MULTIPLE VALUE SYSTEMS FOR ADAPTIVE DECISION-MAKING

MARCOS WILLIAM ECONOMIDES

SUPERVISED BY

RAYMOND JOSEPH DOLAN

SECONDARY SUPERVISOR

KARL FRISTON

WELLCOME TRUST CENTRE FOR NEUROIMAGING

UNIVERSITY COLLEGE LONDON

SUBMITTED FOR THE CONSIDERATION OF A

DOCTORATE IN CLINICAL NEUROSCIENCE

DECLARATION

I, Marcos William Economides, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I can confirm that this has been indicated in the thesis.

ABSTRACT

Values, rewards, uncertainty and risk play a central role in economic and psychological theories of decision-making. Over the past decade, numerous experiments have used neuroimaging techniques to uncover the neural realization of such decision variables while individuals engage in a range of tasks. These have led to a consensus that economic choice involves interplay between multiple systems that enjoy both cooperative and competitive relations. In this thesis, I utilize functional magnetic resonance imaging (fMRI) and computational formalizations of choice to explore how these different brain systems interact to support adaptive decision-making.

In Chapters 4 and 5, I present data from a task in which the inclusion of a dynamic environment required subjects to sometimes approach an option they would normally avoid, or avoid an option they would normally approach. This allowed me to uncover brain systems that track time-varying components of the environment, or immediate reward information, as well as the mechanisms by which these components are integrated. I found that adaptive control in this context involves downstream integration, via functional coupling, of distinct decision components that are computed in separate, often widespread, networks. Yet, choice variables represented in the striatum may in some cases be resistant to modulation, contributing to maladaptive behaviour.

In Chapter 6, I investigate whether task training alters the way in which these different value systems manifest in choice; or more broadly, whether value computations in the brain adapt as humans become more proficient at internalizing models of the world. To address this, I trained subjects on a value-guided decision-making task for 3 consecutive days. The data are suggestive of a shift in the implementation of value-guided planning with training, from a

more cumbersome, resource-dependant mechanism, to a more efficient and robust process that remains resistant to attentional load.

ACKNOWLEDGEMENTS

First and foremost I would like to thank my primary supervisor, Raymond J Dolan, for supporting and entrusting me with the freedom to pursue my own research interests, and for the unprecedented experience that comes with being a part of the esteemed Dolan laboratory. I also thank my secondary supervisor, Karl Friston, for the generous donation of his time and for invaluable guidance during his Monday morning meetings. Further, I owe a huge debt of gratitude to Marc Guitart-Masip and Zeb Kurth-Nelson, who served as mentors and collaborators during my PhD, and without whom none of the work presented in this thesis would have been possible. I thank them not only for instigating countless intellectual discussions, but also for their time, patience, and friendship. I am also grateful to the entire Dolan research group (and to the wider community within the Wellcome Trust Centre for Neuroimaging) for contributing to the overwhelmingly harmonious and collaborative nature that exists amongst the fellows, and with special thanks to Robb Rutledge, Peter Smittenaar, Joel Winston, and Michael Moutoussis for helpful dialogues and advice. Finally I wish to thank Peter Dayan at the Gatsby Computational Neuroscience Unit, who was generous enough to provide comments and theoretical support throughout my PhD.

This work was funded by a 4-year studentship from the Medical Research Council & Brain Research Trust.

CONTENTS

1. INTRODUCTION

2. LITERATURE REVIEW

3. METHODS

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

LIST OF TABLES

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 1

INTRODUCTION

1.1 Introducing neuroeconomics

Humans and other animals are continuously required to make choices between alternative options or courses of action. Neuroeconomics is the study of underlying neurobiological processes that support such decision-making. Importantly, this field draws on economics, psychology, neuroscience and computational modelling, all of which are required for a full understanding of human behaviour. Neuroeconomics is typically studied in the context of behavioural tasks where subjects are provided knowledge of (or must learn about) a set of reinforcers and task contingencies that allow them to maximize rewards earned and minimize punishments or losses.

To date, there is a broad consensus, supported by an array of both human and animal experiments, that choice involves assigning a 'value' to potential alternatives that compete such that the option with the highest expected value can be chosen (Basten, Biele, Heekeren, & Fiebach, 2010; Hunt et al., 2012; Kennerley, Dahmubed, Lara, & Wallis, 2009; D. Lee, Seo, & Jung, 2012; Padoa-Schioppa, 2011; Plassmann, O'Doherty, & Rangel, 2007; Rangel & Hare, 2010; Roesch, Calu, & Schoenbaum, 2007; Schultz, 2000; Strait, Blanchard, & Hayden, 2014; Wunderlich, Rangel, & O'Doherty, 2009). This process can be subdivided into five arbitrary stages, each equally pertinent for optimal decision-making and each evoking distinct computations (see Figure 1.1, p. 11). Although the precise organisation of these stages is still debated, they provide a useful breakdown of the decision-making process into separate components that can be investigated in turn. In the following section I introduce these stages and include a brief synopsis of our understanding of the underlying neurobiology. I then summarize how the work presented in this thesis contributes to that knowledge.

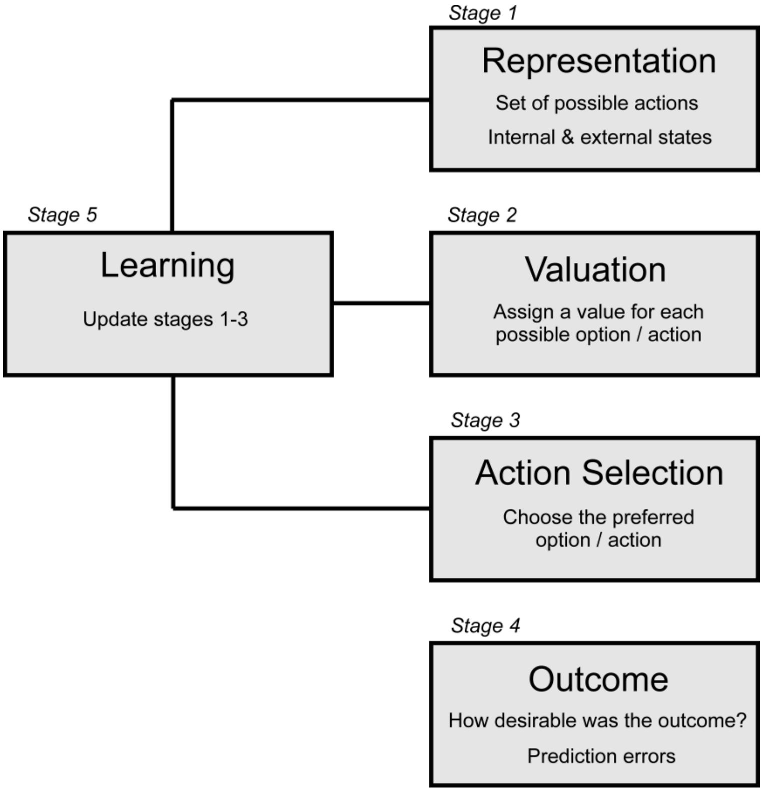


Figure 1.1 Computations involved in decision-making; adapted from (Rangel, Camerer, & Montague, 2008). In order to initiate a decision, an agent must first identify and represent their internal state, the external state of the world, and the possible set of actions available. Next, a value must be assigned to each of these actions, which are compared so that the action with the highest expected utility can be selected. Once the chosen action is executed, the agent can then assess the desirability of the outcome. Any discrepancy between the expected and received outcome is used to inform future choice through learning.

1.2 The stages of value-guided decision-making

1.2.1 Stage 1: Representation

An agent wishing to make decisions within a dynamic environment must identify (and represent) a number of key variables that form an integral part of the decision-making process. First, the current set of internal states or motivations must be recognized. For example, an animal may assign a higher 'value' to water (and indeed experience it as intrinsically more rewarding) when in a thirsty state as opposed to a hungry state. Similarly, animals are more likely to exert effort for a food reward, such as pressing a lever, when hungry compared to when sated. Previously it has been suggested that internal states drive changes in behaviour through negative feedback mechanisms that aim to regulate homeostasis, or rather, to minimize the difference between the current state and a hypothetical physiological setpoint (Toates, 1986).

While many behavioural characteristics are explained well by this framework, it has received wide criticism. Specifically, behaviours typically associated with discrete motivational states often occur in the absence of a negative feedback signal. For example, consumption of food or drink often precedes or anticipates physiological depletion (Toates, 1986). Further, food that is administered intravenously (and thus lacking any associated sensory properties) does not reinforce behaviours otherwise induced by motivated states. In this regard, rats do not learn to enact a response that results in intragastric feeding, but quickly learn to enact the same response when it results in the normal oral consumption of milk (N. E. Miller & Kessen, 1952). Consequently, alternate models have been proposed which describe a more unified role for classical reinforcement learning and homeostatic regulation, whereby rewards are re-defined as action outcomes that reduce subsequent homeostatic drive (Keramati & Gutkin, 2011).

