found that the ACC tracked volatility (Behrens & Woolrich, 2007). Participants were asked to play a restless 2-armed bandit task structured such that one arm delivered rewards more reliably than the other, with probability of reward for arms 1 and 2 being p and $1-p$ respectively. The task dynamics were such that subjects experienced periods of both stable and volatile reward contingencies. During the stable phase of

the experiment, one arm would deliver rewards with 75% reliability (and the other with 25% reliability). During the volatile phase, reward probabilities were flipped between the arms every 30 or 40 trials. Throughout the experiment, rewards for the first arm (r_1) were selected randomly between 0 and 100, and rewards for the second arm were set to $(100 - r_1)$, facilitating dissociations between the magnitude, probability and expected value of reward.

Optimal behaviour demands that previous outcomes be integrated appropriately to guide behaviour on the current trial. This calls for consideration of both the expected value for each arm, and the volatility of the environment. Behrens & Woolrich (2007) defined optimal behaviour according to a Bayesian learner. Given the observation data y, the learner estimated reward probability, r, volatility in the environment, v, and how certain it was of its volatility estimate, k. In a stable environment, the learner came to develop confident reward probability estimates, and a correspondingly confident estimate of low environment volatility. When the learner was exposed to the unstable environment, its volatility estimate increased, as did its distrust in the consistency of the environment's volatility, k.

Estimated volatility had a strong influence on the learning rate; that is, the rate with which observations altered reward contingency predictions. In short, modulating the learning rate in correspondence with volatility allows the learner to adapt to new environments quickly, while also facilitating stable behaviour once the environment has settled, safeguarding from overly reactive policy changes in stable but stochastic environments. Imaging results revealed that ACC activity was modulated according to volatility as estimated by the Bayesian learner, with activity increasing in correspondence with volatility estimates. In light of this evidence, Behrens & Woolrich (2007) proposed that the ACC monitors environment volatility in order to modulate the learning rate, perhaps via output projections to the basal ganglia

(Kunishio & Haber, 1994).

The ACC as action-outcome predictor

A recent model-based approach has reinterpreted a wide range of ACC related phenomena and unified them by suggesting that the ACC is concerned with learning and predicting action outcomes independent of their valance (Alexander & Brown, 2011). The theory proposed that cells within the ACC learn to predict both the timing and possible outcomes of a given action. When outcomes are observed, cells are inhibited in correspondence with how well they predicted what actually occurred. Thus, ACC activity is argued to report both unexpected occurrences (positive surprise) and non-occurrences (negative surprise) of predicted outcomes, regardless of outcome valence. This surprise signal is then used as a teaching signal to improve action outcome predictions.

The authors demonstrate that their model can reproduce effects of conflict in a simulations of the change signal and flanker tasks. Instead of ACC activity being driven by the co-activation of mutually exclusive responses *per se*, conflict effects are generated due to a growing outcome predictions for multiple responses. In the case of the flanker task, the prediction of a correct response is generally high given the simplicity of the task; however, incongruent flaker stimuli are able to influence the system and increase the likelihood of an incorrect response. The summed result is an increase in ACC activity in correspondence with the increased strength of outcome predictions and a lack of outcome driven suppression. This accounts for a range of findings and demonstrates a sensitivity to error likelihood that has been largely unaddressed by other models.

The model also provides a more biologically rooted reinterpretation of the volatility effects reported by Behrens & Woolrich (2007). As reward contingencies change, existing predictions persist while new predictions begin to form. As reversals occur predictions based on previously established associations that are no longer appropriate lead to an increase in activity when those predictions are upset. This includes both negative surprise from the non-occurrence of predicted rewards, and positive surprise from the occurrence of unpredicted rewards. The summed effect of this surprise signal increases the effective learning rate, providing an account for both the behavioural and imaging results described by Behrens & Woolrich (2007).

Summary of ACC function in relation to state uncertainty

Despite its theoretically appealing simplicity and explanatory power, the conflict monitoring theory of ACC function suffers from two shortcomings. First, although conflict can be formally defined either as entropy or Hopfield energy, the concept could be argued to be too general. It is often unclear precisely how conflict might be playing a role in tasks outside the set of tasks designed to probe conflict explicitly. ACC activity has been noted in a wide range of tasks, and it is often unclear exactly how conflict may be playing a role. Second, despite the empirical support found from non-invasive techniques in human, results from monkey neurophysiological studies present inconsistent results, with relatively few clear demonstrations neural activity encoding conflict (Olson & Gettner, 2002; Ito et al., 2003).

The idea that the ACC is somehow tracking volatility in the environment in order to modulate the learning rate offers a potential refinement on the idea of conflict. Indeed, one could argue that conflict is elevated during periods of high volatility; however, the Bayesian quantification of volatility used to predict ACC activity is

rooted in a normative framework geared toward optimal behaviour. Although models using conflict to detect errors (Yeung et al., 2004) and adjust performance (Botvinick et al., 2001) have been proposed, it's not clear how conflict should modulate learning. Conversely, (Behrens & Woolrich, 2007) specify an optimal means of adjusting the learning rate according to volatility estimates, which corresponds well with behaviour of human participants. That said, the restless 2-armed bandit task employed to investigate volatility cannot dissociate learning from task switching. That is, subject behaviour may exhibit rapid behaviour changes because they are learning, or because they have switched to an alternative pre-established action policy. As such, it's unclear whether the ACC is modulating the learning rate or facilitating a strategic adjustment.

The action-outcome prediction theory of ACC activity is particularly appealing with respect to understanding how actions might be selected under state uncertainty (Alexander & Brown, 2011). If the ACC is capable of influencing response behaviour, it could facilitate an action policy guided by outcome predictability as opposed to expected value alone. This may prove to be an adaptive strategy when the current state is uncertain, as I will outline below. I will not attempt a resolution of the ACC's computational role. Any of the previously discussed proposals would serve a beneficial role in detecting and coping with state uncertainty. An increase in conflict would be expected, as would an increased estimate of volatility, and likewise for action-outcome surprise. Generally speaking, all theories point in the same direction in that ACC activity would be expected to increase in line with state uncertainty.

1.3.3 Guiding action selection under state uncertainty

I have outlined the method of TDRL as a general approach to finding an action policy for an MDP (Sutton & Barto, 1998). The success of the TDRL approach depends upon the degree to which the environment adheres to the Markov property, which implies that the conditional probability distribution of future states and rewards depends only on the current state and action. When this property is violated TDRL performance can be severely compromised given its inability to leverage information beyond what is encoded in the current state signal (Littman, 1994).

Despite it's dependence on the Markov property, the TDRL framework has been shown to provide a powerful and efficient solution to action selection problems. This has led to suggestions that perhaps the brain implements a similar strategy. Recent investigations into brain function have found activity corresponding to TDRL signals and structures. The midbrain DA system is theorized to encode a reward prediction error (Montague et al., 1996; Schultz et al., 1997), used to train an action policy in the BG such that the BG will bias action selection toward the most rewarding actions (O'Doherty et al., 2004; Frank & Claus, 2006; Samejima et al., 2005; Hazy et al., 2006; Daw et al., 2006).

Unfortunately, due to noisy sensory systems, unobserved events, or temporal dependencies, many learning and optimal control problems are non-Markovian from the agent's perspective. The POMDP framework addresses this issue by explicitly decoupling the environment's state from observable information, acknowledging the gap between the true state of the world and what is accessible to the learning agent (Kaelbling et al., 1998). Several algorithms for fully solving a POMDP have been proposed; but unfortunately, finding an optimal policy becomes computationally intractable in large, and/or unstable environments.

That said, promising tractable approaches to coping with ambiguity have been developed. An adaptive gating memory mechanism can augment and disambiguate currently available information using previous observations. This effectively transforms a POMDP into an MDP which can be tackled by a standard TDRL approach (Hazy et al., 2006; Todd et al., 2009). A divide-and-conquer strategy may also be used to tackle ambiguity by learning multiple action policies that correspond to possible world states Doya et al. (2002). Here, regularities in the underlying (but unobservable) state structure are exploited in order to decompose the environment into a set of MDPs that can be solved by a standard TDRL algorithm. Ambiguous observations are resolved according to a model-based outcome prediction controller, which serves to select among candidate action policies and learn an appropriate decomposition of the task. A candidate action is nominated by each controller stochastically based on expected reward value, where each controller effectively represents a possible state of the environment given the current observation. One action is then stochastically selected for execution based on each controllers ability to predict action outcomes, preferentially weighting actions from the controller more likely to map onto the true state of the environment.

Relatively little is known as to whether the brain employs a strategy like multiple model-based reinforcement learning. However, recent studies suggests that people do indeed adopt such a strategy by learning task sets; that is, abstract representations that associate stimuli, actions and outcomes (Collins et al., 2014; Collins & Koechlin, 2012; Collins & Frank, 2013). These findings show that human choice behaviour can be accounted for by a combination of controlled and uncontrolled exploration, the former targeting task set selection and creation while the later stochastically selects actions within the active task set.

This collection of work has made significant headway toward our understanding

of the mechanisms and algorithms employed by the brain to facilitate learning and action selection. However, much of this work has used expected value exploitation as a means of investigation; but, evidence reviewed here indicates that uncertainty also plays a significant role in guiding action selection (Markant & Gureckis, 2014; Frank et al., 2007, 2009; Payzan-LeNestour & Bossaerts, 2011). I will now outline a novel task and computational model aimed at probing uncertainty-driven action selection in more detail

short proposal as to how uncertainty reduction may play a role in action selection in partially observable environments.

Chapter Two

Theory Sketch: Experimental design and computational model

2.1 Action selection targeting policy identification

It has been shown that learning may be expedited if sampling (exploration) targets uncertainty (Settles, 2009), and behaviour suggests humans are sensitive to uncertainty (either as a preference or an aversion) (Frank et al., 2009; Acuña & Schrater, 2010; Markant & Gureckis, 2014; Payzan-LeNestour & Bossaerts, 2011). However, the issue of interest here is how action selection should proceed once policies have been shaped in accordance with volatile and unobservable environmental states.

2.1.1 Learning a task decomposition

The approach I describe here builds upon the learning and action selection framework developed by Doya et al. (2002) and extended by Collins & Koechlin (2012). This framework addresses the problem of decomposing partially observable environments into task sets, learning action policies for those task sets, and identifying the most appropriate task set for the current state of the environment. The multiple model-based RL framework is structured such that a task will be decomposed at points of ambiguity where state uncertainty is highest. In the framework outlined by Collins & Koechlin (2012), hidden state volatility introduces periods of high uncertainty when no task set can reliably predict action outcomes. This spurs the creation of a new, though possibly temporary, task set to control learning and action selection. An existing task sets may come to reliably predict action outcomes as more samples are collected, at which point the temporary task set is discarded. However, in the case that no previously defined task is well suited to the current state of the environment, the newly created task set will gain a more permanent status as the expert for the current hidden state value.

Task set creation deserves a detailed discussion given the topic at hand. Creating a new task set under state uncertainty serves two critical functions. First, the new task set will act as the active task set, and since learning only occurs in the active task set, it will absorb learning signals that could otherwise contaminate established action policies tuned to other circumstances. Second, the new task set is endowed with a policy generated from a blend of policies in existing task sets. This supports policy generalization whereby knowledge of other circumstances may be transferred to similar situations. This implies that the model will adopt an exploration strategy that tends to prefer actions that have been rewarded in other similar situations, effectively allowing the knowledge encapsulated in established task sets to act as priors for the new task set's policy. I will return to this issue in more detail in the context of the experimental paradigm outlined next.

2.1.2 Policy identification

If we assume that some learning has already taken place, and the agent has developed a reasonable task set decomposition, then the challenge presented to the agent is not one of policy development, but rather, one of policy selection. If a policy has already been shaped to maximize reward expectation for the current (but unknown) state of the environment, then employing any other policy is sub-optimal. Accordingly, I propose that periods of high state uncertainty should prompt the adoption of an action selection strategy aimed at delivering state information. Doing so will help disambiguate the current state of the environment and facilitate rapid identification of the best known policy.

If we are to pursue the idea that state uncertainty encourages selection of informative actions, then we require some method of measuring an action's informativeness

with respect to a hidden state. One approach is to quantify the mutual information between action outcomes and the environment's state. The mutual information shared between two random variables, X and Y, indicates how reliably one variable predicts the other. In other words, mutual information provides an index of the information gained regarding Y when X is sampled (or vice versa) (Pierce, 2012; Settles, 2009). This is well suited to our problem here since the hidden state cannot be sampled directly; rather, its value must be inferred by sampling other observable variables that (hopefully) have some relationship with the hidden state.

Given an action a, with a set of outcomes, $O_a = \{o_1, o_2, ...o_n\}$, and a hidden state, S_h, mutual information between the two can be quantified as:

$$I(S_h, a) = \sum_{s \in S_h} \sum_{o \in O_a} p(s, o) \log \frac{p(s, o)}{p(s)p(o)} \tag{2.1}$$

If S_h and a are independent; that is, there is no relationship between the two, then $p(s, o) = p(s)p(o)$ and since $\log(1) = 0$, there is no mutual information. Conversely, if S_h predicts a perfectly, then mutual information is maximal and reflects only the uncertainty (entropy) in S_h (or O_a) itself. In our context, high mutual information implies that action outcomes are consistent for a particular hidden state value, and action outcomes differ reliably across possible state values. Conceptually, this idea bears some similarity with the information theoretic framework outlined by (Koechlin, 2007) to explain the hierarchical organization of frontal cortex. The authors note that the information required to select an action is the sum of the mutual and conditional information shared between the current state of the environment and an action, effectively quantifying what is necessary to explain a stimulus/response mapping. However, here we are concerned with the action/outcome mapping given a particular environmental state, searching for actions with the most diagnostic outcomes. I will now outline a simple experiment aimed at probing the role of action-

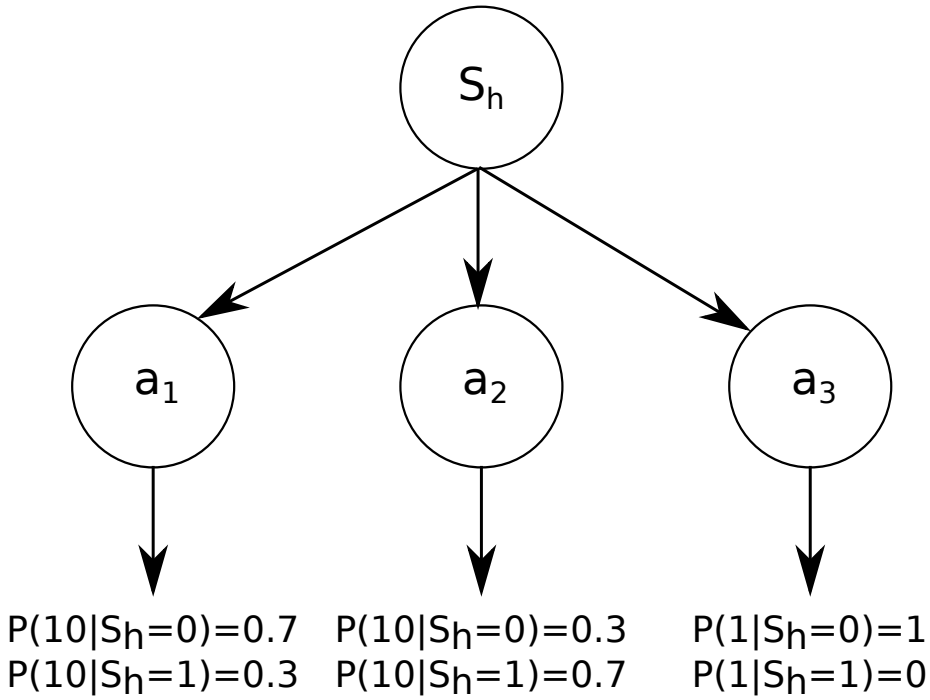

Figure 2.1: Proposed experiment. Hidden state S_h determines action reward probabilities

outcome informativeness to clarify these ideas further and provide an explanatory scaffolding for the modelling portion of the proposal.

2.2 Experimental sketch

Here I outline a simple experimental paradigm that will support a systematic investigation of the interaction between latent state uncertainty and action selection strategies. At the expense of additional structure that will be necessary to probe the issue in greater detail, I will outline a simple volatile 3-armed bandit task, as depicted in Figure 2.1. This provides the most basic structure capable of probing the impact of state uncertainty on behaviour. However, I wish to emphasize that this experimental outline is primarily for clarification purposes.

The core feature of the design is to offer options associated with varying amounts

of reward and/or information. In order to simultaneously investigate reward and information seeking behaviour, the outcome statistics associated with each option will be determined by a dichotomous latent variable that can switch from one value to another at any point during the experiment. The expected reward value of each option varies in accordance with the latent variable's value, and as such, so too does the optimal action selection policy given the latent variable's current value. In addition, the reliability of each option's outcomes vary across values of the latent variable. Thus, although the latent variable's value wasn't directly observable, outcomes can provide information pertaining to the latent variable's current value.