Relatively little is known about the neurobiology of how changes in internal state influence decision-making. The hypothalamus has long been implicated in homeostasis and in particular the regulation of energy intake (Dietrich & Horvath, 2013). For example, neurons in the arcuate nucleus and ventromedial hypothalamus regulate their activity in response to changes in the levels of metabolic fuels including glucose and fatty acids (Lam, Schwartz, & Rossetti, 2005; Minokoshi et al., 2004). But it is not yet clear whether the hypothalamus has a direct role in regulating decision-making. Interestingly, peripheral hormones that play an important role in the regulation of energy intake and appetite have been shown to act within key decision-making regions in addition to regulating hypothalamic function. For example, leptin administration alters activity levels within the human striatum (Farooqi et al., 2007), and it is thought that ghrelin signalling interacts with the striatal dopamine response in rodents (Narayanan, Guarnieri, & DiLeone, 2010). Yet the mechanism via which this might influence the valuation or subsequent action selection stages remains vague.

It has recently been suggested that tonic dopamine, a neuromodulator that plays a crucial role in action, reward, and arousal, encodes the average reward rate of the environment, and is thus closely linked to motivation (Y. Niv, Daw, & Dayan, 2005). This idea has its origin in a proposal that outcomes tend to have higher utilities in more deprived states, generating a higher average expected reward rate, and that this reward rate plays an important role in determining optimal response times. In brief, normative models predict that when an agent interacts with an environment where the average reward rate is higher, all actions should be performed at a faster rate, regardless of their outcomes, to preclude opportunity costs for future rewards induced by slow responses. The proposal that dopamine is involved in tracking this reward rate is corroborated by evidence that administration of L-DOPA, the precursor to dopamine, exacerbates the relationship between average reward and the vigour with which actions are emitted in humans (Beierholm et al., 2013). Yet, several open questions remain, particularly regarding how internal states interact with different value systems to guide motivated decisions.

In addition to the agent's own internal state, other variables that need to be accounted for include the external state of the world, and the range of different possible options or actions that should be included in a putative value comparison process. For example, a different course of action may be chosen when in a volatile compared to stable environment, or in a high threat compared to low threat condition. Evidence from recent studies implicate prefrontal cortex (in addition to the parietal cortex) as important (Behrens, Woolrich, Walton, & Rushworth, 2007; Glascher, Daw, Dayan, & O'Doherty, 2010; Ide, Shenoy, Yu, & Li, 2013; Kolling, Behrens, Mars, & Rushworth, 2012; Rushworth, Noonan, Boorman, Walton, & Behrens, 2011), and this will be discussed in the following sections. Finally, there is a dearth of knowledge regarding how the brain decides which actions to assign values to at this stage of the decision-making process.

1.2.2 Stage 2: Valuation

Given a representation of both internal and external states, and a set of candidate options or actions, the brain then needs to assign a value to each option so that the option likely to maximize the total expectation of reward and minimize the expectation of punishment can be selected. Much of the literature in both animals and humans has focused on the valuation stage. Correlates of the subjective value of goods have been found in a multitude of brain regions including the dorsolateral prefrontal cortex (Kable & Glimcher, 2007; Plassmann et al., 2007; Sokol-Hessner, Hutcherson, Hare, & Rangel, 2012), anterior cingulate cortex (X. Cai & Padoa-Schioppa, 2012; Kennerley et al., 2009), parietal cortex (Hunt et al., 2012; Platt & Glimcher, 1999), amygdala (Jenison, Rangel, Oya, Kawasaki, & Howard, 2011), posterior cingulate cortex (Jocham et al., 2014; Kable & Glimcher, 2007), and orbitofrontal cortex (Hare, O'Doherty, Camerer, Schultz, & Rangel, 2008; Kennerley, Behrens, & Wallis, 2011;

Padoa-Schioppa & Assad, 2006; Plassmann et al., 2007; Schoenbaum, Takahashi, Liu, & McDannald, 2011), with the two most reproducible regions in humans being the ventromedial prefrontal cortex (vmPFC) and striatum (Bartra, McGuire, & Kable, 2013). In the non-human primate field, vmPFC, medial orbitofrontal cortex (mOFC) and lateral orbitofrontal cortex (LOFC) are considered anatomically and functionally distinct (Bouret & Richmond, 2010; Monosov & Hikosaka, 2012; Noonan et al., 2010; Rich & Wallis, 2014), although far more studies have recorded from OFC than vmPFC. By contrast, in human studies the vmPFC and mOFC are often conflated creating confusion over the appropriate nomenclature. This may partly be due to the poor spatial resolution of fMRI which makes it difficult to define a clear-cut anatomical boundary between these regions. Therefore, from this point on, when using the term “vmPFC” in humans I will consider this to include regions from both vmPFC and mOFC, but not LOFC.

Some accounts posit that valuations in human vmPFC signal the difference in value between chosen and unchosen options (Boorman, Behrens, Woolrich, & Rushworth, 2009; De Martino, Fleming, Garrett, & Dolan, 2013; Serences, 2008), and that vmPFC thus acts as a final value comparator (Hunt et al., 2012; Strait et al., 2014; Wunderlich, Dayan, & Dolan, 2012). By contrast, others argue that the comparison process is resolved elsewhere in the brain (Basten et al., 2010; Morris, Dezfouli, Griffiths, & Balleine, 2014; Wunderlich et al., 2009), and that the outcome is transferred to vmPFC, which encodes the final chosen value (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Hampton, Bossaerts, & O'Doherty, 2006b; Kable & Glimcher, 2007; Wunderlich et al., 2009). A recent experiment has also shown evidence that vmPFC may in fact encode the relative value difference between attended and unattended choice options (Lim, O'Doherty, & Rangel, 2011). Further, a multitude of work in both animals and humans now points towards the existence of at least three dissociable value systems: habitual, goal-directed and Pavlovian (Dayan, 2008). The neural and computational basis of the valuation stage will be reviewed and discussed in detail in Chapter 2.

1.2.3 Stage 3: Action selection

Once the brain has assigned a value to the options under consideration, an action must be initiated so that the agent can select the option deemed to generate the largest expected utility. Until recently, little was known about how the brain achieves this. Theoretical models have emerged from studies of perceptual decision-making that model binary choices in the perceptual domain as a race-to-barrier diffusion process (Heekeren, Marrett, & Ungerleider, 2008). Although a variety of proposed model exist, the general principle is built upon the notion that evidence for each alternative option accumulates over time until one option surpasses a predetermined decision threshold, at which point that option is chosen. Further, it is thought that individual populations of neurons encoding each option inhibit each other such that activity only survives in the eventual winning pool.

Indeed it has been shown these models accurately predict single neuron activity within the parietal cortex in non-human primates during perceptual decision-making (Shadlen & Newsome, 2001), though it has remained unclear whether the same mechanisms apply to value-guided decision-making. A recent study tested this precise hypothesis by investigating the temporal dynamics of valuation signals in local field potentials from magnetoencephalography data (Hunt et al., 2012). Interestingly, the authors found that the ventromedial prefrontal cortex (vmPFC) and posterior parietal cortex (PPC) matched the model predictions accurately, suggesting these regions engage in value comparison, whereas other regions associated with value matched poorly, suggesting they perform alternate computations that do not contribute to selection of an action. Follow-up work has shown that the PPC is more likely to support value comparisons during decisions under time pressure, whereas vmPFC takes on the role of a comparator when decisions are made without time pressure, suggesting that parallel cortical mechanisms may resolve the same choices in differing circumstances (Jocham et al., 2014). Lastly, it is worth noting that recent neurophysiological recordings in animals have corroborated the notion from human neuroimaging experiments that vmPFC compares the value of choice options to enact decisions (Strait et al., 2014).

1.2.4 Stage 4: Outcome

Real-life decisions typically result in an immediate reward or punishment and a complex set of delayed consequences. Often, the outcome conveys meaningful information regarding how "good" the choice that led to it was, which is then used to inform future decision-making. In human fMRI studies, activity in the ventromedial prefrontal cortex (and other regions such as the amygdala (LaBar et al., 2001)) has been shown to correlate with subjective ratings of pleasure at the time of reward delivery for primary rewards (Kringelbach, O'Doherty, Rolls, & Andrews, 2003). Further, it has been shown that these signals subside when the subject is first fed to satiation and thus experiences the outcome as less desirable in the context of food rewards, or images of food (Fuhrer, Zysset, & Stumvoll, 2008; Kringelbach et al., 2003). Other regions implicated in outcome evaluation include the anterior cingulate cortex, which has been shown to activate in response to decision errors both in single neuron recordings and fMRI studies (Braver, Barch, Gray, Molfese, & Snyder, 2001; Ito, Stuphorn, Brown, & Schall, 2003). Further, in non-human primates, neurons in anterior cingulate cortex respond to outcomes in a manner that depends on previous reward history, suggesting a role in the evaluation of choice outcomes (Seo & Lee, 2007). However, whether the ventromedial prefrontal cortex and anterior cingulate cortex make entirely distinct contributions to outcome evaluation, or whether other regions are additionally involved, remains unclear. In addition, since the majority of value-guided decision-making paradigms adopt outcomes that consist of immediate rewards or punishments, little is known about how the brain evaluates long-term consequences.

1.2.5 Stage 5: Learning

In order to improve future decision-making on the basis of outcome evaluation, the brain must update one or more of the representations discussed at stage 1, such that "better" actions can be chosen in the future. This is perhaps best understood and illustrated in the context of the habits system, or model-free reinforcement learning (MF-RL). The theoretical and computational premise underpinning this type of learning is that the brain estimates the difference between the expected value of an outcome and the actual value of the received outcome, a quantity termed a prediction error. This prediction error is used to update the value of the action that led to the observed outcome in a manner proportional to the magnitude of the prediction error (governed by a learning rate). In this context, the agent can learn an approximation of the true value of the action after several rounds of choices and outcomes, and thus optimize behaviour in the face of rewards and punishments (Rescorla & Wagner, 1972).

In general, reinforcement learning models, defined in terms of a Markov decision process, are characterized by:

- a set of environment states, $s \in S$
- a set of actions that transfer the agent between states, $a \in A$
- a matrix that characterizes transitions between states, T
- rewards, r, following state-action transitions in the environment
- a policy, π, that assigns an action to each state (e.g. in accordance with a principle of reward maximization)

The goal of the agent is to learn a value function that predicts the sum of future rewards (from all forthcoming states) expected from a particular action at a particular state in the environment. Note that the value of arriving at a particular state has two components, the immediate reward or payoff associated with that state, and the value associated with the state change itself.

Thus, by exploiting the recursive relationship between successive (and deterministic) states, one can define the value of state s_1 as:

$$V(s_1) = r_1 + \gamma V(s_2)$$

where 0 < γ < 1 captures the discounting of future rewards, r_1 is the immediate reward associated with state 1, and s_2 is the sum of all future rewards associated with reaching that state.

However, if state transitions are instead probabilistic, such that action *a* in s_1 can lead to s_2 or s_3, then the value of s_1 depends on the values of both s_2 and s_3, weighted by the probability of reaching either state. In this context, the value of s_1 can be rewritten as:

$$V(s_1) = r_1 + p(s_2)\gamma V(s_2) + p(s_3)\gamma V(s_3)$$

where $p(s_2)$ and $p(s_3)$ are transition probabilities from states 1 to 2 or 3.

In fact, under the Markov assumption that the previous trajectory to a given state has no bearing on future state transition probabilities or future rewards, the value of any state (under a policy π) can be defined by the Bellman equation (Bellman, 1957):

$$V^{\pi}(s) = \sum_{a} \pi(a|s) \sum_{s'} p(s'|s,a)[R(s,a,s') + \gamma V(s')]$$

This equation can be solved by iteratively updating one state after the other until convergence. This requires knowledge of both *T* (transition functions) and *R* (the set of rewards in the environment) so that the average reward over all actions can be computed. Reinforcement learning methods, most notably temporal difference (TD) learning (R. S. B. Sutton, A. G., 1998), attempt to approximate the Bellman equation without the need for explicit models of the world. To achieve this, we must revisit the notion that successive states retain a recursive relationship, where the value of a given state is equal to the immediate reward and the value of the following state. One can calculate the difference between these quantities and formulate the following update equation:

$$\delta = V(s_n) - (r(s_n) + V(s_{n+1}))$$

δ, the prediction error, is used to update the value of the preceding state:

$$V(s_n) = V(s_n) + \ \alpha\delta$$

where 0 ≤ α ≤1 is a learning rate

Thus, rather than storing all past rewards and performing an average every time it is required, temporal difference learning updates the predicted expectation of reward online and then simply stores, or “caches”, this representation.

Remarkably, extremely reliable neural correlates of this prediction error signal have been found in both animals and humans. The first observations came from midbrain dopamine recordings by Schultz and colleagues in non-human primates (Schultz, Dayan, & Montague, 1997). In this experiment, the researchers showed that single neurons increased their firing rate in response to unexpected rewards, but that after several consecutive outcomes the same neurons fired in response to the cue that predicted the same reward. Further, if the reward was subsequently omitted, the same neurons transiently decreased their firing rate in a manner predicted by a negative prediction error, or the unexpected omittance of a rewarding outcome. Since then, several fMRI experiments in humans have identified prediction error signals in the ventral striatum (Glascher et al., 2010; Hare et al., 2008; J. P.

O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003), which receives input from the dopaminergic cell bodies within the ventral tegmental area of the midbrain, supporting a role for dopamine in habitual or model-free reinforcement learning.

So far I have discussed learning associated with a habitual controller, which does not require knowledge of transition or reward functions, but instead relies on trial-and-error. Somewhat more complex is learning associated with goal-directed or model-based choice. In this scheme, an agent makes choices by searching through a decision tree whereby the consequences of all possible sequences of actions and outcomes are simulated so that the best action at any given state can be chosen (Dayan, 2008). Thus, unlike habits which are retrospective, model-based action is prospective. A number of recent experiments have begun to unravel the neural underpinnings of model-based reinforcement learning. For example, neural correlates of state prediction errors that report discrepancies between an agent’s current model of the world and the observed state transitions resulting from an action have been reported in the intraparietal sulcus and lateral prefrontal cortex (Glascher et al., 2010). Further, a recent report has questioned the classical view that the ventral striatum exclusively supports model-free reinforcement learning by demonstrating that the BOLD signal in this region integrates both model-free and model-based prediction errors signals (Daw, Gershman, Seymour, Dayan, & Dolan, 2011).

There remain a number of open questions. For example, it is unknown whether the habit system can learn through observation without directly experiencing outcomes, or whether it can adopt more sophisticated computations with task training. Moreover it is unclear whether the habit system can learn adequately when the delay between action and outcome is temporally extended. Finally, the precise mechanism that supports a model-based system in learning action-outcome and outcome-value representations (that are needed to infer action values) is yet to be elucidated.

1.3 Summary of the work presented in this thesis

Having described the key stages required for value-guided decision-making I now give a brief summary of the experiments reported in this thesis and how they address some of the open questions in the literature.

In daily life humans make adaptive decisions by taking into account the current state of the external world and the future consequences of actions with regards to future states. Together these processes encompass portions of both stage 1 (representing the state of the external world) and stage 2 (evaluating the immediate and delayed consequences of each possible action). Importantly, a change in the current state of the world typically accompanies a change in the value of the options under consideration, which may drive subsequent switches in choice. In my first experiment, I characterized the computational and neural underpinnings supporting these aspects of decision-making. I used a novel sequential decision-making task where actions could bestow immediate rewards but also had delayed consequences. Subjects had to take into account both components when making choices and were often required to switch their responses based on the changing delayed consequences. This allowed me to investigate how the brain tracked changes in the environment to calculate the future costs of acting, and how these computations were integrated with representations of stimulus value. I used computational modelling in combination with a parametric fMRI design.

In my second experiment I used a variant of the same sequential decision-making task to explore how subjects arbitrate between different components of value when they endorse opposing actions. In this version of the task, acting for a large immediate reward could have detrimental future consequences by diminishing the availability of reward later in a trial. Thus, in order to maximize monetary gain across a trial, subjects had to sometimes reject large immediate rewards, creating an incentive for self-control. Thus, I again focused on stage 2 of the decision-making sequence, but with an emphasis on whether and how distinct value systems contribute to the valuation process. Importantly, the task was carefully designed to decorrelate the immediate reward associated with a stimulus from its overall value. Using computational modelling and fMRI, this allowed me to explore the neural correlates of each value component and their manifestation in behaviour. Importantly, I was able to address an ongoing debate in the literature where on the one hand choice is thought to be governed by a single common value system or alternatively by multiple value systems (Hare, Camerer, & Rangel, 2009; Kable & Glimcher, 2007; McClure, Laibson, Loewenstein, & Cohen, 2004).