When the latent variable $S_h = 0$, the optimal policy should exploit a_1 since it delivers a larger reward more reliably than both a_2 and a_3. Conversely, when $S_h = 1$, the optimal policy should exploit a_2. If S_h were observable, then no policy should preferentially select a_3. In this case, a_3 is neither rewarding nor is it informative since S_h predicts a_3 perfectly. But, S_h is both unobservable and volatile, thus, despite a_3's negligible expected reward value, it shares the greatest mutual information with respect to S_h. Thus, according to the hypothesis put forth here, a_3 should be preferentially selected when state uncertainty is high.

2.2.1 The optimal observer

In the task just outlined, the value of the latent state variable must be inferred, and this inferred belief should shape response behaviour. To infer the latent state's value a Bayesian observer was used to incorporate action outcomes into belief regarding which deck was in play. In short, the Bayesian observer maintains the probability that the latent variable is in one of two states (0 or 1) given the task statistics and the outcomes that have been observed. Inference will shape a prior summarizing the

state belief before observing an outcome:

$$P(S_t)' = (1 - P(Sw)) * P(S_{t-1}) + (P(Sw)) * P(S_{t-1}) \tag{2.2}$$

where $P(Sw)$ is the probability that the latent variable will switch values between trials, and $P(S_{t-1})$ is the latent variable belief before starting the trial (i.e. the belief at the end of the previous trial).

Following the observation of each outcome Bayes' rule is used to update the belief that the latent variable is in a particular state (i.e. one of the decks is in play):

$$P(S_t|O_t) = \frac{P(O_t|S_t = 0)P(S_t)'}{P(O_t)} \tag{2.3}$$

where O_t is the outcome observed on trial t, and $P(S_t)'$ is the prior computed by Equation 2.2. In this way the optimal observer maintains the belief that the latent variable is in a particular state (here $S_t = 0$) by chaining the posterior belief from the previous trial, modulated by the probability of a latent variable switch, as the prior on route to updating it's belief after observing the current trial's outcome.

2.3 Models of action selection

Building upon the belief inference offered by the Bayesian observer, I define both a forward Bayesian actor, and a heuristic approximation. The forward Bayesian actor will compute a forward model to generate a utility index akin to the Gittins index, where an action's utility incorporates both the expected value of the action itself, and the expected reward value of optimal action selection at future trials. The heuristic approximation, on the other hand, will forgo the computational complexity

of a forward model in favor of combining an action's expected reward and expected uncertainty reduction. Following a description of each model I will contrast their performance on the task design previously discussed.

2.3.1 The forward Bayesian action utility model

We build upon the Bayesian observer to define the forward Bayesian action utility model. This model will provide us with a useful tool for investigating the task structure, and it will also provide a yardstick to help us understand participant behaviour. The forward Bayesian actor builds a forward model to estimate the rewards it can expect to accrue for each option available to it given it's current belief state. It does so by tracking the expected reward of taking each option available to it, recursively branching forward and updating its belief state according to the probability of observing all possible outcomes, and acting optimally for N steps into the future.

Somewhat more formally, we begin by computing the prior belief using Equation 2.2, and given that belief, computing the expected value for each option ($E[a_i|P(S_t)']$). Using Equation 2.3, we compute a posterior belief for all possible outcomes of each option, weighted by the probability of observing each outcome. This set of posterior beliefs is used to recursively compute the expected value of each option on the subsequent trial. The most rewarding option at depth N is rolled back and summed with the expected value of each option at depth $N-1$. Once value has rolled back to the root level of the recursion we have an estimate of the expected reward value for each option assuming the optimal action will be taken for N trials thereafter. Thus, the forward model integrates reward and latent state information into a singular index

value akin to the Gittins index[1]Gittins & Jones (1979).

As a short example in line with the general thrust of our experimental design, assume that the value of S_t is fairly stable (e.g. $p(Sw) = 0.05$) and that the actor has three options available to it: option 1 has the highest expected reward value when $S_t = 0$ (but low value when $S_t = 1$), option 2 has the highest expected reward value when $S_t = 1$ (but low value when $S_t = 0$), and option 3 is highly diagnostic of the value of S_t (but has low expected value overall). If the agent believes strongly that $S_t = 0$, then option 1 will have the highest expected value on both the current and subsequent trial. If, on the other hand, the agent strongly believes that $S_t = 1$, then option 2 will have the highest expected value on both the current and subsequent trial. However, if the agent is unsure of the state of S_t, option 3 will have the highest expected value. Although option 3 has a low expected value on the current trial, its outcome is diagnostic of the latent variable's state. This implies that state information can be exploited on the subsequent trial and will result in a more rewarding outcome than would otherwise be possible without option 3's hidden state information.

2.3.2 The heuristic uncertainty-reduction action utility model

The forward Bayesian model can computes the tradeoff between reward and latent state information. However, the computational tractability is questionable, and so to is the brain's capacity for complex numerically sensitive computations. As such, I aim to extend the work outlined in Collins & Frank (2013) using an informational heuristic to direct action selection toward informative actions when they may be

[1]The main difference being that the Gittins index is formulated under the assumption that the process has some probability of halting. We formulate the forward model's index by assuming that the process includes a latent variable that has some probability of switching

beneficial.

The multiple model-based framework described by Doya et al. (2002), offers a means decomposing a POMDP into tractable MPDs. This framework was extended to incorporate a mechanism that facilitated task-set creation, which allows the learning agent to dynamically decompose its environment according to the observed statistics (Collins & Koechlin, 2012; Collins & Frank, 2013). When state certainty is high and the current task set can reliably predict action outcomes, the standard stochastic model-free action selection strategy based on expected value can be used. However, should an unobserved state change occur, task set reliability will begin to decrease in the face of increasing prediction violations. As task set reliability dips below acceptable levels, a new task set will be created (Collins & Koechlin, 2012).

However, rather than adopt an action selection strategy generated from a blend of other task set policies that targets expected value, the aim here is to engage a selection strategy targeting actions that share mutual information with the hidden state in order to identify the optimal policy given the (currently unknown) state of the world. A good strategy should not exploit informativeness irrespective of costs associated with gaining that information. As noted previously, a prudent decision maker should never pay more than they stand to profit from acquiring new information (Howard, 1966). Thus, it will be important to introduce a tradeoff between information gain and expected value. An action utility function may take the following form:

$$U(a_i) = E[a_i] + \sigma \cdot I(S_h; O_i) \tag{2.4}$$

where $I(S_h; O_i)$ is the mutual information shared between hidden state S_h and action i's outcomes (O_i), and σ is the belief state uncertainty. $E[a_i]$ is the expected value

of action a_i across all task sets weighted by task set reliability:

$$E[a_i] = \sum_j \phi_j Q^{\pi_j}(a_i) \tag{2.5}$$

where ϕ_j is the reliability of task set j, and $Q^{\pi_j}(a_i)$ is the value of action a_i according to that task set's policy.

The relative weight given to information should be dynamically modulated according to performance demands. The importance of attaining state information increases in accordance with state uncertainty. Thus, if we assume that the value of σ is dynamically adjusted as an index of unexpected uncertainty as outlined by Yu & Dayan (2005), then moments of high uncertainty could bias action selection toward informative actions.

2.3.3 Model response patterns

I now explore various aspects of the experimental design and their influence on the response patterns of both the forward Bayesian model, and the heuristic uncertainty-reduction model by exposing both to an instantiation of the experimental design. The actor had the choice of three possible options on each trial. One option (hA: High-Reward Deck A) offered a reward of 50 points 40% of the time when Deck A was in play, but only 10% of the time when Deck B was in play (and zero points otherwise). A second option (hB: High-Reward Deck B) offered a reward of 50 points 40% of the time when Deck B was in play, but only 10% of the time when Deck A was in play (and zero points otherwise). The third option (Info) always offered a reward of 0.25 when Deck A was in play, and a reward of 0.75 when Deck B was in play. Thus, outcomes from this third card could be used to infer the current deck

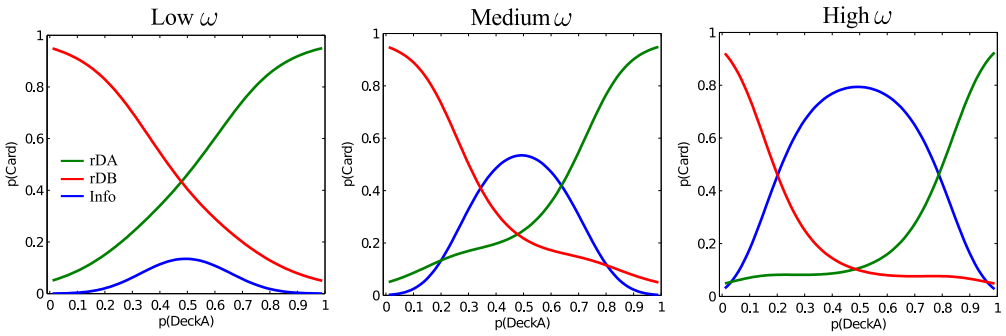

Figure 2.2: Influence of free parameter ω on action selection. As ω increases, the degree to which the uncertainty-reduction utility function will favor informative actions during periods of uncertainty increases, and as such, so to does the probability with which the model will seek information.

in play with certainty. The deck in play had a 5% chance of switching prior to the start of each trial.

Scaling the influence of uncertainty reduction

One motivation for developing a heuristic model of action selection was to facilitate an understanding of individual differences with respect to the valuation of information. In light of this, the uncertainty-reduction model specified in equation 2.6 can be extended to include a free parameter, ω that scales the degree to which expected information gain factors into the utility function:

$$U(a_i) = E[a_i] + \omega \cdot \sigma \cdot I(S_h; O_i) \tag{2.6}$$

Figure 2.2 depicts the effect of increasing ω. As ω increases, the degree of uncertainty necessary before informative actions are favored over rewarding actions is reduced.

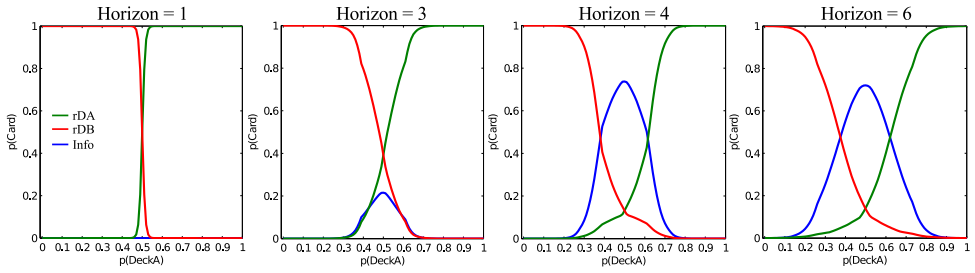

Figure 2.3: The influence of horizon on the optimal actor's action policy. As the forward model's horizon is extended further into the future (1, 3, 4, and 6 steps into the future), information is valued more. The actors policy assymptotes at a horizon of $N = 6$.

Sensitivity to horizon and volatility

The forward Bayesian model computes a utility index by unrolling the expected reward from a Bayesian N-step action look-ahead. To investigate the influence of the N-step horizon on response preferences, the forward Bayesian model was exposed to an instantiation of the experimental design. The Bayesian model had the choice of three possible options on each trial. One option (hA: High-Reward Deck A) offered a reward of 50 points 40% of the time when Deck A was in play, but only 10% of the time when Deck B was in play (and zero points otherwise). A second option (hB: High-Reward Deck B) offered a reward of 50 points 40% of the time when Deck B was in play, but only 10% of the time when Deck A was in play (and zero points otherwise). The third option (Info) always offered a reward of 0.25 when Deck A was in play, and a reward of 0.75 when Deck B was in play. Thus, outcomes from this third card could be used to infer the current deck in play with certainty. The deck in play had a 5% chance of switching prior to the start of each trial.

Figure 2.3 illustrates the influence of limiting the Bayesian model's look-ahead to 1, 3, 4, and 6 actions into the future[2]. As the horizon extends further into the future, the model has more opportunity to exploit information that can move the model into

[2]The expected values converge asymptotically beyond $N = 6$ steps due to the probability of the deck switching between trials

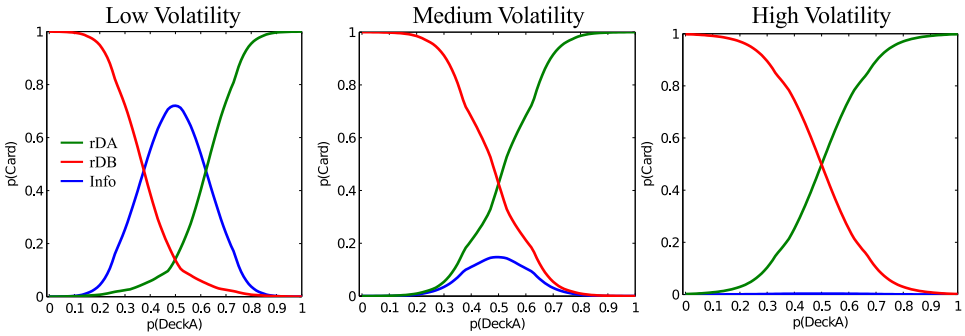

Figure 2.4: The influence of latent state volatility on the forward Bayesian model's action policy. As the environment becomes increasingly unstable (5%, 10% and 20% probability of switch latent state switch), the value of information decreases owing to a dwindling opportunity to leverage any gained information

high states of certainty, offsetting the initial cost of seeking that information. With a horizon of $N = 1$, the optimal actor simply exploits hA and hB according to the deck believed to be in play. At a horizon of $N = 3$, the model starts to consider the value of the Info card at high levels of deck uncertainty (around p(DeckA)=0.5); however, hA and hB are still favored in all possible belief states. At a horizons of $N = 4$ and $N = 6$ the model exhibits a strong preference for the Info card when deck uncertainty is high.

The benefits afforded by knowing which deck is in play on a given trial depends on the environment's stability. Figure 2.3 demonstrates the forward Bayesian model's response pattern as the environment becomes increasingly unstable, with the deck in play having a 5%, 10%, or 20% chance of switching prior to the start of each trial. When the environment is relatively stable (e.g. 5% chance of a deck switch between trials), the Bayesian model favors the Info option when it's fairly uncertain of which deck is in play. However, in a highly unstable environment (e.g. 20% chance of a deck switch between trials) the benefits of information quickly decay; and as such, the forward Bayesian model never favors the Info option, opting to exploit either hA or hB according to the current belief state.

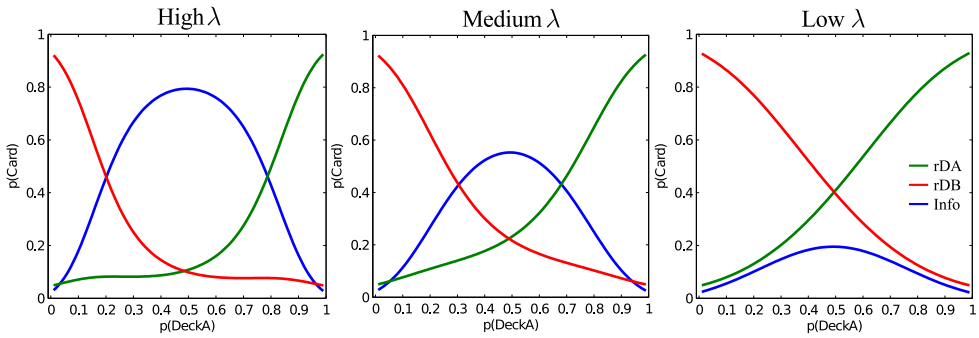

Figure 2.5: The influence of environmental λ on the heuristic actor's policy. As λ decreases, modelling increased volatility or a shortened horizon, the actor's propensity to seek information decreases.

I now turn to the influence horizon and volatility have on the action policy of the heuristic model. In its simplest form, the heuristic model as outlined in Equation 2.6 has no mechanism through which the task horizon can directly shape the action policy. The optimal observer's inferred belief incorporates the environment's volatility, offering an indirect means of shaping the heuristic model's policy, but the actor does not consider volatility explicitly. However, we can incorporate the effects of horizon and/or volatility into the action utility computation as an additional uncertainty-reduction weighting term:

$$U(a_i) = E[a_i] + \lambda \cdot \omega \cdot I(S_h; O_i) \tag{2.7}$$

where $\lambda \rightarrow 0$ to accommodate increasing volatility, and/or decreasing horizon. As illustrated in Figure 2.6, as $\lambda \rightarrow 0$ the action policy exhibits a decreasing preference for the Info option. Although λ does not map onto volatility or horizon directly, it can be scaled to capture the general effects exhibited by the forward Bayesian model. Considering a λ-like term in relation to brain function, its possible that a region like the ACC, which has been shown to be sensitive to volatility (Behrens & Woolrich, 2007), could play the role of weighting the benefit of information in light of estimated volatility or horizon.