In my third experiment, I again investigated how multiple value systems contribute to decision-making, but with an emphasis on the outcome evaluation and learning stages of the decision-making process (see stages 4 and 5). A prominent and contemporary account of learning proposes that one system, the model-based (MB) system, supports goals, whereas a second system, the model-free (MF) system, supports habits (Dolan & Dayan, 2013). It is thought that these systems act in parallel but it has also been shown that MB reasoning is impaired when prefrontal cortex function is disrupted or when working memory demands increase (Otto, Gershman, Markman, & Daw, 2013; Smittenaar, FitzGerald, Romei, Wright, & Dolan, 2013). It is well-established that task training, particularly in the domain of working memory, can increase successive task performance (Jaeggi, Buschkuehl, Jonides, & Shah, 2011). By contrast, it remains unexplored whether frequent task performance alters the degree to which model-free or model-based control is dominant in choice in a similar manner. Here, I used a previously established multi-stage decision paradigm where model-based and model-free strategies make qualitatively different predictions about how outcomes are used to inform future actions. I trained subjects to perform this task for 3 days so as to assess the impact of choice habituation on model-based or model-free decision-making. I hypothesized that following training subjects would adopt a model-based strategy even under high working memory load, suggestive of a change in the mechanism by which model-based calculations are implemented with increasing task exposure.

While a wealth of value-guided decision-making paradigms have characterized neural representations of stimulus value, few have used sequential paradigms. The work presented here therefore offers valuable insight into our understanding of how the brain signals long-term components of value and how this contributes to adaptive decision-making. I discuss my findings in the broader context of decision neuroscience and the implications for both healthy and maladaptive decision processes.

CHAPTER 2

LITERATURE REVIEW

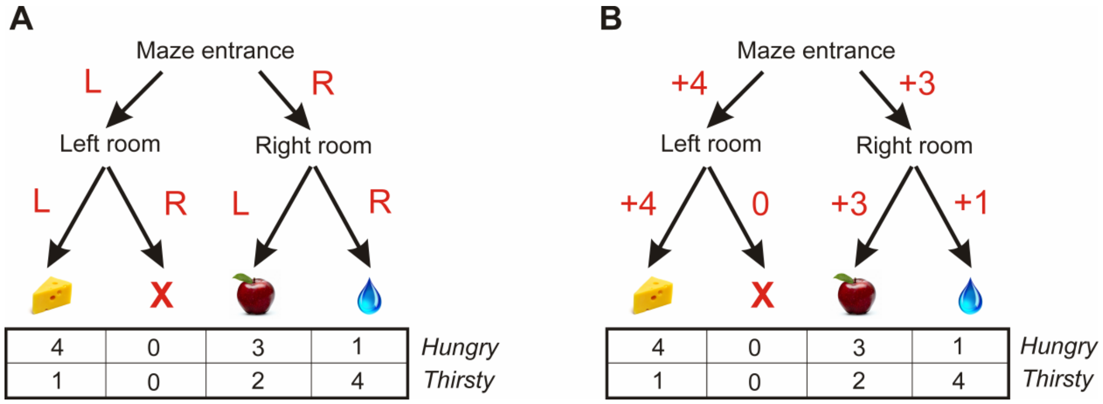
2.1 Multiple value systems

Here I return to the valuation stage of the decision-making process and provide a more detailed overview of the literature. As previously mentioned, there is now ample evidence pointing towards the existence of multiple valuation systems for decision-making, and these are outlined as follows.

2.1.1 Goal-directed control

In his seminal paper (Tolman, 1948), Tolman argued that animals negotiating a maze to harvest rewards develop a cognitive map of the environment (O'Keefe & Nadel, 1978), that enables (through mental search) a pre-emptive evaluation of the best course of action. Contemporary accounts refer to this type of behaviour as goal-directed (Balleine & Dickinson, 1998; Dolan & Dayan, 2013; Rangel & Hare, 2010). Formally, the goal-directed system assigns values to actions by computing action-outcome contingencies, and evaluating the rewards associated with each respective option. Goal-directed behaviour is thus computationally synonymous with model-based choice (Daw et al., 2011; Dayan, 2008). Because the goal-directed system requires forward planning, it is computationally demanding. That is, in a sequential environment, goal-directed actions need not only consider the immediate consequences of an action but the total expected reward that is likely to result from all ensuing states and actions. However, it is this very property that makes the goal-directed system highly flexible. The nature of prospective planning affords an intrinsic revaluation of actions following a change in environmental contingencies without the need for experiencing new outcomes.

Consider, for example, the scenario in Figure 2.1A (p.27), of an animal navigating around the branches of a maze with the aim of locating rewards. In this example, the set of possible outcomes following a sequence of two consecutive actions includes a block of cheese, an apple, a drink of water, or no reward (X). Assuming the animal is placed at the entrance of the maze, the aim is to perform the sequence of actions that will maximize the total expectation of reward overall. The challenge here (even in this relatively simple example), is that maximizing total reward requires planning through both actions in the sequence, rather than just one action-outcome. Presuming the animal knows the layout of the maze and the locations of the respective rewards they can simply plan through each route and choose the pair of actions that will yield the largest reward under a given motivational state. Suppose the animal is routinely placed in the maze in a hungry state. The optimal action sequence in this case would be L-L which results in a block of cheese, their preferred food outcome (+4 utils of reward). In contrast, suppose the animal is first fed to satiety and then placed in the maze. Here the optimal action course would be R-R resulting in water (+4 utils of reward). This is the essence of goal-directed control. The set of outcomes and the corresponding actions that yield them are explicitly represented, allowing online calculation of the best action under the present motivational state.



4	0	3	1	*Hungry*
1	0	2	4	*Thirsty*

4	0	3	1	*Hungry*
1	0	2	4	*Thirsty*

Figure 2.1 An illustration of goal-directed versus habitual valuation in a maze-like paradigm; adapted from (Y. Niv, Joel, & Dayan, 2006). The goal of the animal is to acquire the outcome that is most valuable given the current motivation state (in this example hungry or thirsty), by navigating from an initial to a terminal state via an intermediate state. (**A**) Assuming the animal is familiar with the layout of the maze, it knows, under a goal-directed system, that choosing L-L will lead to the block of cheese, the most desirable outcome if hungry (+4 utils of value), but that it should choose R-R if thirsty, leading to water (+4 utils of value). In essence, the animal can use knowledge of action-outcome contingencies in the maze to plan through each route and assign a value to each action pair. (**B**) Supposing the animal receives daily training in the maze and experiences all actions and outcomes on multiple occasions, the animal can then calculate an expected value for each action in each state, and cache this value for future use. Thus, if the animal receives training in the hungry state, then choosing L at the maze entrance acquires a value of +4 (based on the expectation of cheese at the terminal state), whereas choosing R acquires a value of +3 (based on the expectation of an apple at the terminal state). Thus the decision to go L at the maze entrance becomes habitual. While computationally efficient, this can be maladaptive under a change in motivation state.

2.1.2 Habitual control

The second value system, the habitual system, is thought to be synonymous with model-free decision-making (Daw, Niv, & Dayan, 2005; Dolan & Dayan, 2013; Yin & Knowlton, 2006). The habit system learns to assign values to stimulus-response (S-R) associations, without explicitly representing outcomes. Thus, actions that lead to rewards or the avoidance of punishments are repeated, whereas those that lead to the converse are extinguished. S-R accounts of learning date back to experiments conducted by Thorndike (Thorndike, 1911), who concluded that animals can learn instrumental contingencies that are "blind to changes in the environment". Unlike the goal-directed system, the habit system learns the value of actions slowly through trial-and-error. Further, while the habit system may assign the same value to an action as the goal-directed system in a stable environment, actions may be incorrectly valued following a change in environmental contingency, until the correct value has again been learnt.

Let us revisit the maze-task in Figure 2.1B (p. 27). Supposing both the environment (the locations of the rewards) and the motivational state (hungry rather than thirsty) of the animal are constant over a prolonged period of time, the animal can learn that the action sequence L-L yields the preferred result. Thus, the animal can store or 'cache' a value for going L at the maze entrance, which represents the total expectation of reward at the end of the sequence. This is because the animal knows from experience that going L from the left room results in the cheese reward, and so by extension, going L at the maze entrance has an eventual expected value of +4 utils of reward. By contrast, the best possible outcome from choosing R at the maze entrance is the apple, resulting in +3 utils of reward. Thus, the decision to choose L at the maze entrance becomes habitual. The animal no longer has to plan through each route, and ceases to represent each individual action-outcome.

In a stable environment, the habit system is computationally efficient and highly advantageous. However, suppose that on one occasion, the animal is fed to satiety before entering the maze, devaluing the block of cheese from +4 to +1 utils of reward. The preferred outcome is now water (+4 utils of reward), which requires choosing R-R. Yet under habitual control, the animal will draw on cached values which incorrectly motivate choosing L-L. Thus, the habit system relies on re-learning the set of Q-values (the value of each action in each state) following a change in motivational state, and is less intrinsically flexible.

2.1.3 Pavlovian control

The third value system is the Pavlovian system. This system assigns values to a discrete set of actions that are evolutionarily advantageous, such as approaching stimuli that are predictive of water or food, and avoiding stimuli that are predictive of threat (Dickinson, 1980; Huys et al., 2011). However, in some cases these innate or 'hard-wired' responses can be maladaptive (Guitart-Masip, Duzel, Dolan, & Dayan, 2014). For example, in a famous experiment by Hershberger, a food tray was made to recede at twice the speed with which the animal approached, but would move towards the animal at twice the speed if the animal were instead to recede (Hershberger, 1986). The animals were unable to learn to overcome the prepotent drive to approach the tray for food, presumably demonstrating the influence of a Pavlovian bias. While it is thought that the Pavlovian system largely controls responses to a narrow set of predetermined stimuli, evidence indicates that through sufficient training, animals can learn to deploy Pavlovian responses to relatively novel stimuli. It is currently unknown whether there is a simple common Pavlovian system, or multiple systems that interact during choice.

2.1.4 Other value systems

It is worth noting that a number of researchers have reported evidence in favour of multiple value systems operating in parallel in a context where it is not clear which of the three established value systems (if any) they map on to. For example, McClure and colleagues demonstrated that a different 'value' network is active when people choose small and immediate rewards compared to large but delayed rewards in a temporal discounting task (McClure et al., 2004). One might predict that a preference for immediate rewards would be subserved by brain regions previously shown to support habitual behaviour (Yin & Knowlton, 2006) whereas a preference for delayed rewards would be subserved by brain regions previous shown to support goal-directed behaviour (Balleine & Dickinson, 1998; Morris et al., 2014; Rangel & Hare, 2010), yet the pattern of neural activations reported in the study did not necessarily support this account. Further, other researchers have noted that there is extensive overlap between the neural correlates of model-free and model-based valuation and that these systems may be more integrated than previously hypothesized (Doll, Simon, & Daw, 2012). For example, in rodents it has been shown that dorsolateral striatum is required for habitual control (Yin & Knowlton, 2006), whereas dorsomedial striatum is required for goal-directed control (Yin, Knowlton, & Balleine, 2005; Yin, Ostlund, Knowlton, & Balleine, 2005). Although these neighbouring regions are structurally similar, the expression of divergent functional computations is not surprising given their distinct anatomical connections (see *Striatum*, p 39).

2.1.5 Arbitrating between the different systems

A natural question that follows the preceding discussion is why are multiple value systems needed, and how does one arbitrate between the different systems when they promote divergent actions? I have already discussed that the model-based system is most useful in changing environments where one can update actions without the need for experiencing outcomes. However, the difficulty of a tree search increases exponentially with increasing depth, making several model-based computations near-intractable. By contrast, the model-free system is highly efficient in stable environments where only a limited set of values are cached, but can induce erroneous decisions in non-stable environments. Similarly, the Pavlovian system hard-codes a number of evolutionarily advantageous responses, but is limited to a narrow set of actions and can be maladaptive in complex environments. Given that each system possesses a unique set of advantages and disadvantages, it seems intuitive that being able to utilize all three would produce complimentary results.

However, this still does not address how these systems interact during decision-making. One influential theory by Daw and colleagues proposed that the model-based and model-free systems trade-off according to two forms of uncertainty - knowledge and computational noise - which are tracked by each system (Daw et al., 2005). Then, the relative contribution of each system during choice should be directly proportional to their respective levels of uncertainty, with the system demonstrating the least uncertainty presiding. Very recently, this hypothesis was formally tested by Lee and colleagues using a clever behavioural paradigm in which on different trials the structure of the task favoured control by the model-based or model-free system respectively (S. W. Lee, Shimojo, & O'Doherty, 2014). The authors demonstrated that the inferior lateral prefrontal cortex and frontopolar cortex encode the reliability (or uncertainty) of each system in addition to the relative comparison between the two quantities. Further, they reported changes in functional connectivity between these areas and regions within the striatum that support model-free control. Thus, model-free processing could be subject to top-down control by the prefrontal cortex when the output of a model-based system is deemed sufficiently reliable.

2.1.6 Summary

In summary, there are thought to be at least three separate value systems for guiding choice distinguished as goal-directed (model-based), habitual (model-free) and Pavlovian. Broadly, goal-directed choice is thought to be subserved by the prefrontal cortex (and subregions of the striatum) while habitual choice is subserved by other regions within the striatum. The neural basis of Pavlovian valuation remains less well-understood. In section 2.3 I will review the neural evidence and validity of this conjectural distinction.

2.2 Anatomy of the prefrontal cortex and striatum

Today, there is a rich body of work in both humans and animals that implicates distinct regions of the prefrontal cortex and striatum in different facets of value-guided decision-making, including goal-directed and habitual control (Balleine & Dickinson, 1998; Balleine & O'Doherty, 2010; Dolan & Dayan, 2013; Rangel & Hare, 2010; Valentin, Dickinson, & O'Doherty, 2007; Yin & Knowlton, 2006). In the following section, I will provide a brief (and admittedly highly simplified) anatomical description of these regions, including their respective delineations and projections.

Crudely, the prefrontal cortex (PFC) can be anatomically divided into dorsolateral, ventrolateral, dorsomedial, ventromedial, frontopolar and orbitofrontal components, each with distinct anatomical projections and functions (Badre, 2008; Koechlin, Ody, & Kouneiher, 2003; E. K. Miller & Cohen, 2001; Petrides, 2005; E. E. Smith & Jonides, 1999; Tanji & Hoshi, 2008; Walker, 1940; Wise, 2008). Most knowledge about the anatomical connections of prefrontal cortex comes from experimental work in monkeys, whereas the functional computations subserved by these regions comes from both single unit recordings in monkeys and human neuroimaging studies. For this reason it is essential to have architectonic maps that are based on the application of similar criteria in the delineation of areas in both the human and monkey cerebral cortex. One example of a contemporary numerical scheme used to delineate regions of PFC that is comparable between species is shown in Figure 2.2, where panel A shows the human brain and panel B the macaque monkey brain (Petrides & Pandya, 1999). This work by Petrides and Pandya builds on historical cytoarchitectonic maps in the human and monkey brains by Brodmann (Brodmann, 1909) and then Economo and Koskinas (Economo & Koskinas, 1925), and later on by Walker (Walker, 1940).

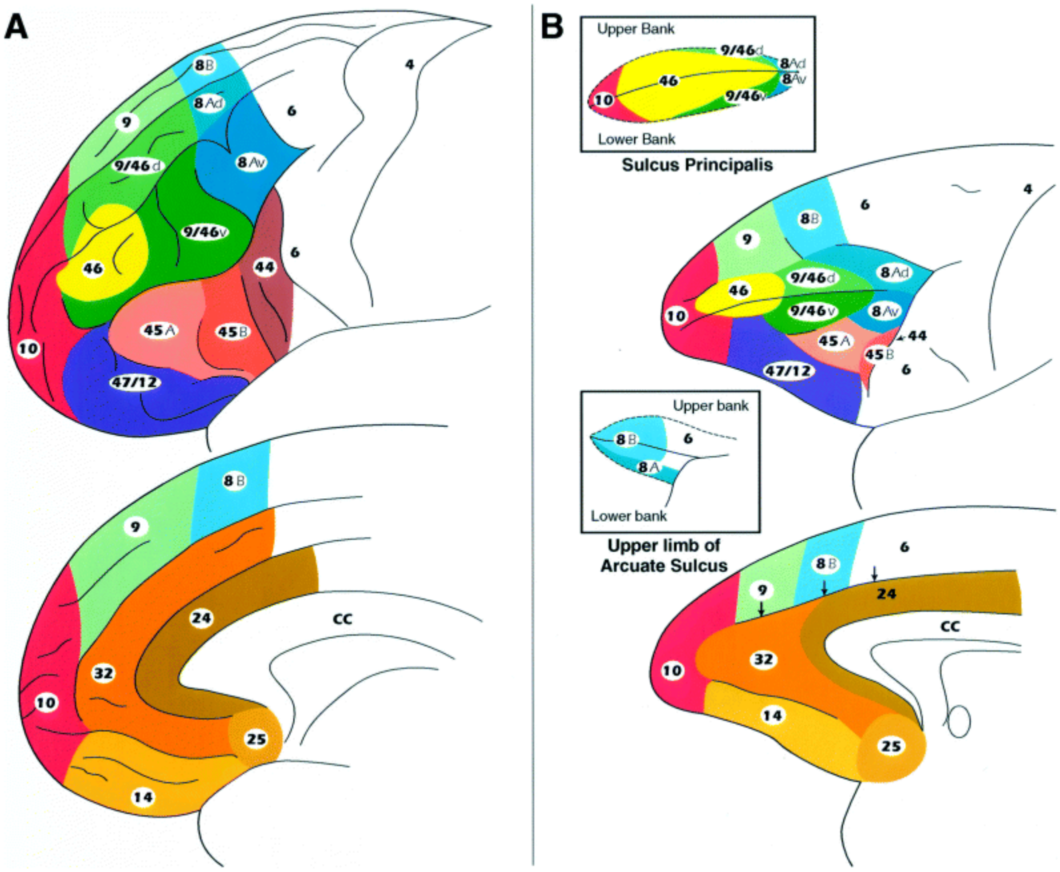


Figure 2.2 Cytoarchitectonic maps of the lateral and medial surfaces of the frontal lobe; taken from (Petrides & Pandya, 1999). (**A**) Human brain, and (**B**) Macaque monkey brain. Note that the present numerical scheme provides a basis for a closer integration of findings from functional neuroimaging studies in human subjects with experimental work in the monkey.