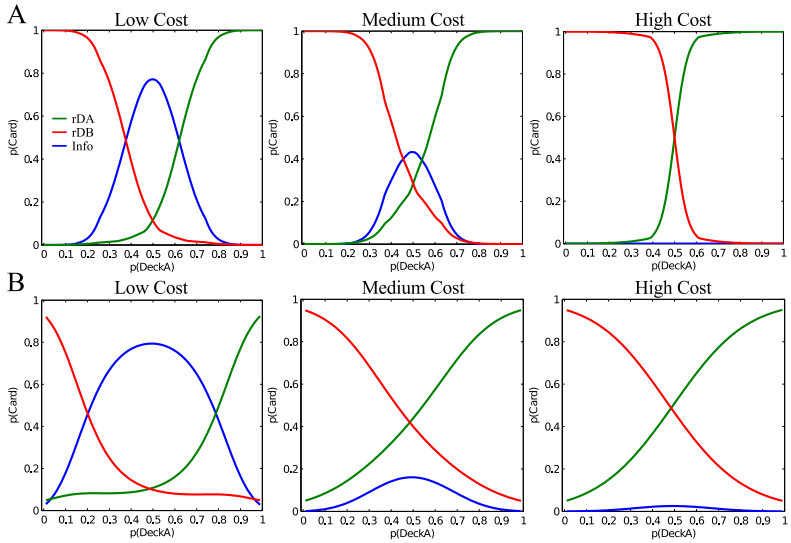

Figure 2.6: Effect of information cost on the action policy. Both the optimal (A) and heuristic (B) actor's exhibit a sensitivity to the cost of information relative to the potential rewards that can be gained. As the cost of information increases (modelled here as an increasing expected value for rDA and rDB), both actors exhibit a decreasing propensity to seek information

Sensitivity to the cost of information

The cost of information relative to the rewards plays an important role in shaping the optimal actor's action policy. Figure 2.6A shows the optimal actors action policy as we parametrically increase the probability of reward in both the rDA and rDB options. As information cost moved from Low to Medium the actor's policy demands higher levels of uncertainty to warrant selecting the Info card, until eventually it forgoes ever playing the Info card at High cost. Without modifying the heuristic utility equation 2.6, the heuristic actor exhibits the the same sensitivity to the cost of information (while holding ω constant, see Figure 2.6B). This pattern emerges as the expected reward factor in equation 2.6 increases, while the informative term remains constant (and it's scaling relative to expected reward), leads to a decreasing propensity to pursue informative actions as the cost of doing so increases.

Sensitivity to the benefit of information

The forward Bayesian model ascribes value to informative actions in line with the potential benefits that can be gained by reducing latent state uncertainty. That is, the Bayesian model will favor informative actions when it can leverage the information gained to attain more rewarding outcomes than would otherwise be possible. Thus, the Bayesian model doesn't value uncertainty reduction outright; rather, it values information as a means of future reward exploitation. Conversely, the heuristic uncertainty-reduction model ascribes value to uncertainty reduction without considering the future implications of doing so by assigning an expected uncertainty-reduction bonus to the expected reward value of each action. Thus, the heuristic model may pursue informative actions in situations where information will have no bearing on the action policy.

This distinction between models is made clear by exposing them to an environment where the optimal action policy does not depend on the latent variable's value. For example, when Deck A is in play, card 1 (High-Reward) has a 30% chance of awarding 50 points (and zero points otherwise), and card 2 (Low-Reward) has a 10% chance of awarding 50 points (and zero points otherwise). When Deck B is in play, card 1 (High-Reward) has a 60% chance of awarding 50 points (and zero points otherwise), and card 2 (Low-Reward) has a 30% chance of awarding 50 points (and zero points otherwise). Card 3 (Info) offers 1 or 2 points when Deck A and B are in play respectively. Here, the optimal action policy (when only reward is considered) simply selects the High-Reward card on all trials.

As illustrated in Figure 2.7A, the forward Bayesian model opts to select the High-Reward card regardless of it's inferred belief of the deck in play. It does so because there is no portion in the belief space where the Info card can pull the policy into

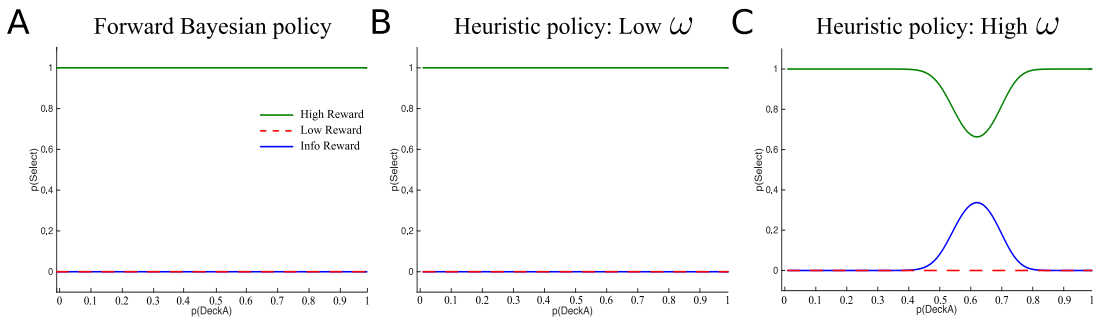

Figure 2.7: Action selection policy in a single policy environment. A) The probability that the forward Bayesian model will select the High-Reward, Low-Reward or Info card as a function of inferred belief that Deck A is in play. The Bayesian model's policy selects the High-Reward card across the entire belief space. B-C) The probability that the heuristic model will select each card with low (B) and high (C) emphasis on expected information gain. B) When information has no influence on the action policy ($\omega = 0$), the heuristic model exploits the High-Reward option regardless of belief state. C) As information is given a greater role in the utility function ($\omega = 4$), the heuristic model assigns some value to the Info card when uncertainty is high, and when Deck A (the least rewarding deck) is more likely to be in play.

a belief state where any card other than the High-Reward card is the best choice. Figure 2.7B shows that the action policy of the heuristic model mirrors that of the forward model when the uncertainty-reduction weight parameter, ω, is low. That is, the heuristic model exploits rewarding actions when information gain plays no role in the utility function. However, as ω increases and uncertainty reduction plays a larger role in the utility function, the heuristic model's action policy shifts to increase the probability of selecting the Info card (see Figure 2.7C). Noting that the expected value of Deck A (p(50 points)=0.3) is lower than the expected value of Deck B (p(50 points)=0.6), the heuristic model begins to express interest in the Info card when uncertainty is relatively high, and when the cost of Information is relatively low (i.e. when Deck A is more likely to be in play). This pattern emerges because expected reward is augmented according to the expected reduction of uncertainty. When expected values are lower (i.e. when Deck A is more likely to be in play), the uncertainty reduction terms plays a larger role in shaping each actions utility.

2.3.4 Conclusions

I have outlined an experimental design aimed at assessing the conditions under which information is sought in order to identify the appropriate policy for a given state of the environment. The task can be solved by building upon an optimal observer that blends latent state volatility and outcome statistics to infer the latent state of the environment. By leveraging the optimal observer's inferred belief regarding the deck in play, the forward Bayesian actor quantify the expected value of the options available to it by considering the implications of its actions toward a horizon some number of actions into the future. However, computing this forward model is computationally taxing. The heuristic approximation outlined here derives its action selection policy as a linear combination of expected reward and expected uncertainty reduction. This heuristic can be accommodate sensitivity to latent state volatility, task horizon, and the cost of information. I will now outline three experiments with human participants; one aimed at demonstrating the experimental design's validity, and the second aimed at quantifying sensitivity to the cost of information, and the third probes human performance in an environment where latent state has no influence on the optimal action selection policy.

Chapter Three

Behavioural Patterns of Probing

Having established our experimental design, and some of the factors that shape the value of information, I now turn to an investigation of human performance using the same design. In Experiment 1 I show that human participants seek latent state information when it is available and beneficial to do so, and that they leverage gained information to shape their action policy. Experiment 2 investigates human sensitivity to the cost of information, and demonstrates that humans are more willing to seek latent state information when it is less costly to do so. Finally, Experiment 3 probes the valuation of information, demonstrating that human participants seek latent state information even in circumstances where that information should have no impact of their action selection policy.

3.1 Experiment 1: Assessment of task sensitivity to probing

I begin with an assessment of the design's sensitivity to human information valuation. To do so, the experimental structure was encoded as a card game. The latent variable was encoded as possible decks from which cards could be drawn, and the various cards, who's values depend on the deck from which they are drawn, represent options available to the participant.

In this first experiment, the valuation of information was probed by exposing participants to two conditions. In the Informative condition, participants had to forgo possible reward in order to harvest information pertaining to the latent state of the environment (i.e. the deck in play). This was accomplished by structuring the outcomes for one of the cards (referred to as the Info card) such that it offered a very low expected reward value, but its outcomes were unique to each deck. As

such, outcomes from the Info card could be used to infer which deck was in play. In order to expose circumstances under which reward is waived in favor of information, participants were also exposed to a second Uninformative condition. The expected value of each card in this Uninformative condition was the same as the Informative condition; however, outcomes associated with what was the Info card (referred to as the NoInfo card in the Uninformative condition) were randomized. As such, the NoInfo card offered no information pertaining to the deck in play. Thus, the emergence of conditionally unique response patterns allow us to identify factors that contribute to the valuation of information, and how information is leveraged to inform action selection.

3.1.1 Methods

Fifteen participants were recruited from the Brown University subject pool. Participants were instructed that they would be dealt four symbolically unique cards on each trial. They were asked to select one card, which would be flipped to reveal the number of points earned. Participants were told that cards could be drawn from one of two decks; Deck A or Deck B. They were explicitly informed of each card's reward statistics in both decks, that all cards on a given trial would be drawn from same deck, but the deck in play had a 5% chance of switching before each deal . The participant's goal was to accumulate as many points as they could over the course of the game, which would be converted into a cash bonus at the end of the game.

Participants were exposed to two conditions. In the informative condition, one card (hA: high value Deck A) had a high expected value when drawn from Deck A (15% of 50 points, 0 points otherwise: expected value = 7.5 points), but a lower expected value when drawn from Deck B (5% of 50 points, 0 points otherwise:

expected value = 5 points). A second card (hB: high value Deck B) had a high expected value when drawn from Deck B (15% of 50 points, 0 points otherwise: expected value = 7.5 points), but a lower expected value when drawn from Deck A (5% of 50 points, 0 points otherwise: expected value = 5 points). A third card (Info: informative of the deck in play) offered small but reliable rewards for a given deck. It offered a reward of 0.25 points when drawn from Deck A, and 0.75 points when drawn from Deck B (both with 100% reliability). Thus, outcomes from this card could be used to infer the which deck that trial's cards had been drawn from. Since the Info card offered a reliable reward, albeit with very low expected values, participants might simply select it as a sure win. To address this concern, a fourth and final card (Const: constant value) offered a constant reward of 1 point regardless of the deck it was drawn from. This constant reward card gave participants the opportunity to seek a reliable reward, but it gave no information regarding the deck it was drawn from.

The Uninformative condition's reward statistics were identical to the Informative condition except for the Info card. In the Uninformative condition, the third card (NoInfo: randomized outcomes) offered the same reward (0.25 or 0.75 points), but did so randomly across both decks (50% of 0.25 points, 0.75 points otherwise). Thus, the reward outcomes offered from this card were no longer diagnostic of the deck they were draw from.

In order to familiarize participants with the game structure, and to give them a feel for the reward statistics, participants were initially given 20 practice trials in each condition. They were explicitly told which deck was in play during these trials, half of which were dealt from Deck A and the remaining half from Deck B. To reduce the possible impact of sampling biases (e.g. only sampling the most rewarding cards), participants were not only shown the reward of the card they selected, but

they were also shown counterfactual rewards from the other cards they did not select (although they did not accumulate points from those cards).

Before moving into the test phase of the experiment, participants were asked a series of probe questions using a two-forced choice design. They were repeatedly asked which of two possible cards had the highest expected value if drawn from a specified deck, and which of two possible cards was the most informative of the deck they were drawn from. Of the 15 individuals that participated in the task, four performed below chance on the probe questions and were removed from further analysis, leaving 11 participants for the final analyses.

Each test phase consisted of 200 trials (400 trials total across both conditions). Participants were reminded of the reward statistics for each card in both decks, that there was a 5% chance the deck would switch prior to each trial, but they would no longer be explicitly told which deck was in play. They were also informed that they would only observe the outcome of the card they selected.

3.1.2 Results

During the training phase, when participants knew which deck was in play, they preferred the card with the highest expected value (see Figure 3.1A). Card type (High or Low expected value), was entered as a predictor of response in a mixed-effects logistic regression, which confirmed that the Low value card was selected significantly less often during both the Informative (Beta = -1.96, $p<0.01$) and the Uninformative condition (Beta = -2.87, $p<0.01$). Thus, participants responded in line with the specified reward statistics, and exhibited different action selection policies according to the deck in play.

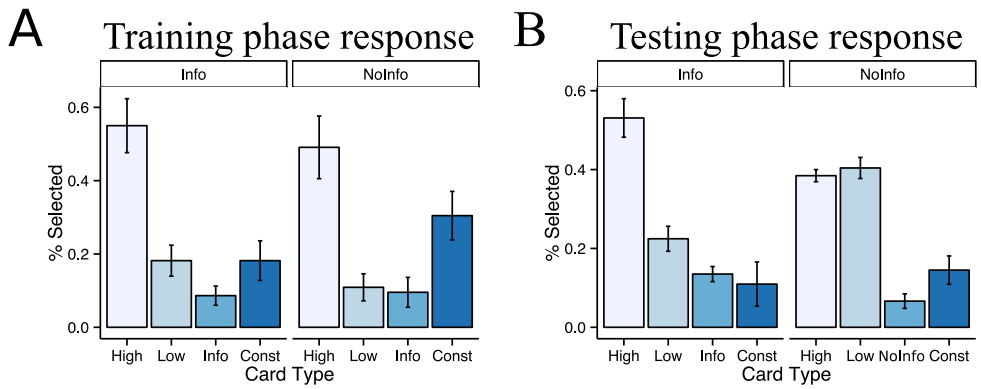

Figure 3.1: Training and testing phase response as a function of the true deck in play. A) Training phase. The proportion of trials in which each card type was selected when the deck in play was explicitly known during the Informative and Uninformative conditions (High: hA and hB when Deck A and B are in play respectively; Low: hB and hA when Deck A and B are in play respectively; Info: informative of deck in play; NoInfo: uninformative of deck in play; Const: constant value). B) Testing phase: The proportion of trials in which each card type was selected when the deck in play was not explicitly known during the Informative and Uninformative conditions.

Once participants had completed the training phase, they were exposed the testing phase of the task were they were no longer explicitly told which deck was in play. Mirroring their training phase performance, there was an overall preference for the rewarding cards (hA and hB) over less-rewarding cards (Info/NoInfo and Const) during both the Informative (Beta=1.6, $p<0.01$) and Uninformative condition (Beta=1.88, $p<0.01$). However, although participants exhibited a strong preference for the High value over the Low value card during the Informative condition (Beta=1.43, $p<0.01$), they show no such preference during the Uninformative condition (Beta=-0.07, $p=0.5$). Given that the only difference between the two conditions was the reliability of the Info/NoInfo card's outcomes as a function of the deck in play, this response pattern suggests that participants leveraged Info card outcomes in order to adopt an action selection policy better suited to the true, but unobservable, state of the environment.

Despite not having direct access to which deck cards were drawn from, participants exhibited a clear policy shift during the test phase of the Informative condition. Like the Bayesian observer, a participant's only recourse was to leverage previously

observed outcomes to infer which deck was in play on any given trial. This suggests that the Bayesian observer's inferred belief should be a better predictor of response than the veridical deck in play, and indeed this was confirmed by comparing the fit of a model using Deck, and a model using inferred belief predictors of response (Deck AIC=11256, Belief AIC=9441, $p<0.01$). Figure 3.2A illustrates the group probability of selecting each card as a function of the optimal observer's inferred belief. Mirroring the effect of deck on choice during the Informative condition, the probability of selecting the hA card increased (hA Beta=6.55, $p<0.01$), while the probability of selecting hB decreased (hB Beta=-5.95, $p<0.01$) as a function of the probability that Deck A was in play. As previously noted, when response was considered as a function of deck, participants did not differentiate between the High and Low cards during the Uninformative condition; however, they did exhibit a selective preference for the High valued card when response is considered as a function of inferred belief. Although the effects are much weaker in the Uninformative condition, the probability of selecting the hA card increased (hA Beta=2.13, $p=0.01$), while the probability of selecting hB tended to decrease (hB Beta=-1.33, $p=0.1$) as the probability that Deck A being in play increased.

Inferred belief not only shaped preference for the rewarding cards (hA and hB), but during the Informative condition, it also shaped preference for the Info card as well. Participants were more likely to select the Info card as they became increasingly uncertain of which deck was in play; that is, when their inferred belief did not strongly favor either deck being in play (e.g. $p(\text{DeckA}) = 0.5$, see Figure 3.2A). Uncertainty can be quantified as the belief distribution's entropy, which is maximal when there is an equiprobable belief that either Deck A or B is in play ($p(\text{Deck A}) = 0.5 \rightarrow H(\text{Deck A}) = 1$), and is minimal when either Deck A or Deck B is strongly believed to to be in play ($p(\text{Deck A}) = 1$ or $p(\text{Deck A}) = 0 \rightarrow H(\text{Deck A}) = 0$). Belief entropy

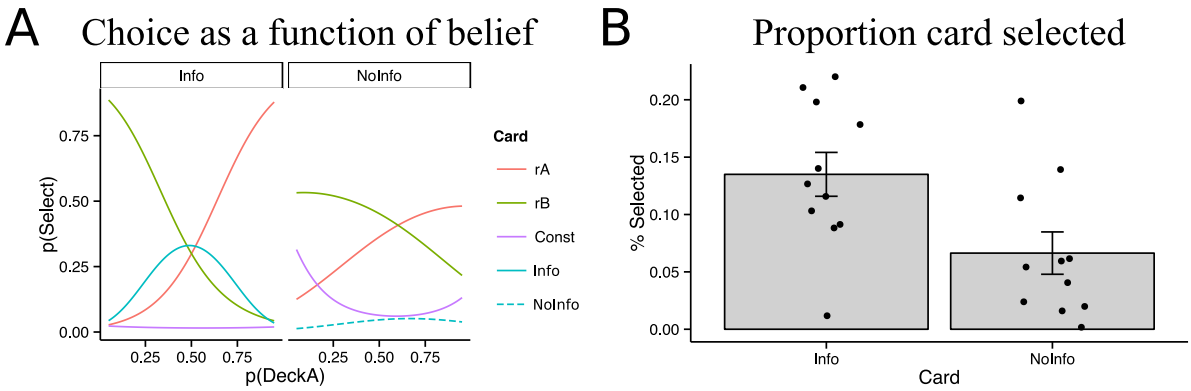

Figure 3.2: Effects of inferred belief on choice. A) The probability of selecting each card as a function of the optimal observer's inferred belief that Deck A was in play (hA: high value in Deck A; hB: high value in Deck B; Info: informative of deck in play; NoInfo: uninformative of deck in play; Const: constant value). B) The proportion of trials where the Info (Informative condition) and NoInfo (Uninformative condition) cards were selected. Points represent individual participants, bars represent group mean, error bars represent standard error of the mean.

was entered as a predictor of Info card response in a mixed-effect logistic regression, demonstrating that the likelihood of selecting the Info card increased significantly as a function of increasing uncertainty (Beta=4.07, $p<0.01$); however, no such effect was found for the NoInfo card (Beta=1.95, $p=0.42$). Furthermore, as depicted in Figure 3.2B, participants selected the NoInfo card significantly less than the Info card (Beta=-1.06, $p<0.01$).

3.1.3 Discussion

Participants preferred the card with the highest expected value during the training phase of both conditions, demonstrating that their response selection strategy depended on the deck in play when such information was provided. The same general pattern was also observed during the testing phase, where participants shifted their preference between hA and hB in accordance with the inferred belief that either Deck A or B was in play. Although they were able to adapt their response patterns as a function of inferred belief during both conditions, albeit much more reliably during

the Informative condition, participants were only able to reliably exploit the High reward card during the Informative condition. This suggests that participants did a better job of inferring which deck was in play when they had access to the Info card, and that they leveraged that information to direct their choices toward the card with the highest expected value.

An investigation into circumstances when the Info card was selected, a card that always had a lower expected value than all other options, showed that the Info card was increasingly preferred with growing uncertainty regarding which deck was in play. Moreover, the Info card was selected significantly more often than the NoInfo card, suggesting that participants were indeed selecting the Info card because of its informativeness. Thus, in summary, participants sought latent state information in accordance with inferred latent state uncertainty, and they used that information to guide their action selection policy, demonstrating that the experimental design is sensitive to directed exploration.

3.2 Experiment 2: Sensitivity to the cost of information

Both the Bayesian forward model, and the heuristic uncertainty-reduction model are sensitive to the cost of information. The forward model balances the cost of information against the potential benefits that may be harvested in the future, whereas the heuristic model augments the expected reward value according to expected uncertainty-reduction. While the mechanisms may differ, as the cost of information increases both models demands higher degrees of uncertainty before informative actions are favored over rewarding action. The purpose of Experiment 2 was to probe

human sensitivity to the cost of information, or in this case, the forgone reward of not exploiting more rewarding options.

3.2.1 Methods

Five participants were recruited from the Brown University subject pool. The design of Experiment 2 mirrored Experiment 1 with the exception of the Constant card which was removed. Feedback from Experiment 1 suggested that participants found it challenging to keep track each card's reward statistics. To avoid further compounding this issue in Experiment 2, where participants were exposed to additional conditions, the design was simplified by removing the Constant value card. However, since all conditions in Experiment 2 included the same informative option, any preference for a predictable outcome should be unaffected by manipulating the cost of information and should not reflect a conditionally varying desire for a guaranteed outcome.

Participants were instructed that they would be dealt three symbolically unique cards on each trial. As in Experiment 1, they were asked to select a single card, which would be flipped to reveal the number of points earned. Participants were told that cards could be drawn from one of two decks; Deck A or Deck B. They were explicitly informed of each card's reward statistics in both decks, that all cards on a given trial would be drawn from same deck, but the deck in play had a 5% chance of switching before each deal. The participant's goal was to accumulate as many points as they could over the course of the game, which would be converted into a cash bonus at the end of the game.

Participants were exposed to three conditions; Low-cost, Medium-cost, High-cost

information. In the Low-cost condition, card 1 (hA) had the highest expected value when Deck A was in play (40% chance of 100 points, 0 points otherwise; expected value = 40), and a lower expected value when Deck B was in play (10% chance of 100 points, 0 points otherwise; expected value = 10). Card 2 (hB) had the highest expected value when Deck B was in play (40% chance of 100 points, 0 points otherwise; expected value = 40), and a lower expected value when Deck A was in play (10% chance of 100 points, 0 points otherwise; expected value = 10). Card 3 (Info) offered small but reliable rewards for a given deck. It offered a reward of 1 point when drawn from deck A, and a reward of 2 points when drawn from deck B (both with 100% reliability).

In the Medium-cost and High-cost conditions, the reward statistics for the Info card remained the same while the probably of reward for the hA and hB cards increased. In the Medium cost condition, the hA and hB cards had a 70% chance of awarding 100 points when drawn from Decks A and B respectively (or 0 points otherwise; expected value = 70), and a 40% chance of 100 points when drawn from Decks B and A respectively (or 0 points otherwise; expected value = 40). The high-cost condition increased the probability of reward yet again, with the hA and hB cards offering a 90% chance of 100 points when drawn from Decks A and B respectively (or 0 points otherwise; expected value = 90), and a 60% chance of 100 points when drawn from Decks B and A respectively (or 0 points otherwise; expected value = 60).

As in Experiment 1, participants were given 20 practice trials in each condition, 10 drawn from Deck A, and 10 drawn from Deck B. Participants were also shown the outcome of all cards on each training trial. Each test phase consisted of 120 trials each (360 trials across all three conditions). Participants were reminded of the reward statistics for all cards in both decks, that there was a 5% chance the deck

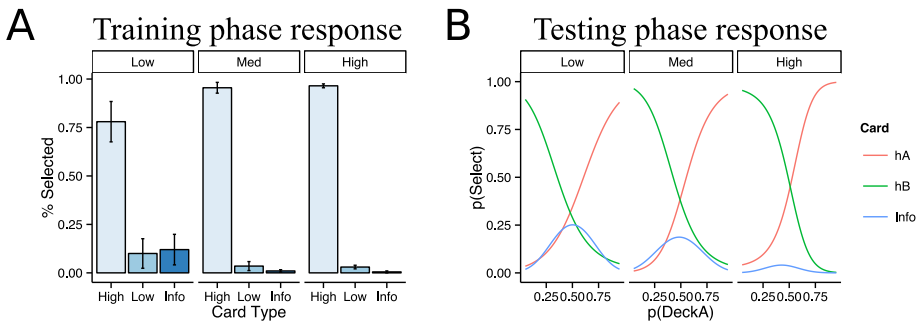

Figure 3.3: Training and Testing phase performance during Low, Medium, and High-cost conditions. A) The proportion of trials in which each card type (High reward, Low reward, and Info) was selected. Participants prefer cards with the highest expected reward value during all three conditions. B) The likelihood of selecting each card as a function of inferred probability that Deck A was in play. Participants show a strong preference for hA when Deck A was believed to be in play, and for hB when Beck B was believed to be in play (when p(Deck A) is low). Preference for the Info card is modulated according to deck uncertainty; however, willingness to seek information was tempered according to its cost.

would switch prior to each trial, and that they would no longer be explicitly told which deck was in play. They were also informed that they would only observe the outcome of the card they selected.

3.2.2 Results

Participants preferred the High value card (hA when Deck A was in play, and hB when Deck B was in play) during the training phase of all three conditions (see Figure 3.3A), indicating that the reward statistics and task structure were understood. Card type, (High or Low expected value), was entered as a predictor of response in a mixed-effects logistic regression, confirming that participant selected the Low value card significantly less often during the Low-cost (Beta=-7.15, p=0.02), Med-Cost (Beta=-8.03, p<0.01), and High-Cost conditions (Beta=-6.79, p<0.01).

As in Experiment 1, the optimal observer's inferred belief that Deck A was in play was a better predictor of testing phase response than the true but unobservable deck

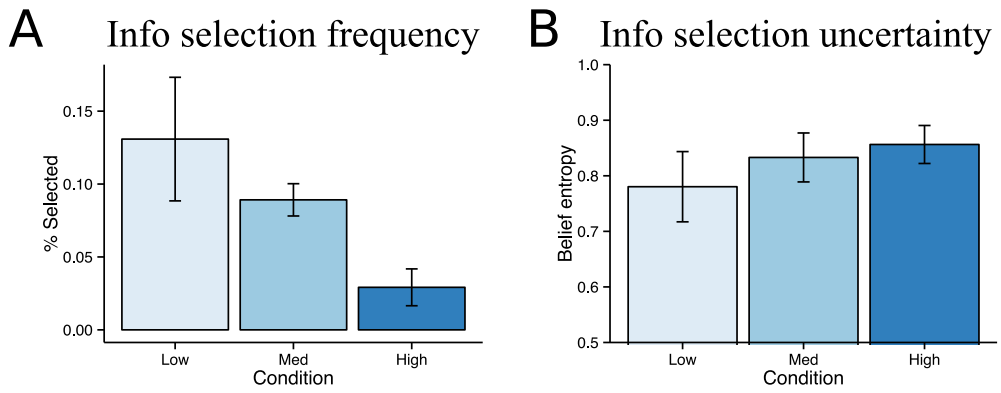

Figure 3.4: Info card selection. A) The frequency with which the Info card was selected in each information cost condition. Participants select the Info card less frequently as the cost of information increased. B) Mean uncertainty when the Info card was selected. As the cost of information increased, participants were more uncertain when they selected the Info card. Bars represent group means, error bars represent standard error of the mean

in play (Deck AIC = 10892, Belief AIC = 7919, $p<0.01$). Figure 3.3B illustrates the likelihood of selecting each card as a function of inferred belief. Participants were significantly more likely to selecting hA (Beta>6, and p-values<0.01 for all three conditions), and less likely to select hB (Beta<-6, and p-values<0.01 for all three conditions) as their inferred belief that Deck A was in play grew.

Inferred belief also influenced preference for the Info card. During the Low-cost condition participants were more likely to select the Info card as they became increasingly uncertain of the deck in play (Beta=4.62, $p<0.01$), replicating the effect of uncertainty observed in Experiment 1. However, in line with patterns predicted by both the Bayesian and heuristic models, participants were less willing to seek information as it became more costly to do so. As the cost of information increased participants selected the Info card significantly less often (Beta=-0.69, p=0.02, see Figure 3.4A), and they were significantly more uncertain when they did so ($\chi^2(1)=3.7$, p=0.05, see Figure 3.4B).

3.2.3 Discussion

During the training phase, participants exhibited a clear preference shift as a function of the deck in play in all three information cost conditions. As observed in Experiment 1, this response pattern is consistent with a strategy that engaged unique action policies for each value of the latent Deck variable. Participants exhibited the same preference during the test phase, when the deck in play was not explicitly known. Replicating effects observed in Informative condition of Experiment 1, participants sought information in line with their degree of uncertainty. Following Bayesian and heuristic model predictions, willingness to seek information was influenced by its cost. As the cost increased, participants required higher levels of uncertainty before information was favored over reward. Thus, consistent with predictions from both the Bayesian and heuristic models, participants exhibit a sensitivity to the relative costs/benefit tradeoff between information and reward.

3.3 Experiment 3: Non-Pecuniary value of information

The Bayesian forward model considers the implications of reducing its uncertainty by embedding the expected future benefits/detriments of doing so into its action utility value. As such, the forward model will only reduce its uncertainty in circumstances where doing so has a positive impact on its prospective action policy; that is, where information can be leveraged to pick more rewarding actions in the future. Conversely, the heuristic uncertainty-reduction model defines its action utility by adding an uncertainty-reduction bonus to each action's expected reward value. A unique

characteristic of this model's utility function is that it does not explicitly consider the implications of reducing its uncertainty, and as such, it may assign value to actions with informative outcomes in circumstances where doing so has no prospective benefit in terms of expected reward value. Thus, when exposed to environments where the action policy is unaffected by a latent variable's value, these two models differ in terms of whether or not uncertainty reducing outcomes are sought. Experiment 3 was designed to probe human propensity to reduce latent state uncertainty in circumstances where doing so should have no impact on their action selection policy.

3.3.1 Methods

Six participants were recruited from the Brown University subject pool. Participants were instructed that they would be dealt three symbolically unique cards on each trial. They were asked to select one card, which would be flipped to reveal the number of points earned. Participants were told that cards could be drawn from one of two decks; Deck A or Deck B. They were explicitly informed of each card's reward statistics in both decks, that all cards on a given trial would be drawn from same deck, but the deck in play had a 5% chance of switching before each deal . The participant's goal was to accumulate as many points as they could over the course of the game, which would be converted into a cash bonus at the end of the game.

Participants were exposed to two conditions; 2-policy and 1-policy environments. The 2-policy condition mirrored the low-cost condition of Experiment 2 where the most rewarding action was determined by the deck in play. Card 1 (hA) had the highest expected value when Deck A was in play (40% chance of 100 points, 0 points otherwise), and a lower expected value when Deck B was in play (10% chance of 100 points, 0 points otherwise). Card 2 (hB) had the highest expected value when Deck

B was in play (40% chance of 100 points, 0 points otherwise), and a lower expected value when Deck A was in play (10% chance of 100 points, 0 points otherwise). Card 3 (Info) was diagnostic of which deck the cards were drawn from, offering 1 point when drawn from Deck A, and 2 points when drawn from Deck B (both with 100% reliability).

In the 1-Policy condition, card 1 had the highest expected value in both decks (30% chance of 100 points Deck A, 60% chance of 100 points Deck B, 0 points otherwise), while card 2 offered a lower expected value in both decks (10% chance of 100 points Deck A, 30% chance of 100 points Deck B, 0 points otherwise). As in the 2-Policy condition, card 3 (Info) offered 1 point when drawn from Deck A, and 2 points when drawn from Deck B (both with 100% reliability).

Participants were exposed to 20 practice trials in each condition, 10 drawn from Deck A, and 10 drawn from Deck B. Participants were shown the outcome of all cards on each training trial. Each test phase consisted of 120 trials each (360 trials across all three conditions). Participants were reminded of the reward statistics for each card in both decks, that there was a 5% chance the deck would switch prior to each trial, but they would no longer be explicitly told which deck was in play. They were also informed that they would only observe the outcome of the card they selected.

3.3.2 Results

Mirroring training phase performance in Experiments 1 and 2, participants selected the Low reward card significantly less often than the High reward card during both the 2-Policy (beta=-0.67, p=0.01) and 1-Policy conditions (Beta=-4.3, p<0.01, see

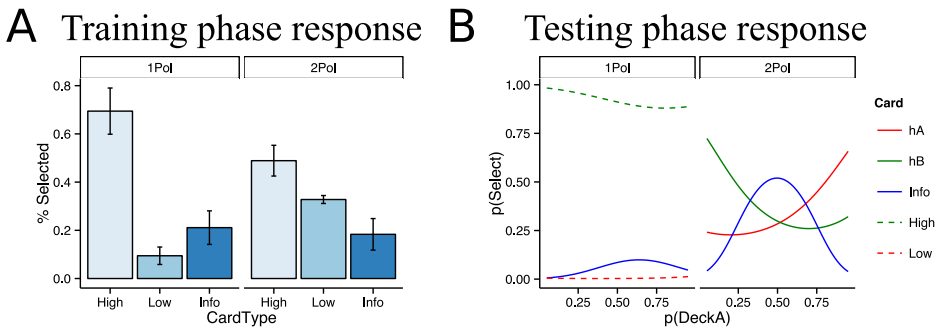

Figure 3.5: Task performance during 2-Policy and 1-Policy conditions. A) Training phase: The proportion of trials in which each card type (High reward, Low reward, Info) was selected. B) The probability of selecting each card as a function of inferred belief that Deck A is in play. During the 2-Policy condition, participants adjust their preference for rA and rB cards as a function of belief, and preferentially select the Info card when uncertain of the deck in play. During the 1-Policy condition, participants prefer the most rewarding card across the belief space, but also show a tendency to select the Info card when they are most uncertainty of the deck in play, and when information is less costly

Figure 3.5A). Figure 3.3B illustrates the likelihood of selecting each card as a function of the inferred belief the Deck A was in play during the testing phase. During the 2-Policy condition, as was observed in Experiments 1 and 2, the hA card was significantly more likely to be selected (Beta=2.67, $p<0.01$), and the hB card was significantly less likely to be selected (Beta=-2.47, $p<0.01$) as the inferred belief that Deck A was in play grew. Furthermore, participants were willing to forgo potential reward in favor of information when deck uncertainty was high (effect of belief entropy on Info card response: Beat=4.63, $p<0.01$).