2.2.1 Lateral prefrontal cortex

In lateral prefrontal cortex (LPFC), areas 9, 46 and 9/46 in Figure 2.2 constitute the dorsolateral prefrontal cortex (DLPFC) while areas 44 and 45 constitute the ventrolateral prefrontal cortex (VLPFC). The basic architecture and anatomical connectivity of these regions is thought to be similar in the human and macaque brains (Petrides & Pandya, 1999, 2002) (for a detailed cross-species review see (Wise, 2008)).

DLPFC and VLPFC are often viewed as part of two distinct, large-scale networks within the PFC respectively. DLPFC is part of a mediodorsal network originating from the periallocortex in the medial PFC, whereas VLPFC is part of an orbitoventral network originating from the periallocortex in the orbital PFC (Tanji & Hoshi, 2008). The orbitoventral network is characterized by multiple sensory inputs, including visual, auditory, somatosensory, gustatory, and olfactory (E. K. Miller & Cohen, 2001). This pattern of connections suggests that this network plays a major role in receiving multiple sensory signals to retrieve and integrate necessary information. In contrast, the mediodorsal network receives inputs from multimodal areas in the temporal cortex or auditory areas in the superior temporal gyrus, and from the parvocellular lateral part of the mediodorsal thalamic nucleus (Tanji & Hoshi, 2008). This suggests that the dorsal network receives signals that are already processed and are multimodal in nature. Thus the dorsal and ventral parts of the LPFC seem to process information based on distinct inputs. Additionally, there are extensive interconnections between the two networks (Barbas & Pandya, 1989; Petrides & Pandya, 2002).

The DLPFC has preferential connections to motor structures including the supplementary motor area (SMA), pre-supplementary motor area (pre-SMA), the rostral cingulate, the premotor cortex, the cerebellum and superior colliculus, which may be important for its control over action (Bates & Goldman-Rakic, 1993; Lu, Preston, & Strick, 1994; E. K. Miller & Cohen, 2001). The VLPFC by comparison is linked with the ventral premotor cortex (Petrides

& Pandya, 2002). Wide areas of LPFC project to the dorsal striatum of the basal ganglia. These connections are topographically organised such that DLPFC projects mainly to dorsal and central caudate nucleus whereas VLPFC projects mainly to ventral and central caudate nucleus (Haber, Kunishio, Mizobuchi, & Lynd-Balta, 1995; Parent & Hazrati, 1995). These connections, as well as other major inputs and outputs of LPFC are summarized in Figure 2.3 (taken from (Tanji & Hoshi, 2008)).

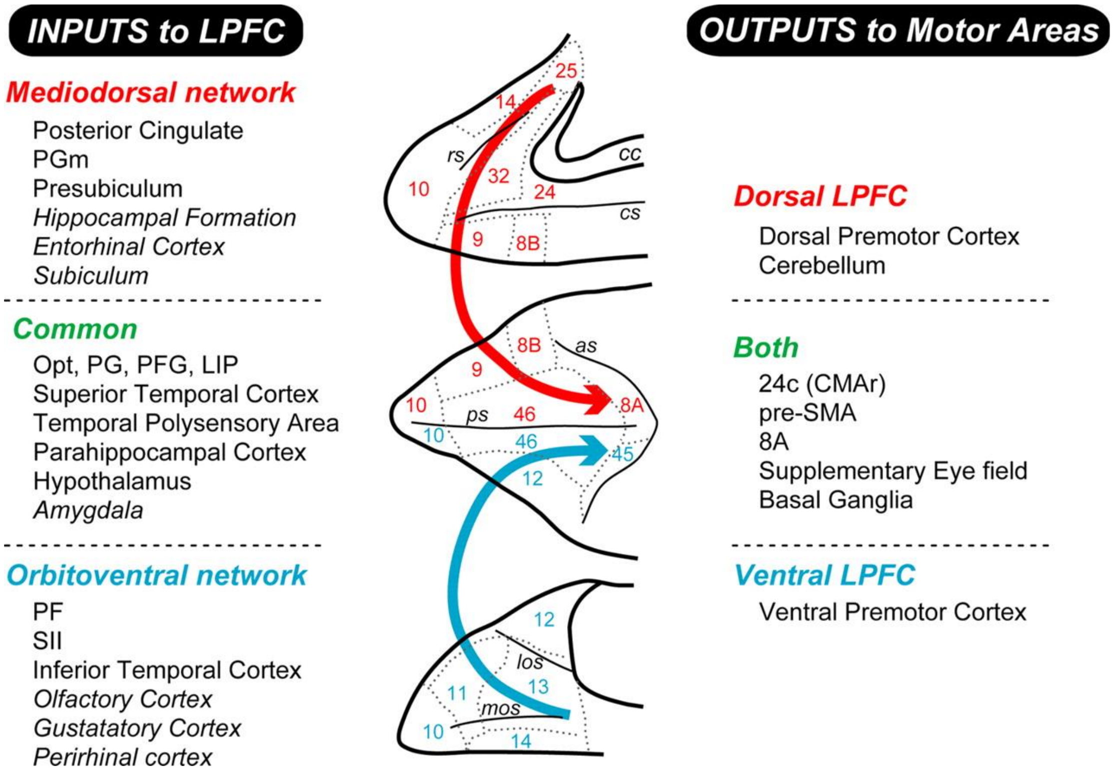


Figure 2.3 Schematic of the major input-output organization and cytoarchitecture of the lateral prefrontal cortex; taken from (Tanji & Hoshi, 2008). The top panel refers to the mediodorsal network (red), of which the DLPFC forms an integral part, whereas the bottom panel refers to the orbitoventral network (blue), of which the VLPFC forms an integral part. The middle panel (green) refers to inputs and outputs that are common to both networks, where areas chiefly projecting to the orbital or medial prefrontal cortex, but less to the LPFC, are italicized. The left column refers to input structures, the right column to output structures, and the middle column to cytoarchitectonic boundaries. *Rs = rostral sulcus; cs = cingulate sulcus; cc = corpus callosum; as = arcuate sulcus; ps = principal sulcus; mos = medial orbital sulcus; los = lateral orbital sulcus. PF, PFG, PG, PGm, and Opt are subareas in the parietal cortex* (see (Pandya & Seltzer, 1982))*. SII = secondary somatosensory area; LIP = lateral intraparietal area; CMAr = rostral cingulated motor area; pre-SMA = pre-supplementary motor area.*

2.2.2 Medial prefrontal cortex

The medial prefrontal cortex (mPFC) can be split into dorsal and orbital (ventral) components. The dorsomedial prefrontal cortex (DMPFC) spans areas 24, 32 and 33, and for simplicity will be considered as synonymous with the anterior cingulate cortex (ACC) in this thesis. Neurons in ACC receive afferent projections from the anterior medial, medial dorsal and parafascicular thalamic nuclei (Gabriel, Burhans, Talk, & Scalf, 2002). Other major inputs come from visual cortex, hippocampus, subiculum, entorhinal cortex and amygdala (Gabriel et al., 2002; Vogt, Rosene, & Pandya, 1979). There is also significant reciprocal connectivity between anterior and posterior cingulate cortices (Vogt et al., 1979). Anterior cingulate neurons largely project to most of the aforementioned thalamic areas, the subiculum, entorhinal cortex, pons, the basal ganglia (including the caudate nucleus and nucleus accumbens), and in primates, to multiple areas of the motor and pre-motor cortex (Gabriel et al., 2002). It is also worth noting that Goldman-Rakic and colleagues have demonstrated direct reciprocal projections of cingulate cortical neurons (in primates) to the lateral prefrontal and parietal cortex (Goldman-Rakic, 1988).

The orbitofrontal cortex (OFC) has been proposed to span areas 10, 11, 12, 13 and 14 (Walker, 1940), though more recent studies have subdivided these regions further (Carmichael & Price, 1994). In humans the term ventromedial prefrontal cortex (vmPFC) is typically used to refer to a region that spans the medial OFC (mOFC) and other areas on the medial wall, though not the central and lateral regions of OFC (Kringelbach, 2005). OFC receives sensory inputs from gustatory, olfactory, somatosensory, auditory and visual regions, and is perhaps the most polymodal region of the cerebral cortex (Kringelbach, 2005). It has direct reciprocal connections with the amygdala, cingulate cortex, insula, hypothalamus, hippocampus, striatum, periaqueductal grey and DLPFC (Cavada, Company, Tejedor, Cruz-Rizzolo, & Reinoso-Suarez, 2000; Kringelbach, 2005).