Most participants chose the Info card during the 1-Policy condition despite there being no advantage to knowing which deck was in play. As in the 2-Policy condition, participants were more likely to select the Info card as their uncertainty regarding the deck in play grew (effect of belief entropy on Info card response: Beta=1.23, $p<0.01$). However, there was an asymmetrical propensity to select the Info card across decks, with the Info card being selected preferentially when it was less costly to do so (when Deck A was believed to be in play, see Figure 3.5B). Indeed, as illustrated in Figure 3.6A, the inferred belief was strongly biased toward Deck A being in play

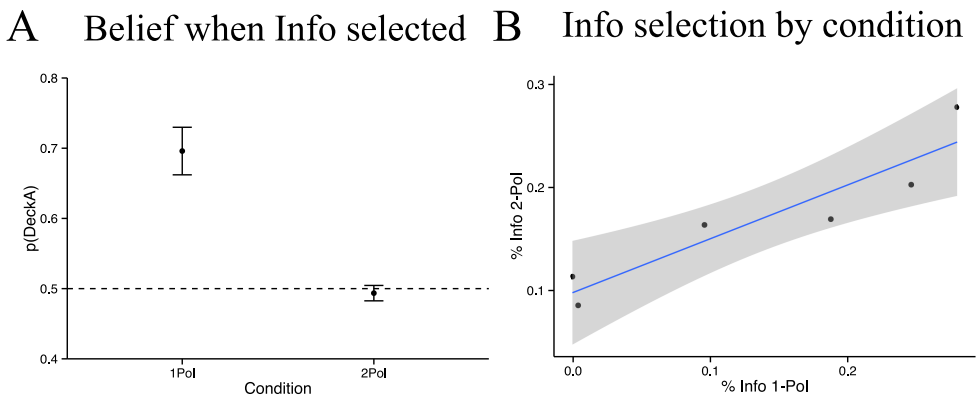

Figure 3.6: Conditions and consistency of information valuation. A) The mean inferred belief when the Info card was selected. Deck belief was balanced across decks during the 2-Policy condition (i.e. 50% chance that either deck was in play), but strongly biased in favor of belief that Deck A was in play during the 1-Policy condition. B) Correlation between the proportion of trials in which subjects selected the Info card in each condition, suggesting information valuation consistency within individuals.

when the Info card was selected during the 1-Policy condition ($t(5)=5.79$, $p < 0.01$); there was no such bias during the 2-Policy condition ($t(5)=-0.5$, $p=0.58$). Finally, individuals that valued information more heavily in the 2-Policy condition did so in the 1-Policy condition as well (see Figure 3.6B, $r(5)=0.92$, $p<0.01$), suggesting that individuals valued information consistently throughout the experiment.

3.3.3 Discussion

Behavioural response patterns during the 2-Policy conditions were consistent with predicted behaviour from both the forward Bayesian model and the heuristic uncertainty-reduction model. Participants were more likely to select the most rewarding card when they were relatively certainty of the deck in play, but exhibited an increasing propensity to seek information as deck uncertainty increased. Participants preferred the most rewarding card during the 1-Policy condition as well. However, in contrast to the response pattern predicted by the forward Bayesian model, where neither the deck nor the degree of uncertainty had any influence over the action selection policy,

participants sought deck information nonetheless. Participants were drawn to the Info card when deck uncertainty was relatively high, and particularly when information was believed to be less costly (when Deck A was more likely to be in play). The heuristic uncertainty-reduction model, where each action's expected reward value is augmented according to its expected uncertainty reduction, generates this same response pattern.

As in Experiment 2, the 1-Policy response pattern demonstrates that participants were sensitive to the relative cost of information, preferring to seek latent state information when it was less costly to do so. Behaviour during the 1-Policy condition also indicates that participants did not employ the strategy formalized by the Bayesian forward reward model. However, I caution that this response pattern does not necessarily imply that humans do not employ a forward model. It's possible that the problem optimized by the forward Bayesian model is an over-simplification of the challenge faced by participants; and indeed, recent work has shown that human performance is near-optimal if inference over task structure is also taken into consideration (Acuña & Schrater, 2010).

The brain is faced with the challenge of identifying and executing rewarding actions. This presents the brain with two potentially conflicting constraints: a) take the most rewarding actions, and b) integrate your observations into a veridical representation of the environment. In the task structure outlined here sample integration is facilitated by reducing deck uncertainty, which allows observed outcomes to be segregated as a function of the deck believed to be in play. Although action values were explicitly given, making outcome integration redundant for a rational agent, some brain regions may not work with explicit numerical information efficiently. As such, the brain may assign some value to knowing the latent state of the environment in order to improve its action value estimates when outcomes are observed.

However, integrating the additional factors required to optimize learning (what ever those factors might be) brings the forward model's utility function into a realm of complexity and numerical precision almost certainly beyond the grasp of a biological implementation. Conversely, a computation like the uncertainty-reduction bonus as implemented by the heuristic model offers a straight-forward method of addressing the brain's eagerness to learn from its observations. It's possible that the brain embodies this learning constraint not by explicitly computing the expected learnability associated with an outcome, but by using uncertainty reduction as an simplified and generally beneficial index of learnability.

Chapter Four

Neural correlates of information value

4.1 Introduction

In Experiments 1, 2 and 3, human participants sought latent state information as a function of their uncertainty, and they used that information to adjust their action selection policy. Participants were also sensitive to the cost of information, and exhibited a desire for latent state information in circumstances when it had no impact of the optimal action selection policy. I turn now to the question of how the brain tracks and processes the variables necessary to determine the value of information.

The forward Bayesian model computes an action utility value that blends expected reward and information gain, where information is 'valued' according to the reward that can be expected from moving into different regions of the belief space. Alternatively, the heuristic model computes and combines separate terms for expected reward and expected uncertainty reduction. Thus, the forward and heuristic actor models make different predictions regarding the kinds of signals we might expect to find in the brain: the forward model suggest that we should find a singular index value signal, whereas the heuristic model suggest that we might find separable signals associated with expected reward and expected uncertainty reduction.

To probe the characteristics of the brain's value computation, the scalp electroencephalography (EEG) was monitored while participants played a variant of the card game task. The scalp EEG offers high temporal precision (on the order of milliseconds), but relatively poor spatial precision (on the order of centimeters at best). Thus, separable reward and information signals should be detectable if they follow different time courses, and/or they engage regions of the brain that produce different dipole capable of generating a cohesive electrical signal detectable at the scalp.

In order to provide a basis from which more experimental analyses will be rooted,

a traditional event-related potentials (ERP) approach will be used to analyze the EEG data. In brief, ERPs are created for conditions of interest by averaging the scalp EEG across trials around some event of interest (e.g. stimulus onset). By averaging the EEG signal from many trials together, signals produced by cognitive processes of interest remain, while non-event coherent noise associated with other brain function or environmental interference (e.g. heartbeat, line noise) is reduced.

A wealth of research has focused on various ERP components and their relationship to cognitive processes. Of particular interest here is a component referred to as the N2, a negative deflection in the ongoing EEG at approximately 200-300 ms following stimulus onset. Previous work has linked modulation of then anterior N2 to various aspects of cognitive control, including attention, reward, conflict monitoring, and visual template matching (Folstein & Van Petten, 2008). Of particular relevance, a collection of studies have suggested that central anterior N2 modulation is driven in part by phasic dopamine signals, situating N2 modulation as an index of reinforcement learning processes in the brain (Cockburn & Frank, 2011; Holroyd & Coles, 2002; Holroyd et al., 2011; Krigolson & Holroyd, 2007; Krigolson et al., 2011; Baker & Holroyd, 2011).

Though greatly understudied, recent work has begun to identify EEG signals associated with an option's subjective value. Both Harris et al. (2011) and Larsen & O'Doherty (2014) applied an functional magnetic resonance imaging (fMRI) style analysis to identify a late-wave frontal modulation of the scalp EEG. Rather than averaging several trials together (as is done with standard ERP analysis), both of these studies applied a separate general linear model (GLM) to each channel/time-point in the data set. This approach facilitated the use of model-based parametric predictors to identify a late (700ms) anterior (Fz) signal associated with model-derived subjective value. The study reported by (Larsen & O'Doherty, 2014) also

collected fMRI data, which they leveraged to identify the ventromedial and lateral prefrontal cortex as potential sources of the EEG signal.

4.2 Methods

Forty participants drawn from the Brown University subject pool and the local community received course credit or financial compensation for taking part in the study. The task structure was similar to that of Experiments 1 with a few modifications to accommodate scalp EEG recording. Participants were told that they would be dealt one of three symbolically identifiable cards on each trial, and asked to choose between playing that card or a 50/50 lottery. On each trial, a card stimulus was presented center screen (700-1000ms), after which 'Card' and '50/50' buttons were presented to the right and left of the card stimulus (50/50 and Card buttons were randomly placed on the left or right each trial in order to prevent response preparation). Once a choice was made, all stimuli were cleared from the screen feedback (i.e. reward won) was presented (1000ms).

Participants were told that cards could be drawn from one of two decks; Deck A or Deck B. They were explicitly informed of each card's reward statistics in both decks, and that the deck in play had a 5% chance of switching before each deal. One card (high reward Deck A: hA) had a high expected value when drawn from Deck A (40% of 40 points, 5 points otherwise: expected value = 19 points), but a lower expected value when drawn from Deck B (15% of 40 points, 5 points otherwise: expected value = 10.25 points). A second card (high reward Deck B: hB) had a high expected value when drawn from Deck B (75% of 25 points, 5 points otherwise: expected value = 20 points), but a lower expected value when drawn from Deck A

(25% of 25 points, 5 points otherwise: expected value = 10 points). A third card (high deck information: Info) offered small but reliable rewards for a given deck. It offered a reward of 3 points when drawn from Deck A, and 4 points when drawn from Deck B (both with 100% reliability). Participants were given the choice of playing the card they were dealt, or a 50/50 lottery that awarded 22 or 8 points (expected value = 15 points).

Participants were also told that there was a 10% chance they would be asked to explicitly note their belief that either deck was in play after each trial. The belief probe portion of the task provided a rough measure of the degree to which the optimal observer's inferred belief matched that of the participant. Since participant were only dealt one card at random on each trial there was no guarantee that knowledge of the deck in play could be leveraged on the next trial. As such, participants were motived to track the value of knowing which deck was in play in that the computer would use their noted belief to make 10 choices on their behalf, and that they would be awarded all points accumulated by the computer (without observing any outcomes, and without risk of the deck switching between trials). Using this scheme, knowing the deck in play would allowed participants to give a more accurate estimate of the true deck in play, which in turn, would result in better choices and more rewards on behalf of the computer. As before, the participant's goal was to accumulate as many points as they could over the course of the game, which would be converted into a cash bonus at the end of the game.

As in previous experiment, the optimal policy across decks shifted from playing card hA (and the lottery otherwise) when Deck A was in play, to playing card hB (and the lottery otherwise) when Deck B was in play. However, unlike our previous experiments, the level of information across all three cards varied. The mutual information shared between the Info card's outcomes (O_I) and the deck in play was perfect

($I(O_I; Deck) = 1$). The hA card's outcomes (O_{hA}) provided very little information regarding the deck in play ($I(O_{hA}; Deck) = 0.058$), whereas the hB card's outcomes (O_{hB}) provided an intermediary degree of information ($I(O_{hB}; Deck) = 0.19$). As such, both the expected value and the expected information gain from each card varied as a function of belief. Although expected value and expected information gain share some variance (as deck uncertainty increases, expected value of hA and hB decrease), systematically varying both reward and mutual information across cards drove some separation between signals associated with reward and information (should they exist).

Participants were exposed to three interleaved blocks of training for each deck (20 trials in each block, for a total 120 trials). We extended the training protocol owing to the fact that participants were not given counterfactual outcome information, and to ensure a sufficient number of trials for EEG analyses. Following the training phase, participant entered the test phase (400 trials). Throughout the test phase, participants were intermittently reminded of all reward statistics, and that the deck in play had a 5% chance of switching prior to each trial.

4.2.1 EEG recording and preprocessing

:

EEG was recorded continuously across 0.1 – 100 Hz with a sampling rate of 500 Hz and an online CPz reference on a 64-channel Brain Vision system. Data were epoched around the onset of card presentation (-500 to 2000 ms), visually inspected to identify and remove bad channels and epochs, then submitted to Eeglab's independent component analysis decomposition algorithm. Components were visually

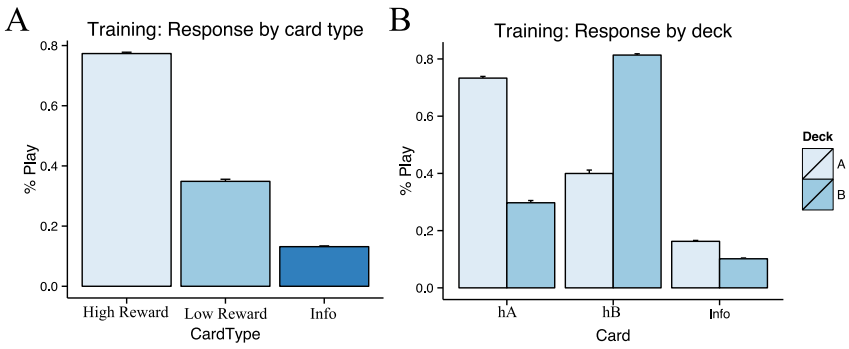

Figure 4.1: Training phase performance. A) The proportion of trials each card type (High-Reward, Low-Reward, and Info) was played. B) The proportion of trials each card was played as a function of the deck in play.

inspected to identity eye artifacts (identified by low-frequency, large, frontal signals). Eye artifact components were then removed from the original dataset, epochs were down-sampled to 250 Hz, re-filtered to 0.1 – 25 Hz, limed to -200 to 700 ms, and re-referenced to the average scalp voltage. Bad channels were interpolated, epochs were baseline corrected (-200 to 0 ms), and epochs with artifacts were removed (voltage $\pm$ 75 μV).

4.3 Results

4.3.1 Behaviour

As illustrated in Figure 4.1A, participants preferred the most rewarding card across decks. Card type (High-Reward, Low-Reward, Info), as defined by the deck in play, was entered as a predictor of response in a mixed-effects logistic regression, showing that participants exhibited a strong preference for the High-Reward card over the others (contrasts relative to High-Reward, all Beta's < -2, all p-value < 0.01). Inspection of the response patterns for the rewarding cards (hA and hB) showed the expected preference shift across decks (see Figure 4.1B). Card (hA and hB) and

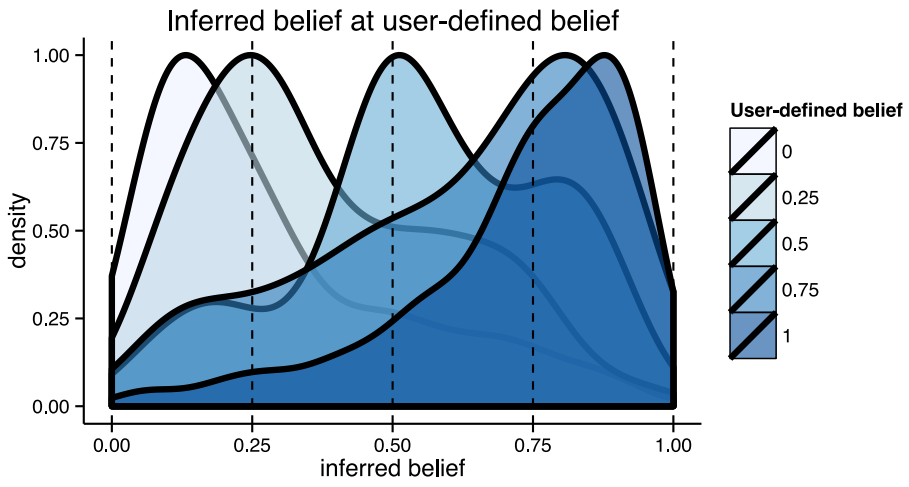

Figure 4.2: Relationship between inferred and user-defined belief. Participants were asked to mark their belief that Deck A was in play (0%, 25%, 50%, 75%, or 100%). The distribution of inferred beliefs for each level of explicitly noted belief across all participant peak in proximity to the user-defined belief.