It has been proposed that OFC forms part of a larger functional network known as the orbital and medial prefrontal cortex (OMPFC), which includes parts of ACC and has unique anatomical connections with the rest of the brain (Carmichael & Price, 1994, 1996). Based on local cortico-cortical connections, two connectional systems or networks were recognized within OMPFC, which are referred to as the 'orbital' and 'medial prefrontal networks' respectively. The areas within each network are preferentially interconnected with other areas within the same network, and also have common connections with other parts of the cerebral cortex (Carmichael & Price, 1994, 1996; Ongur & Price, 2000). The orbital network is characterized by connections with several areas of sensory cortex, whereas the medial prefrontal network is characterized by its outputs to visceral control areas in the hypothalamus and periaqueductal grey (Price & Drevets, 2010). It also has connections with specific regions or cortex that include the rostral part of superior temporal gyrus, the anterior and posterior cingulate cortex, the entorhinal cortex and parahippocampal cortex (Price & Drevets, 2010; Saleem, Kondo, & Price, 2008).

2.2.3 Frontopolar cortex

The frontopolar cortex (FPC), commonly associated with Brodmann area 10 (though its precise cytoarchitectonic boundaries are debated (Ramnani & Owen, 2004)), is the anterior most region of PFC and one of the least well understood regions of the human brain (Christoff & Gabrieli, 2000). However, there is a broad consensus that FPC is important for high-level cognition, including the learning and representation of abstract actions and task rules, and influences processing in more posterior prefrontal regions (Badre & D'Esposito, 2009; Boschin, Piekema, & Buckley, 2015; Koechlin, Basso, Pietrini, Panzer, & Grafman, 1999). FPC is unique in that it seems to lack connections with ‘downstream’ regions in the way that other cortical regions are connected. Instead, it shares reciprocal connections with supramodal cortex in the PFC, anterior temporal cortex and cingulate cortex (Ramnani & Owen, 2004).

2.2.4 Striatum

The striatum is a subcortical structure that acts as a major input station to the basal ganglia. Anatomically it can be divided into the dorsomedial striatum (caudate nucleus in humans), dorsolateral striatum (putamen in humans), and ventral striatum (nucleus accumbens and olfactory tubercle, though the term ventral striatum is often synonymous with the former) (Balleine, Delgado, & Hikosaka, 2007; Haber & Knutson, 2010; Parent & Hazrati, 1995). The striatum interacts with the cortex via recurrent networks referred to as corticostriatal loops, classically divided into four loops: motivational, executive, visual and motor (Seger, 2008). Almost all of the cortex sends projections to the basal ganglia (including the striatum and subthalamic nucleus), which themselves project to output structures of the basal ganglia such as the globus pallidus, internal segment and substantia nigra pars reticulata. These regions project to the thalamus and then back to the cortex forming “loops” (see Figure 2.4) (Tekin & Cummings, 2002).

Different cortical areas have predominant connections to different striatal regions and these are summarised in Figure 2.5, taken from (Seger, 2008). These connectivity profiles are thought to underlie differences in the role of the caudate, putamen and ventral striatum in decision-making (Balleine et al., 2007; Haber, 2011; J. O'Doherty et al., 2004; Yin, Knowlton, & Balleine, 2004), and these will be explored in detail in the following section (2.3). Importantly, the so-called “reward circuit”, first identified by the observation that rats would work for electrical stimulation in specific brain sites (Olds & Milner, 1954), forms an integral part of the cortico-basal ganglia system. The key structures in this network are the anterior cingulate cortex, orbitofrontal cortex, ventral striatum, ventral pallidum and the midbrain dopamine neurons. Their anatomical projections (focusing on inputs to, and outputs from the ventral striatum) are shown in detail in Figure 2.6, taken from (Haber & Knutson, 2010).

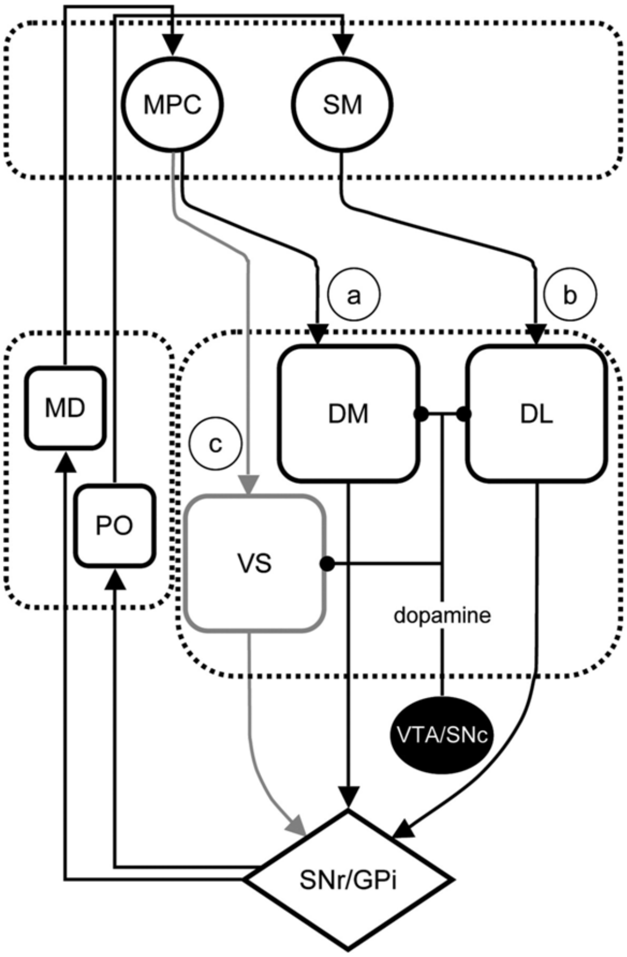


Figure 2.4 Corticostriatal circuits involved in decision-making; taken from (Balleine et al., 2007). Recurrent loops from sensorimotor (SM) and medial prefrontal cortex (MPC) to dorsolateral (DL) and dorsomedial (DM) striatum mediate the acquisition of habitual and goal-directed control respectively. These connections feed back to the cortex via the substantia nigra pars reticulata/internal capsule of the globus pallidus (SNr/GPi) and mediodorsal/posterior (MD/PO) nuclei of the thamalus. A further corticostriatal loop involving the ventral striatum (VS) (and tonic dopamine release) influences the performance of the DL and DM loops through encoding reward acquisition and reward prediction. *VTA/SNc = ventral tegmental area / substantia nigra pars compacta*.

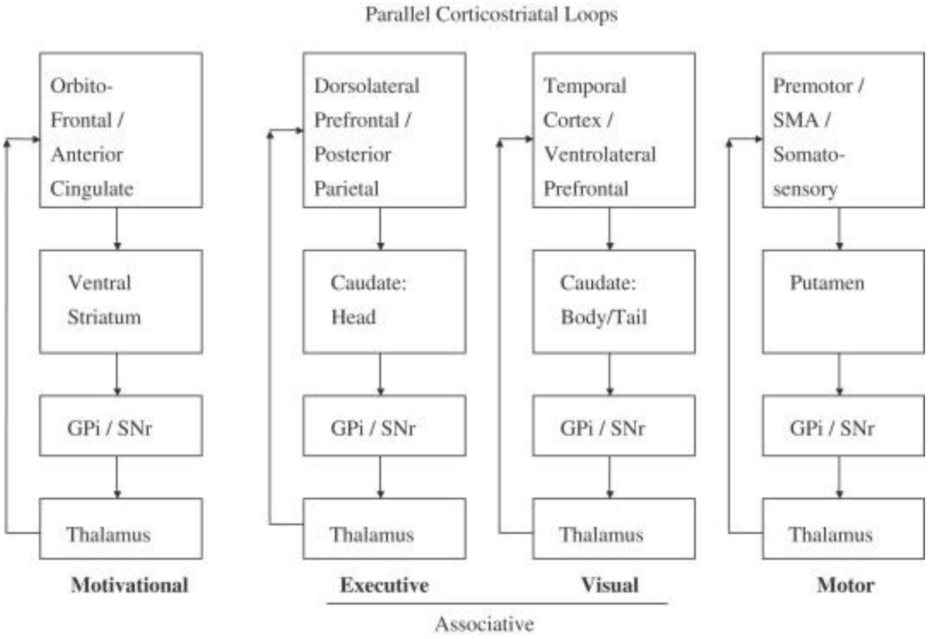

Parallel Corticostriatal Loops
Orbito-Frontal / Anterior Cingulate
Ventral Striatum
GPi / SNr
Thalamus
Motivational
Dorsolateral Prefrontal / Posterior Parietal
Caudate: Head
GPi / SNr
Thalamus
Executive
Temporal Cortex / Ventrolateral Prefrontal
Caudate: Body/Tail
GPi / SNr
Thalamus
Visual
Premotor / SMA / Somato-sensory
Putamen
GPi / SNr
Thalamus
Motor
Associative