Deck (A and B) were entered as predictors of response during training, revealing a significant Card by Deck interaction (Beta=1.23, $p<0.01$). As in Experiments 1, 2 and 3, these results demonstrate that participants understood the task, responded in line with the specified reward statistics, and exhibited different action selection policies according to the deck in play.

Ensuring the validity of the optimal observer's inferred deck belief as a predictor of response, model comparison demonstrated that a model including the optimal observer's inferred belief was a better predictor of response than a model that included the true deck in play on each trial (Deck AIC=16199, Prior AIC=15987, $p<0.01$). With the addition of the belief-probe portion of the task, the relationship between inferred belief and the participant's explicitly noted belief could be probed. Figure 4.2 depicts the distributions of inferred belief for each level of explicitly marked belief. Notably, there are clear peaks in the inferred belief distributions roughly centered at the noted levels of belief, indicating a relatively good match between inferred and participant denoted deck belief.

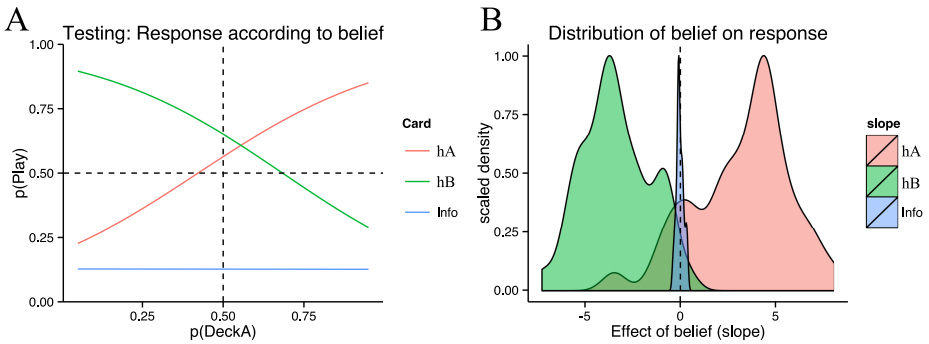

Figure 4.3: Test phase performance as a function of inferred belief. A) Fixed-effects extracted from a mixed-effects logistic regression illustrating the probability of selecting each card as a function of inferred belief. B) The distribution of random-effects across all subject.

As illustrated in Figure 4.3A, participants shifted their preference between hA and hB as a function of inferred belief. A mixed-effects logistic regression showed a significant interaction between belief and their propensity to play each card (Beta=3.41, $p<0.01$). However, there was also a main effect of card (Beta=-1.89, $p<0.01$), indicating that participants played the hB card significantly more than the hA card. This effect will be investigated in more detail shortly, but this preference is in line with an information seeking strategy whereby participant opt to play hB (which offers a moderate amount of latent state information) more often than the hA card (which offers almost no latent state information).

Turning now to the effect of uncertainty on response patterns, Figure 4.4A depicts the effect of uncertainty on the groups response, and Figure 4.4B illustrates the distribution of uncertainty effects on each card across participants. Statistical analyses show only a slight and non-significant interaction between uncertainty on Info card response propensity (Beta=0.82, p=0.17); however, the distribution of this effect this varies considerably across participants (see Info Figure 4.4B), perhaps due to adjustments made to the task design to accommodate a EEG recordings that make the information gleaned from the Info card more challengig to exploit. In support of the EEG analyses to follow, a median split on propensity to probe formed Probe

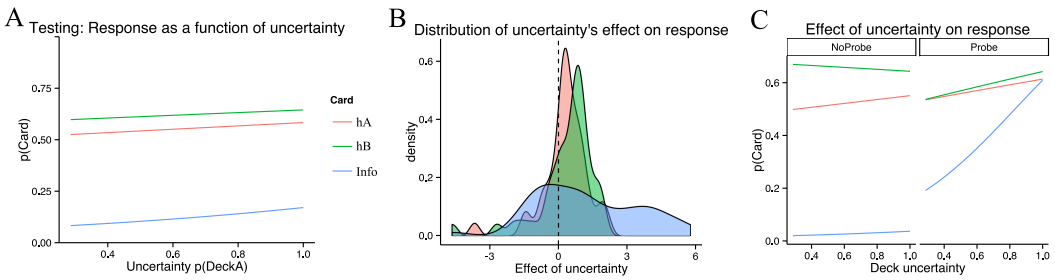

Figure 4.4: Test phase performance as a function of inferred belief uncertainty. A) Fixed-effects extracted from a mixed-effects logistic regression illustrating the probability of selecting each card as a function of inferred belief entropy. B) The distribution of random-effects across all subject. C) Fixed-effects extracted from a mixed-effects logistic regression illustrating the probability of selecting each card as a function of inferred belief entropy for the Probe and NoProbe groups.

and NoProbe groups. As depicted in Figure 4.4C, the NoProbe group exhibited no effect of uncertainty on Info card response (Beta=0.58, p=0.64), whereas the Probe group show a strong effect of uncertainty (Beta=2.19, p<0.01).

To summarize, participants exhibited a strong preference for the card with the highest expected value when the deck in play was explicitly known (during training), and when uncertainty was low during testing. Their explicitly noted deck belief matched that of the optimal observer's, indicating that participants integrated information their observations rationally. Finally, participants exhibited considerable variance with respect to their preference for the Info card. However, by dividing the group according to their overall propensity to probe, an uncertainty-driven information seeking response pattern emerged in the Probe group.

4.3.2 EEG correlates of stimulus value

Turning now to the scalp EEG proximal to the point of card stimulus presentation, I begin with a traditional ERP analysis of the data with respect to behaviorally defined preferences. I then move to a model-based analysis to identify signal related to expected value and uncertainty reduction.

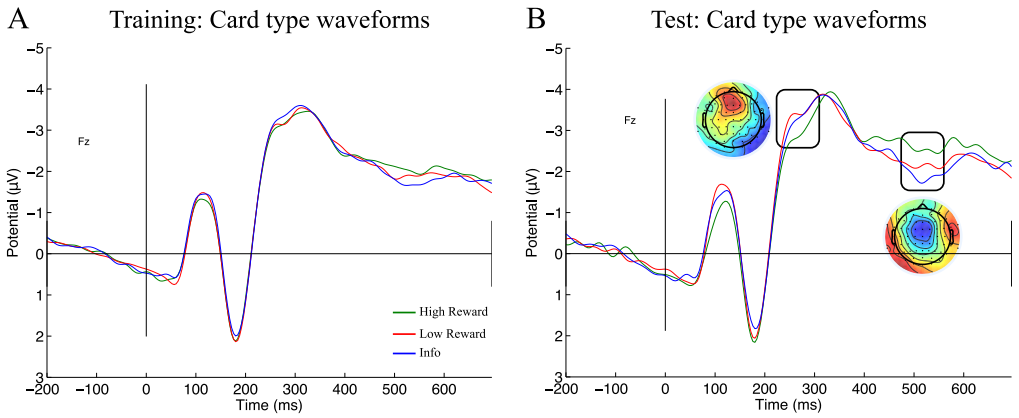

Figure 4.5: Front-Central (Fz) ERP waveforms locked to card stimulus onset A) Training phase waveforms for as a function of card type (High-Reward, Low-Reward, Info). No effect of card type was observed during training. B) Testing phase waveforms when inferred uncertainty was low. The N2 component was modulated according to expected value of choice, and the late-wave (proximal to 500ms) was modulated according to expected value of the stimulus.

ERP analyses

Participants were exposed to an extended training phase (120 trials across both decks) in order to collect a sufficient number of trials to facilitate an analysis focusing on reward expectancy in the absence of any deck uncertainty. Previous work has found modulation of the front-central N2 component in line with expected reward, where stimuli associated with less rewarding outcomes drive a negative deflection 250ms post stimulus onset. However, as illustrated in Figure 4.5A, there was no frontal effect of card type whatsoever. In short, there no effect of stimulus, response, reward magnitude, reward probability or expected value during the training phase. Although somewhat disappointing, this lack of effect may be a result of the task simplicity during the training phase, or from averaging over an evolving strategy as training progressed.

Shifting focus to the ERPs during the testing phase when inferred uncertainty was low (i.e. a strong belief that either deck was in play, p(Deck1)>0.7 or p(Deck2)>0.7), and limiting the data to only include participants with sufficient trials for analysis

(20 trials per condition, for N=23 subject), Figure 4.5B shows clear modulation of the N2 as a function of the trial's expected value. During periods of low deck uncertainty, participants generally opted to play the High-Reward card, but opted for the 50/50 lotto when presented with either the Low-Reward or Info card. A analysis of voltage at front-central sites (AFz, F1, Fz, F2, FCz) during the time period of 230-270 ms contrasting High-Reward vs. Other (Low-Reward and Info) revealed a significant effect (adjusted $F(1, 18)=4.64$, $p=0.05$), with Other card presentation being associated with a negative deflection of the waveform.

There also appears to be separation among waveforms associated with High-Reward, Low-Reward, and Info cards later in the epoch (proximal to 500ms). Indeed, previous studies have found late-wave effects associated with subjective value (Larsen & O'Doherty, 2014; Harris et al., 2011). In line with these studies, voltages at front-central sites (AFz, F1, Fz, F2, FCz) during the time period of 450 to 550 ms contrasting card Type (High-Reward, Low-Reward, Info) were submitted to a robust ANOVA, revealing a significant effect of card type ($F(2, 16)=5$, $p=0.037$, with Large<Small<Info). These results hint at separable value based EEG components. An early component that appears to be associated with the expected value of the choice, and a later component associated with the expected value of the stimulus itself. However, since the expected value of each stimulus is a continuous function of the belief state (i.e. belief of which deck is in play), I leave a more detailed analysis of this late component for a model-based analysis.

Behavioural analyses revealed a high degree of individual differences with respect to the propensity to play the Info card. Focusing on those individuals that did select the Info card (the Probe group), although there was an effect of expected choice value on the N2, there was no such modulation of the N2 component when uncertainty was high (adjusted $F(2,8)=0.91$, $p=0.45$, see Figure 4.6)A. However, as observed when

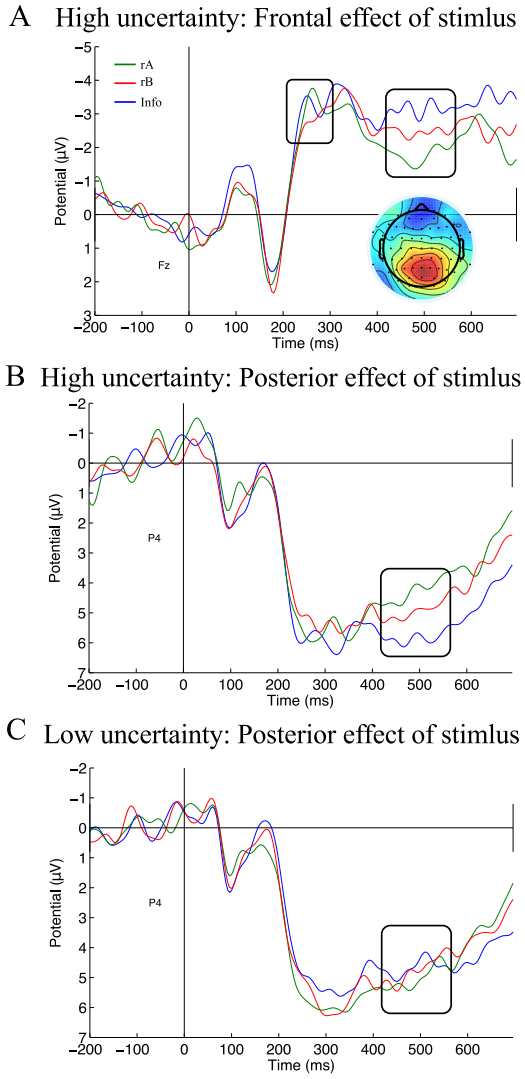

Figure 4.6: Effects of uncertainty at A) frontal and B) posterior channels in the Probe group. A) Waveforms at frontal sites (here Fz) when uncertainty was high. There is no significant modulation of the N2, but the late-wave component is modulated in accordance with subjective value (as exhibited by behavioural preference of Info>hB>hA). B) Waveforms at posterior sites (here P4) when uncertainty was high. There is significant modulation of the late-wave component in line with expected information gain for each stimulus (Info>hB>hA). C) Waveforms at posterior sites (here P4) when uncertainty was low. No modulation of the late-wave component was observed.

uncertainty was low (see Figure 4.5)B), frontal waveforms were modulated according to subjective preferences when uncertainty was high. As illustrated in Figure 4.6)A, there is a clear modulation in the 450 – 550ms time range, with Info<hB<hA (adjusted $F(2,8)=5.32$, $p=0.02$). This mirrors late-wave expected value modulation observed when uncertainty was low (High-Reward>Low-Reward>Info).

The scalp plot proximal to 500ms (see Figure 4.5)A), suggests two loci of activity; the frontal signal just discussed, and a second posterior signal. An investigation of posterior sites revealed a strong late-wave effect at approximately the same time as the frontal signal just discussed in line with the expected information gain of each stimulus (see Figure 4.5B, Info<hB<hA, adjusted $F(2,8)=6.33$, $p<0.01$). Furthermore, this posterior signal is only observed when uncertainty is high (Figure 4.5C, $F(2,8)=0.46$, $p=054$).

In summary, the ERP analyses revealed a modulation of the N2 component during the testing phase of the task. This modulation appears to track the expected value of choice on the trial, with lower valued choices eliciting a negative deflection of the N2. A late-wave component associated with expected reward value was also identified during periods of low uncertainty (when a stimuli's expected value was relatively well known). However further investigation of this late-wave component during periods of high uncertainty suggest that it may be tracking a more robust subjective value that also takes a stimuli's expected uncertainty reduction into account. Finally, a posterior late-wave component was also identified. This posterior modulation does not appear to be associated with expected reward value, but does appear to track the expected uncertainty reduction associated with a particular stimulus. I know turn to a novel model-based analysis of the EEG data to provide a more comprehensive understanding of the trial-by-trial value dynamics.

Model-based analyses of the EEG

Traditional ERP analyses, which are principled and well understood, require some form of trial averaging. As demonstrated in the previous section, this approach can give some insight into the characteristics of the scalp EEG and it's relationship to cognitive processes involved with computing value. However, this approach does not take advantage of the trial-by-trial dynamics of the task, where expected value, expected information gain, and action utility are continuously valued functions. Here I outline and apply a novel model-based approach to analyzing EEG data, the inspiration for which is rooted in model-based fMRI analysis, and applications outlined by Harris et al. (2011) and Larsen & O'Doherty (2014).

Rather than create condition-wise averages for each subject and analyze those averages across subjects, a set of GLMs are defined for each subject, using a separate GLM for each channel and time-point in the ongoing scalp EEG. Thus, for a given channel (e.g. Fz) and time-point (e.g. 180ms post stimulus onset), a GLM is constructed using a set of variables (e.g. expected value and expected information gain for each trial as derived from the uncertainty-reduction model) to predict the the channel/time-point voltage on each trial. This produces a map of beta weights for each predictor in the GLM corresponding to each channel/time-point pair for each subject. In summary, the trial-by-trial estimates for a model-based parameter estimate will act as a GLM's parametric regressors, and the trial-by-trial voltage for a given channel/time-point will provide the dependent variable being predicted.

The next step is to analyze the beta maps across subject. By applying a GLM to each channel/time-point pair (for a total of 64*225=14400 measurements per regressor), we are faced with daunting multiple comparisons problem. A Bonferroni correction would severely compromise the power to detect real effects. As such, a

more powerful approach is to apply a modified cluster-correction strategy often used in fMRI voxel data. In essence, each cell of all subject's beta maps (where a cell is comprised of a single channel/time-point pair) for a given predictor can be submitted to a t-test. If a cell's t-test is significant to a specified threshold across subject a cluster is formed. In the second phase, neighboring channels and/or time-points in the same channel that exceeded threshold are joined to form a single cluster. The end result is a number of channel/time-point clusters that are separated by by regions where the t-test did not reach the specified threshold level. Significance testing was performed by permutation test. A t-weight was computed for each cluster (the mean t-score for a given cluster, which accounts for both cluster size and effect size simultaneously). A t-distribution was then computed by repeatedly submitting sign-permuted beta values to a t-test. A cluster was deemed significant if its t-weight fell beyond the 95th percentile of the t-weight distribution.

The first model-based analysis of the EEG is a two-step process using the model definition outlined in Equation 4.1. The expected value ($E[a_i]$)and expected uncertainty-reduction terms ($\sigma \cdot I(S_h; O_i)$) as computed by the heuristic model were used to define two regressors. The first regressor provides a significance mask and is used to identify cells associated with either expected value and/or expected information gain terms. Those cells that survived the significance thresholding on the first regressor had the second regressor submitted to a sign test. Owing to the regression definition, cells associated with information gain should have a positive sign, while cells associated with expected value will have a negative sign.

$$V_{c_i,t_i} = \beta_0 + \beta_1(\sigma \cdot I(S_h; O_i) + E[a_i]) + \beta_2(\sigma \cdot I(S_h; O_i) - E[a_i]) \qquad (4.1)$$

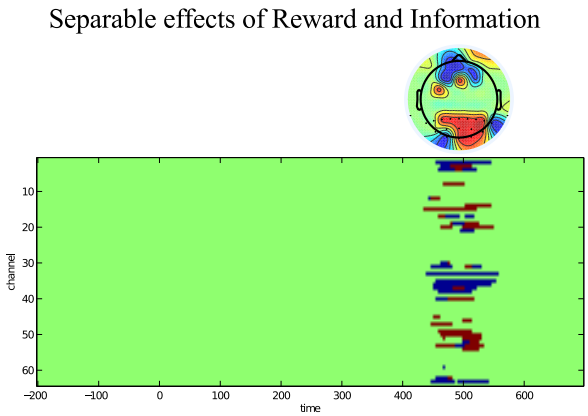

Figure 4.7: Separable effects of expected reward and expected information gain. Beta-weight sign (+1/-1) for cells that passed clustering threshold from the β_1 term in equation 4.1. The y-axis denotes all channels, and the x-axis denotes all time-points. All green cells failed to pass threshold, red cells indicate stronger association with information gain, and blue cells indicate a stronger association with expected reward. The scalp map illustrates the scalp topography of the β_2 terms sign proximal to 500ms.

This analysis identified two significant clusters during the 450 – 550ms time range; one anterior and the other posterior. As depicted in Figure 4.7, a sign test of cells with a significant β_1 term reveals negative β_2 weights associated with the frontal cluster, and positive β_2 weights associated with the posterior cluster. In keeping with the ERP analyses previously discussed, this result suggests that voltage changes at posterior sites exhibits a greater association with expected information gain. This analysis also suggests that voltage change at anterior sites is more tightly associated with expected reward. However, this result must be interpreted with caution since participants tended to spend more time in a state of relatively low uncertainty, potentially biasing the a subjective value toward an expected reward computation. Indeed, our ERP analyses suggested that frontal sites tracked a value more akin to subject value (not just expected reward).

To investigate effects of expected reward and information gain further, an orthogonalization approach was employed to identifying signal uniquely associated with terms of interest. To probe for signal uniquely associated with expected reward, expected information gain was entered as the first regressor, and expected reward

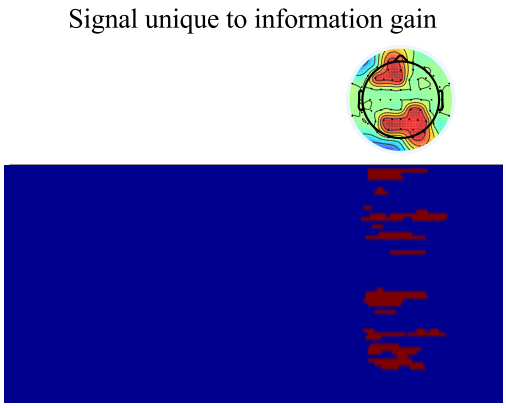

Figure 4.8: Variance uniquely associated with information gain. Clusters of cells that passes significance testing(red) for the orthogonalized information gain term. The scalp map illustrates the scalp topography of the beta-weight terms proximal to 500ms

orthogonalized to information gain (residual variance in expected reward after regressing it against expected information gain) as the second. An investigation of orthogonalized expected reward term revealed no significant effects, suggesting that neither the anterior nor the posterior effects were uniquely associated with expected reward. The same strategy was used to probe for signal uniquely associated with expected information gain; however, this model included expected reward as the first regressor, and expected information gain orthogonalized to expected reward (residual variance in expected information after regressing it against expected reward) as the second. As Figure 4.8 shows, both posterior and anterior clusters share unique variance with expected information gain.

4.4 Discussion

As expected, participants preferred the most rewarding stimulus when they were fairly certain of the deck in play (hA when Deck A was in play, and hB when Deck B was in play). However, some participants shifted their preference, favoring the Info option to a greater degree as their uncertainty as to which deck was in play grew. It's

not clear exactly why approximately half of the study's participants did not leverage the Info card (as the majority did in Experiments 1 & 2). However, it's possible that the extensive training regime or the fact that information might not necessarily be exploitable in the near future (owing to a random card being dealt on each trial) were to blame. Anecdotally, post-task surveys showed that most participants that did not use the Info card had employed a 'never pick the low reward card' rule after training. Nonetheless, an analysis of participants that did use the Info card (the Probe group), revealed that they preferred the Info card as their uncertainty grew.

An analysis of the ongoing EEG proximal to stimulus presentation revealed no effect of expected value during the training phase. However, during the testing phase, when the task was more challenging, the N2 component was modulated in line with the expected value of choice. This effect mirrors previous results linking front-central ERP components in that time range to cognitive control (Yeung et al., 2004; Holroyd & Coles, 2002; Folstein & Van Petten, 2008). The results reported here are in line with the reinforcement learning hypothesis of the FRN, where stimulus-locked N2 modulation reflects the degree to which the stimulus is better or worse than expected (Holroyd & Coles, 2002). Indeed, the results reported here show greater negative deflection of the N2 when either the Low-Reward or the Info card were presented, which are both less rewarding than the High-Reward stimulus. However, this results could also be considered to be in line with the conflict monitoring theory of N2 modulation, where increased response conflict drives greater negative N2 deflection (Yeung et al., 2004). We might consider the High-Reward option to offer the most obvious response, whereas the Low-Reward and Info options offer an expected value closer to the 50/50 gamble, making the choice more difficult. Nevertheless, N2 modulation appears to dichotomize the available options in line with preference for them.

A late-wave modulation of the EEG was also identified using both standard ERP analyses and novel applications of model-based multiple regression. These analyses revealed an anterior component of the EEG, emerging 450–550ms following stimulus onset, that appears to reflect a stimuli's subjective value. When uncertainty was low (i.e. the deck in play was relatively well known), this frontal signal reflected the stimuli's expected reward. However, when uncertainty was high (i.e. the deck in play was not known), the signal at these same sites no longer reflected expected reward; rather, it appeared to reflected something akin to expected uncertainty-reduction, or expected future reward. Recent work has identified these same sites as reflecting subjective value (Harris et al., 2011), which has been linked to value computations associated with the vmPFC (Larsen & O'Doherty, 2014). What is not clear is whether, under conditions of high uncertainty, this frontal signal is associated with information gain *per se*, or whether this signal emerges from a computation of expected future rewards since information pertaining to the deck in play is likely to deliver greater reward in the future.

Coinciding with this frontal late-wave modulation, a posterior signal tracking expected uncertainty-reduction was also identified when uncertainty was high, which to my knowledge, is the first report of such a signal. A model-based analysis of these late-wave dynamics suggests that the posterior signal is more closely tied to expected information gain. Conversely, the frontal signal appeared to vary in accordance with both expected information, and expected reward, suggesting it is more closely tied to subjective value (where value is derived according to an individual's desired ends).

Chapter Five

Conclusion

The exploration/exploitation tradeoff is a fundamental challenge facing decision making agents. This tradeoff is most often described in a stationary environment where a policy can be established and followed once exploration has sufficiently reduced uncertainty regarding action outcomes and/or state transition. While questions remain regarding efficient exploration/exploitation boundary detection in a stable environment, the issue is complicated considerably if we consider unstable environments. In a volatile world, current knowledge may no longer apply in the future, making it challenging to quantify the value of accruing said knowledge.

However, recent work has demonstrated that the explore/exploit tradeoff can be tackled when the environment's volatility is simultaneously estimated and incorporated into the action utility function, and the human brain appears to track similar signals (Behrens & Woolrich, 2007). Humans also appear to be sensitive to the horizon with which information can be applied (Wilson et al., 2014). In an elegant set of studies, Wilson et al. (2014) demonstrate that human participants use a mixture of directed (uncertainty-reducing actions) and undirected (random choice) exploration, and the degree to which people were willing to explore depended on the horizon of actions to which information gained might be applied. And yet, while we are making gains with respect to the factors that drive human exploration, we have yet to pinpoint characteristics of the cognitive processes guiding action valuation in a volatile environment.

Using a volatile partially observable environment, human participants were shown to employ a directed uncertainty-reducing action selection strategy. In Experiment 1, participants were found to seek information at the expense of reward as their uncertainty regarding the latent state of the environment grew, and they leveraged that information to exploit their environment more effectively. In Experiment 2 participants were shown to be sensitive to the cost of information, with participants

demanding higher levels of uncertainty before seeking more expensive information. In Experiment 3 participants were willing to forgo reward in favor of latent state information even in an environment where the latent state had no influence over the optimal policy.

A model-based investigation of the task structure and its various parameterizations focused on two models. The first, a forward Bayesian inference model, identifies the Bayes optimal policy by considering the expected value of current and future actions in potential future belief states. The second, a heuristic uncertainty-reduction model, assigns an expected uncertainty-reduction bonus to the expected reward value of each action. Both models were found to value latent state information in line with uncertainty regarding that variable, increasing the the utility ascribed to informative actions as belief in possible states of the world broadened. Both models, like human participants, were sensitive to the cost of information relative to potential rewards, demanding higher degrees of uncertainty before pursuing more costly information. However, the models differed in their approach to an environment where the optimal policy was independent of the latent state. The forward Bayesian model performed as one would expected from a Bayes optimal model. It found no reason to select anything other than the High-Reward card irrespective of the belief state. However, mirroring human performance, the uncertainty-reduction model could be tuned such that latent state information was valued as a function of both uncertainty and information cost.

Although it's unclear as to exactly why human participants value latent state information in an environment where the optimal policy is independent of the state, at least two options seem plausible. One possibility is that human participants don't employ an optimal policy; rather, they opt for a probability matching strategy. Although the most rewarding option was the same across the entirety of belief space,

the probability of reward shifted according to the deck in play. Knowledge of the latent variable's value is required to determine the probability of reward for each action, and as such, latent state information may be valued so as to facilitate something akin to a probability matching response strategy in a sparsely rewarded region of the environment. However, more data will be required to determine if participants are in fact adopting such a strategy.

A second option might be that information valuation comes from a model-free source, making it more a product of habitually driven response than a goal-directed system. Over the course of our lives, more often than not, knowing the state of the world helps us pursue reward and avoid danger. Thus, regions of the brain may learn to prefer states of low uncertainty, and come to value actions that reduce uncertainty (i.e actions that move from a highly uncertainty region of the belief space to a more certain region). A tension can emerge, as captured by the heuristic uncertainty-reduction model, whereby reward and uncertainty-reduction pull the agent in different directions. Together, the expected reward and expected uncertainty-reduction components of the utility function work together to produce behaviour that can appear model-based at first glance (i.e. seeking information due to the advantage reaped in future states), but relies on model-free value computations.

Regardless of the underlying mechanisms driving the valuation of information, the heuristic uncertainty-reduction model is better able to capture human response patterns in the single-policy experiment. This finding motivated a model-based analysis of the EEG data, which found evidence for separable signals related to stimulus value. Approximately 500ms after stimulus onset, a front-central signal associated with stimulus utility was identified. This signal appears to track expected reward value when uncertainty is low, but for those individuals that sought information, this late-wave front-central signal also appears to track expected uncertainty-reduction

during periods of high uncertainty. However, a second posterior signal was found to track expected uncertainty-reduction alone. These finding are more in line with the idea that separable processes or systems are associated with computing reward value and information gain, lending some support to the hypothesis that information is valued in and of itself. However, it should be noted that signal uniquely associated with expected reward value was not identified, leaving room for alternative models of the brain's utility function.

Future work will aim to further demystify the computational characteristics of the brain's valuation of information. One question of interest is whether information value is derived from a mechanism like temporal difference learning, where informative actions accumulate value because they tend to move the agent into more rewarding states, or whether the brain has developed an independent strategy to motivate information seeking behaviour. One such mechanism may be the anxiety associated with uncertainty, which motives one to take action in order to reduce said uncertainty, but without necessarily considering how that information might be leveraged (e.g. performing a particular data analysis even though the results would never be included in the paper). An ongoing study is probing the relationship between propensity to seek information and a response bias referred to as ambiguity aversion (where humans have been found to avoid ambiguity disproportionately with respect to expected reward) and delayed discounting (where the value ascribed to an outcome depends on the delay between the choice and outcome deliver).

Future work will also focus on the model-free/model-based characteristic of information valuation. Recent work supporting the theory that model-based and model-free systems can guide behaviour but the model-based system had greater cognitive resource demands demonstrated that the model-based system can be comprised by taxing the system with a difficult secondary task (Otto et al., 2013). This suggests

that, if information valuation is driven by a model-free system, then it should be preserved when the system is taxed; however, other model-based reasoning should be compromised. Of course, it's possible that information is valued by a model based system (just as the forward Bayesian model does), but further exposing a model-free information valuation system could increase an individuals propensity to seek information when it may not be to their advantage to do so.

In summary, this work has defined the Bayes optimal solution to a partially observable and volatile environment, and has also offered a computationally attractive heuristic approximation using an uncertainty-reduction bonus mechanism that augments an action's expected reward value. The model's were shown to mirror human preferences under various parameterizations of the environment, but only the heuristic model could match the seemingly 'sub-optimal' behaviour exhibited by participants in the single policy environment. EEG signals related to this heuristic model were also identified, with a front-central signal tracking action utility, and a posterior signal tracking expected uncertainty reduction.

Together, I would argue that these results suggest that the brain had adopted a modular action utility computation strategy, with one component focusing on expected reward, and another on uncertainty reduction (and possibly factors as well). Given the computational complexity of the forward Bayesian solution to the task defined here, and keeping in mind that this is only one of possibly infinite task structures encountered in the world, a modular solution may do well to meet the biological constraints pressed upon a living organism. A mix-and-match strategy that approximates optimality lends itself to a more adaptable and reliable system than an optimal solution applied to the wrong problem.

Bibliography

Acuña, D., & Schrater, P. (2010). Structure Learning in Human Sequential Decision-Making. *PLoS computational biology*, *6*(12).

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature neuroscience*, *14*(10), 1338–44.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annual review of neuroscience*, *28*, 403–50.

Aston-Jones, G., Rajkowski, J., & Cohen, J. (1999). Role of locus coeruleus in attention and behavioral flexibility. *Biological psychiatry*, *46*(9), 1309–20.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends in cognitive sciences*, *12*(5), 193–200.

Badre, D., & D'Esposito, M. (2007). Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. *Journal of cognitive neuroscience*, *19*(12), 2082–99.

Badre, D., Doll, B., Long, N., & Frank, M. (2012). Rostrolateral Prefrontal Cortex and Individual Differences in Uncertainty-Driven Exploration. *Neuron*, *73*(3), 595–607.

Baker, T. E., & Holroyd, C. B. (2011). Dissociated roles of the anterior cingulate cortex in reward and conflict processing as revealed by the feedback error-related negativity and N200. *Biological psychology*, *87*(1), 25–34.

Behrens, T., & Woolrich, M. (2007). Learning the value of information in an uncertain world. *Nature ...*, *10*(9), 1214–1221.

Berridge, C. W., & Waterhouse, B. D. (2003). The locus coeruleus-noradrenergic system: modulation of behavioral state and state-dependent cognitive processes. *Brain research. Brain research reviews*, *42*(1), 33–84.

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain research. Brain research reviews*, *28*(3), 309–69.

Björklund, A., & Dunnett, S. B. (2007). Dopamine neuron systems in the brain: an update. *Trends in neurosciences*, *30*(5), 194–202.

Boorman, E. D., Behrens, T. E. J., Woolrich, M. W., & Rushworth, M. F. S. (2009). How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron*, *62*(5), 733–43.

Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature*, *402*(6758), 179–81.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652.

Botvinick, M. M., Niv, Y., & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition*, *113*(3), 262–80.

Braver, T. S., & Cohen, J. D. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working memory. *Control of cognitive processes: Attention and performance XVIII*, (pp. 713–737).

Bromberg-Martin, E. S., & Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, *63*(1), 119–26.

Bromberg-Martin, E. S., Matsumoto, M., & Hikosaka, O. (2010). Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron*, *68*(5), 815–34.

Brown, J. W., & Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science (New York, N.Y.)*, *307*(5712), 1118–21.

Bryant, C., Muggleton, S., Kell, D., Reiser, P., King, R., & Oliver, S. (2001). Combining inductive logic programming, active learning and robotics to discover the function of genes. *Electronic Transactions in Artificial Intelligence*, *5*(B), 1–36.

Bunge, S. a., & Wendelken, C. (2009). Comparing the bird in the hand with the ones in the bush. *Neuron*, *62*(5), 609–11.

Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning.*. John Wiley & Sons, Inc.

Caplin, A., & Dean, M. (2008). Axiomatic methods, dopamine and reward prediction error. *Current opinion in neurobiology*, *18*(2), 197–202.

Carter, C. S., Macdonald, a. M., Botvinick, M., Ross, L. L., Stenger, V. a., Noll, D., & Cohen, J. D. (2000). Parsing executive processes: strategic vs. evaluative functions of the anterior cingulate cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(4), 1944–8.

Cavanagh, J. F., Figueroa, C. M., Cohen, M. X., & Frank, M. J. (2011). Frontal Theta Reflects Uncertainty and Unexpectedness during Exploration and Exploitation. *Cerebral cortex (New York, N.Y. : 1991)*.

Cockburn, J., Collins, A. G. E., & Frank, M. J. (2014). A reinforcement learning mechanism responsible for the valuation of free choice. *Neuron*, *83*(3), 551–7.

Cockburn, J., & Frank, M. J. (2011). Reinforcement learning, conflict monitoring, and cognitive control: An integrative model of cingulate-striatal interactions and the ERN. In R. B. Mars, J. Sallet, M. F. S. Rushworth, & N. Yeung (Eds.) *Neural basis of motivational and cognitive control*, vol. 9, chap. 16, (pp. 311–331). MIT Press.

Collins, A., & Koechlin, E. (2012). Reasoning, learning, and creativity: frontal lobe function and human decision-making. *PLoS biology*, *10*(3).

Collins, A. G. E., Cavanagh, J. F., & Frank, M. J. (2014). Human EEG uncovers latent generalizable rule structure during learning. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *34*(13), 4677–85.

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological review*, *120*(1), 190–229.

Cools, R., Altamirano, L., & D'Esposito, M. (2006). Reversal learning in Parkinson's disease depends on medication status and outcome valence. *Neuropsychologia*, *44*(10), 1663–1673.

Cragg, S. J., & Rice, M. E. (2004). DAncing past the DAT at a DA synapse. *Trends in neurosciences*, *27*(5), 270–7.

Dagan, I., & Engelson, S. P. (1995). Committee-Based Sampling For Training Probabilistic Classifiers. In *In Proceedings of the Twelfth International Conference on Machine Learning*, (pp. 150–157). MORGAN KAUFMANN PUBLISHERS, INC.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–9.

Dayan, P., & Sejnowski, T. J. (1994). TD() converges with probability 1. *Machine Learning*, *14*(3), 295–301.

Dayan, P., & Yu, A. J. (2006). Phasic norepinephrine: a neural interrupt signal for unexpected events. *Network (Bristol, England)*, *17*(4), 335–50.

Debener, S., Ullsperger, M., Siegel, M., Fiehler, K., von Cramon, D. Y., & Engel, A. K. (2005). Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *25*(50), 11730–7.

Delgado, M., & Nystrom, L. (2000). Tracking the Hemodynamic Responses to Reward and Punishment in the Striatum. *Journal of ...*, (pp. 3072–3077).

Doll, B. B., Hutchison, K. E., & Frank, M. J. (2011). Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *31*(16), 6188–98.

Doll, B. B., Jacobs, W. J., Sanfey, A. G., & Frank, M. J. (2009). Instructional control of reinforcement learning: a behavioral and neurocomputational investigation. *Brain research*, *1299*, 74–94.

Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural computation*, *12*(1), 219–45.

Doya, K., Samejima, K., Katagiri, K.-i., & Kawato, M. (2002). Multiple model-based reinforcement learning. *Neural computation*, *14*(6), 1347–69.

Fedorov, V. (1972). *Theory Of Optimal Experiments*. London: Academic press.

Floresco, S. B., West, A. R., Ash, B., Moore, H., & Grace, A. A. (2003). Afferent modulation of dopamine neuron firing differentially regulates tonic and phasic dopamine transmission. *Nature neuroscience*, *6*(9), 968–73.

Folstein, J. R., & Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: a review. *Psychophysiology*, *45*(1), 152–70.

Foote, S. L., & Morrison, J. H. (1987). Extrathalamic modulation of cortical function. *Annual review of neuroscience*, *10*(1), 67–95.

Frank, M. J., & Claus, E. D. (2006). Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review*, *113*(2), 300–326.

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature neuroscience*, *12*(8), 1062–1068.

Frank, M. J., Moustafa, A. a., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(41), 16311–6.

Frank, M. J., Seeberger, L. C., & O'reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science (New York, N.Y.)*, *306*(5703), 1940–3.

Gittins, J. C., & Jones, D. M. (1979). A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika*, *66*(3), 561–565.

Goldman-Rakic, P. S., Leranth, C., Williams, M. S., Mons, N., & Geffard, M. (1989). Dopamine synaptic complex with pyramidal neurons in primate cerebral cortex. *Proc. Natl. Acad. Sci. USA*, *86*, 9015–9019.

Goto, Y., & Grace, A. A. (2008). Limbic and cortical information processing in the nucleus accumbens. *Trends in neurosciences*, *31*(11), 552–8.

Grace, a. a. (2000). The tonic/phasic model of dopamine system regulation and its implications for understanding alcohol and psychostimulant craving. *Addiction (Abingdon, England)*, *95 Suppl 2*(January), S119–28.

Grace, A. A., & Bunney, B. S. (1983). Intracellular and extracellular electrophysiology of nigral dopaminergic neurons. Identification and characterization. *Neuroscience*, *10*(2), 301–315.

Gureckis, T. M., & Love, B. C. (2009). Short-term gains, long-term pains: how cues about state aid learning in dynamic environments. *Cognition*, *113*(3), 293–313.

Hansen, E. (1998). Solving POMDPs by searching in policy space. In *Proceedings of the fourteenth conference on uncertainty in artificial intelligence*, (pp. 211–219).

Harris, A., Adolphs, R., Camerer, C., & Rangel, A. (2011). Dynamic construction of stimulus values in the ventromedial prefrontal cortex. *PloS one*, *6*(6), e21074.

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2006). Banishing the homunculus: making working memory work. *Neuroscience*, *139*(1), 105–118.

Hernández-López, S., Bargas, J., Surmeier, D. J., Reyes, a., & Galarraga, E. (1997). D1 receptor activation enhances evoked discharge in neostriatal medium spiny neurons by modulating an L-type Ca2+ conductance. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *17*(9), 3334–42.

Herrnstein, R. J. (1974). Formal properties of the matching law. *Journal of the experimental analysis of behavior*, *21*(I), 159–164.

Hillman, K. L., & Bilkey, D. K. (2010). Neurons in the Rat Anterior Cingulate Cortex Dynamically Encode Cost-Benefit in a Spatial Decision-Making Task. *Journal of Neuroscience*, *30*(22), 7705–7713.

Hirvonen, M. M., Laakso, A., Nå gren, K., Rinne, J. O., Pohjalainen, T., & Hietala, J. (2009). C957T polymorphism of dopamine D2 receptor gene affects striatal DRD2 in vivo availability by changing the receptor affinity. *Synapse (New York, N.Y.)*, *63*(10), 907–12.

Holroyd, C. B., & Coles, M. G. H. (2002). The Neural Basis of Human Error Processing: Reinforcement Learning, Dopamine, and the Error-Related Negativity. *Psychological Review*, *109*(4), 679 –709.

Holroyd, C. B., Krigolson, O. E., & Lee, S. (2011). Reward positivity elicited by predictive cues. *Neuroreport*, *22*(5), 249–52.

Howard, R. (1966). Information value theory. *... Science and Cybernetics, IEEE Transactions on*, (1), 22–26.

Ito, S., Stuphorn, V., Brown, J. W., & Schall, J. D. (2003). Performance monitoring by the anterior cingulate cortex during saccade countermanding. *Science (New York, N.Y.)*, *302*(5642), 120–2.

Jocham, G., Klein, T. a., & Ullsperger, M. (2011). Dopamine-Mediated Reinforcement Learning Signals in the Striatum and Ventromedial Prefrontal Cortex Underlie Value-Based Choices. *Journal of Neuroscience*, *31*(5), 1606–1613.

Kaelbling, L., Littman, M., & Cassandra, A. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(1-2), 99–134.

Kakade, S., & Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural networks : the official journal of the International Neural Network Society*, *15*(4-6), 549–59.

King, R., Whelan, K., Jones, F., Reiser, P., Bryant, C., Muggleton, S., Kell, D., & Oliver, S. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, *427*(6971), 247–252.

Kobayashi, S., & Schultz, W. (2008). Influence of reward delays on responses of dopamine neurons. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *28*(31), 7837–46.

Koechlin, E. (2007). Anterior prefrontal function and the limits of human decision-making. *Science*, *318*(5850), 594–8.

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science (New York, N.Y.)*, *302*(5648), 1181–5.

Krebs, J., Kacelnik, A., & Taylor, P. (1978). Test of optimal sampling by foraging great tits. *Nature*, *275*(5675), 27–31.

Krigolson, O. E., & Holroyd, C. B. (2007). Predictive information and error processing: the role of medial-frontal cortex during motor control. *Psychophysiology*, *44*(4), 586–95.

Krigolson, O. E., Pierce, L. J., Holroyd, C. B., & Tanaka, J. W. (2011). Learning to become an expert: Reinforcement learning and the acquisition of perceptual expertise. *Annals of Neurosciences*, *18*, 113–114.

Kringelbach, M. L. (2005). The human orbitofrontal cortex: linking reward to hedonic experience. *Nature reviews. Neuroscience*, *6*(September), 691–702.

Kunishio, K., & Haber, S. N. (1994). Primate cingulostriatal projection: limbic striatal versus sensorimotor striatal input. *The Journal of comparative neurology*, *350*(3), 337–56.

Larsen, T., & O'Doherty, J. (2014). Uncovering the spatio-temporal dynamics of value-based decision-making in the human brain: a combined fMRIEEG study. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, (369).

Ledoux, J. (2007). The amygdala. *Current Biology*, *17*(20), 868–874.

Littman, M. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the eleventh international conference on machine learning*, vol. 157, (p. 163).

Loch, J., & Singh, S. (1998). Using Eligibility Traces to Find the Best Memoryless Policy in Partially Observable Markov Decision Processes. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)*, (pp. 323–331).

MacDonald, a. W. (2000). Dissociating the Role of the Dorsolateral Prefrontal and Anterior Cingulate Cortex in Cognitive Control. *Science*, *288*(5472), 1835–1838.

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of experimental psychology. General*, *143*(1), 94–122.

McClure, S. M., Berns, G. S., & Montague, P. R. (2003a). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, *38*(2), 339–46.

McClure, S. M., Daw, N. D., & Read Montague, P. (2003b). A computational substrate for incentive salience. *TRENDS in Neurosciences*, *26*(8), 423–428.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science (New York, N.Y.)*, *306*(5695), 503–7.

McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of mathematical biology*, *5*, 115–133.

Meyer-Lindenberg, A., Kohn, P. D., Kolachana, B., Kippenhan, S., McInerney-Leo, A., Nussbaum, R., Weinberger, D. R., & Berman, K. F. (2005). Midbrain dopamine and prefrontal function in humans: interaction and modulation by COMT genotype. *Nature neuroscience*, *8*(5), 594–6.

Meyer-Lindenberg, A., Straub, R. E., Lipska, B. K., Verchinski, B. A., Goldberg, T., Callicott, J. H., Egan, M. F., Huffaker, S. S., Mattay, V. S., Kolachana, B., & Others (2007). Genetic evidence implicating DARPP-32 in human frontostriatal structure, function, and cognition. *Journal of Clinical Investigation*, *117*(3), 672–682.

Mink, J. (1996). The basal ganglia: Focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, *50*(4), 381–425.

Montague, P., Dayan, P., & Sejnowski, T. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*(5), 1936.

Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, *191*(3), 507–20.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science (New York, N.Y.)*, *304*(5669), 452–4.

Olson, C. R., & Gettner, S. N. (2002). Neuronal activity related to rule and conflict in macaque supplementary eye field. *Physiology & behavior*, *77*(4-5), 663–70.

Orr, H. (2009). Fitness and its role in evolutionary genetics. *Nature Reviews Genetics*, *10*(8), 531–539.

Otto, a. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological science*, *24*, 751–61.

Palminteri, S., Lebreton, M., Worbe, Y., Grabli, D., Hartmann, A., & Pessiglione, M. (2009). Pharmacological modulation of subliminal learning in Parkinson's and Tourette's syndromes. *Proceedings of the National Academy of Sciences*, *106*(45), 19179.

Pavese, N., Evans, A. H., Tai, Y. F., Hotton, G., Brooks, D. J., Lees, A. J., & Piccini, P. (2006). Clinical correlates of levodopa-induced dopamine release in Parkinson disease: a PET study. *Neurology*, *67*(9), 1612.

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS computational biology*, *7*(1), e1001048.

Pedarzani, P., & Storm, J. F. (1995). Dopamine modulates the slow Ca(2+)-activated K+ current IAHP via cyclic AMP-dependent protein kinase in hippocampal neurons. *Journal of neurophysiology*, *74*(6), 2749–53.

Peshkin, L., Meuleau, N., & Kaelbling, L. (2001). Learning Policies with External Memory. *arXiv preprint cs/0103003*.

Pierce, J. R. (2012). *An introduction to information theory: symbols, signals and noise*. Courier Dover Publications.

Reynolds, J. N. J., Hyland, B. I., & Wickens, J. R. (2001). A cellular mechanism of reward-related learning. *Nature*, *413*(6851), 67–70.

Ribas-Fernandes, J. J. F., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y., & Botvinick, M. M. (2011). A neural signature of hierarchical reinforcement learning. *Neuron*, *71*(2), 370–9.

Ridderinkhof, K. R., Ullsperger, M., Crone, E. a., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science (New York, N.Y.)*, *306*(5695), 443–7.

Roesch, M. R., Calu, D. J., & Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature neuroscience*, *10*(12), 1615–1624.

Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and flexible cognitive control: rules without symbols. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(20), 7338–43.

Rutledge, R. B., Dean, M., Caplin, A., & Glimcher, P. W. (2010). Testing the reward prediction error hypothesis with an axiomatic model. *The Journal of Neuroscience*, *30*(40), 13525–13536.

Sajikumar, S., & Frey, J. U. (2004). Late-associativity, synaptic tagging, and the role of dopamine during LTP and LTD. *Neurobiology of learning and memory*, *82*(1), 12–25.

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science (New York, N.Y.)*, *310*(5752), 1337–40.

Satoh, T., Nakai, S., Sato, T., & Kimura, M. (2003). Correlated coding of motivation and outcome of decision by dopamine neurons. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *23*(30), 9913–23.

Sawamoto, N., Piccini, P., Hotton, G., Pavese, N., Thielemans, K., & Brooks, D. J. (2008). Cognitive deficits and striato-frontal dopamine release in Parkinson's disease. *Brain*, *131*(5), 1294–1302.

Schultz, W. (1998). Predictive Reward Signal of Dopamine Neurons. *Journal of neurophysiology*, *80*(1), 1.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, *36*(2), 241–63.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science (New York, N.Y.)*, *275*(5306), 1593–9.

Seamans, J. K., & Yang, C. R. (2004). The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Progress in neurobiology*, *74*(1), 1–58.

Servan-Schreiber, D., Printz, H., & Cohen, J. (1990). A network model of catecholamine effects: Gain, signal-to-noise ratio, and behavior. *Science*, *249*(4971), 892.

Settles, B. (2009). Active learning literature survey. Tech. Rep. 2, University of Wisconsin-Madison.

Settles, B., Craven, M., & Friedland, L. (2008). Active learning with real annotation costs. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, (pp. 1–10). MORGAN KAUFMANN PUBLISHERS, INC.

Shen, W., Flajolet, M., Greengard, P., & Surmeier, D. J. (2008). Dichotomous dopaminergic control of striatal synaptic plasticity. *Science*, *321*(5890), 848.

Smith, A. D., & Bolam, J. P. (1990). The neural network of the basal ganglia as revealed by the study of synaptic connections of identified neurones. *Trends in neurosciences*, *13*(7), 259–265.

Stipanovich, A., Valjent, E., Matamales, M., Nishi, A., Ahn, J.-H. H., Maroteaux, M., Bertran-Gonzalez, J., Brami-Cherrier, K., Enslen, H., Corbillé, A.-G. G., Filhol, O., Nairn, A. C., Greengard, P., Hervé, D., Girault, J.-A., & Others (2008). A phosphatase cascade by which rewarding stimuli control nucleosomal response. *Nature*, *453*(7197), 879–884.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge Univ Press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*(1-2), 181–211.

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science (New York, N.Y.)*, *307*(5715), 1642–5.

Todd, M., Niv, Y., & Cohen, J. (2009). Learning to use working memory in partially observable environments through dopaminergic reinforcement. *Advances in Neural Information Processing ...*.

Tsai, H.-C., Zhang, F., Adamantidis, A., Stuber, G. D., Bonci, A., de Lecea, L., & Deisseroth, K. (2009). Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science (New York, N.Y.)*, *324*(5930), 1080–4.

Usher, M. (1999). The Role of Locus Coeruleus in the Regulation of Cognitive Performance. *Science*, *283*(5401), 549–554.

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*(3-4), 279–292.

Wickens, J., Begg, A., & Arbuthnott, G. (1996). Dopamine reverses the depression of rat corticostriatal synapses which normally follows high-frequency stimulation of cortex in vitro. *Control*, *70*(1), 1–5.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The Neural Basis of Error Detection: Conflict Monitoring and the Error-Related Negativity. *Psychological Review*, *111*(4), 931–959.

Yoshida, W., & Ishii, S. (2006). Resolution of uncertainty in prefrontal cortex. *Neuron*, *50*(5), 781–9.

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681–92.