Figure 2.5 Illustration of the four major corticostriatal loops; taken from (Seger, 2008). Cortical input from different regions is kept separate as it projects to basal ganglia output structures, then back to cortex. This figure summarises the projection paths of different regions of frontal cortex to different parts of the striatum. These anatomical distinctions may underlie the varying functions of these cortico-striatal loops in value-guided decision-making. *GPi = Globus pallidus, internal portion. SNr = Substantia nigra pars reticulata.*

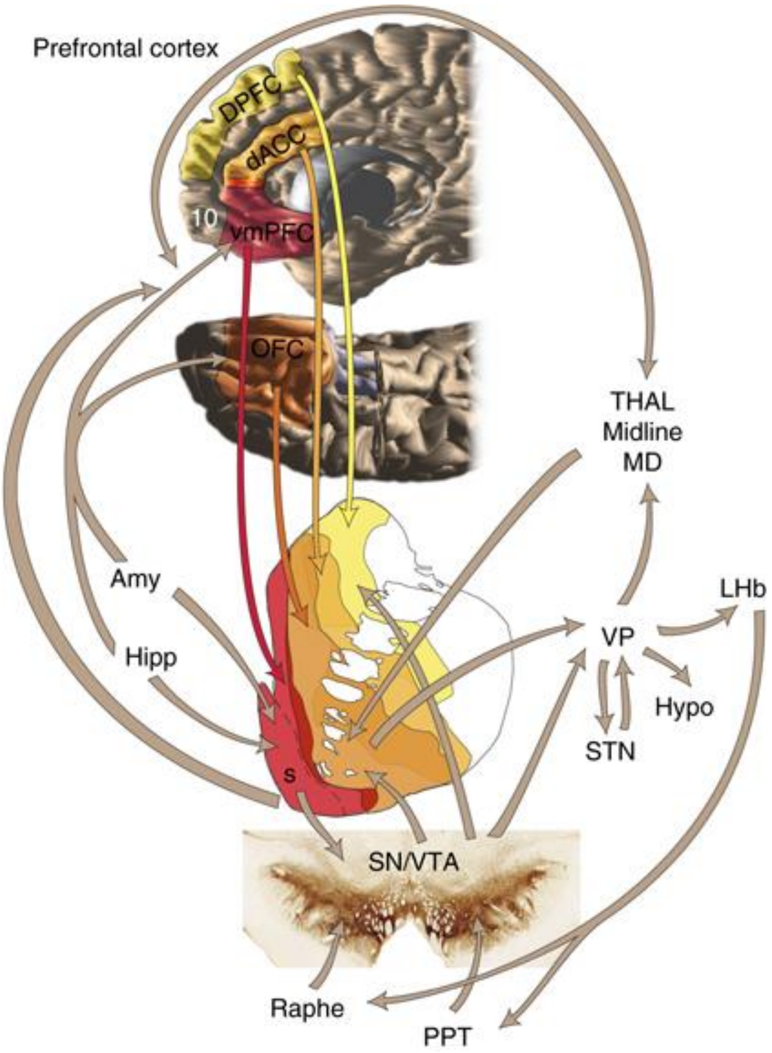


Figure 2.6 Schematic illustrating key structures and pathways of the reward circuit; taken from (Haber & Knutson, 2010). The red arrow refers to input from the ventromedial prefrontal cortex (vmPFC); the dark orange arrow refers to input from the orbitofrontal cortex (OFC); the light orange arrow refers to input from the dorsal anterior cingulate cortex (dACC); the yellow arrow refers to input form the dorsal prefrontal cortex (dPFC); the brown arrows signal other main connections of the reward circuit. *Amy = amygdala; dACC = dorsal anterior cingulate cortex; dPFC = dorsal prefrontal cortex; Hipp = hippocampus; LHb = lateral habenula; hypo = hypothalamus; OFC = orbitofrontal cortex; PPT = pedunculopontine nucleus; S = shell, SNc = substantia nigra, pars compacta; STN = subthalamic nucleus; Thal = thalamus; VP = ventral pallidum; VTA = ventral tegmental area; vmPFC = ventromedial prefrontal cortex.*

2.3 The neural mechanisms of value-guided choice

The investigation of the neural substrates of goal-directed and habitual control in humans would not have been feasible had it not been for a generation of behavioural experiments in animals that helped to characterize and validate these two systems.

2.3.1 Background

Early experiments probed goal-directed and habitual control using instrumental conditioning paradigms (Colwill & Rescorla, 1986). Typically, rodents learn to enact a response, such as pressing a lever, in order to receive a rewarding outcome, such as a food pellet. After a period of training, the rewarding outcome would be devalued, e.g. a food outcome could be fed to satiety, or paired with a noxious substance. Next, the instrumental contingency is tested in extinction, i.e. the animal is now free to press the lever but without delivery of the associated outcome, or any continued reinforcement.

Importantly, goal-directed and habitual systems predict a different course of action following outcome devaluation. If the animal's choice is guided by a stimulus-response (S-O) association, then the animal should continue to enact the conditioned response even if the outcome is undesired. In other words, the disposition to press the lever should be under habitual control. By contrast, if the animal has learnt an action-outcome (A-O) association, less instrumental responding should occur as the animal has an explicit representation of an outcome that is no longer motivationally salient. Thus, behaviour in the latter case is said to be goal-directed. It is now more than 30 years since Adams and Dickinson first showed that rats trained to press a lever for sucrose reduced their responding in an extinction test following devaluation of the sucrose, confirming that animals are indeed capable of forming A-O associations (C. D. Adams & Dickinson, 1981). Interestingly however, S-O (habitual) control prevails in a context where the instrumental contingency is over-trained prior to testing in extinction (Dickinson, Nicholas, & Adams, 1983), suggesting that behaviour is initially goal-directed (R-O driven), but transitions to habitual (S-R driven) with increasing exposure.

These experiments were not only important as a proof-of-concept for dual systems in decision-making, but also formed the conceptual basis for a number of lesion and pharmacological manipulations that aimed to uncover their neural bases. These have revealed strong evidence in rodents that dorsomedial striatum is required for goal-directed behaviour (Yin, Ostlund, et al., 2005), whereas dorsolateral striatum supports habitual behaviour (Yin et al., 2004). These two regions correspond to the caudate nucleus and putamen in primates, respectively. Interestingly, lesions of the PFC, in particular the insular and prelimbic regions of PFC, induce similar deficits in the ability to appropriately adapt to changes in outcome contingency as those seen with lesions to dorsomedial striatum (Balleine & Dickinson, 1998), implying these regions may form part of a common functional network.

In this section I will provide a detailed overview of what is currently known about the neural basis of value-guided decision-making, taking each anatomical region in turn.

2.3.2 Lateral prefrontal cortex

As previously mentioned, the LPFC can be anatomically delineated into the dorsolateral prefrontal cortex (DLPFC, Brodmann areas 9, 46 and 9/46) and ventrolateral prefrontal cortex (VLPFC, Brodmann areas 44 and 45) (see Figure 2.2, p. 33). Although much of this section will focus on DLPFC, these regions have distinct functional roles in decision-making and these will be discussed in turn.

Some of the earliest evidence relating neurobiology to human behaviour came from clinical observations of the effects of brain injury (Fellows, 2013), and It is well-documented that damage to the LPFC in humans is associated with a cascade of cognitive deficits, including an impaired ability to recognize changes in the external environment, deficits in inhibitory control, and a reduced capacity for bridging together temporally segregated events (Manes et al., 2002; Owen, 1997). Further, lesion experiments in both humans and monkeys have revealed both a rostral-caudal axis in the organisation of cognitive control in LPFC, and a dorsal-ventral axis in the mid-lateral region of PFC (Petrides, 2005). These distinctions, first discovered in the late 1980s (Petrides, 1987) and early 1990s (Petrides, 1994) respectively, have since been supported by an array of functional MRI experiments in humans (Badre, 2008; Koechlin et al., 2003; Petrides, Alivisatos, & Frey, 2002).

Briefly, it is thought that the most caudal region of LPFC is involved in fine motor control and sensorimotor mappings, whereas more rostral region of LPFC are involved in higher-order control processes that regulate selection among multiple competing responses and stimuli based on conditional operations (Petrides, 2005). In this manner, posterior-anterior LPFC mediates progressively abstract, higher-order, and most likely hierarchical, control (Badre & D'Esposito, 2009). While the mid-dorsolateral PFC is thought to primary be involved in the monitoring of information in working memory (Petrides, 2000), the mid-ventrolateral PFC is thought to be involved in the active retrieval and manipulation of information held in posterior cortical association regions (Petrides, 1996, 2005).

From a neuroimaging perspective, fMRI experiments suggest LPFC plays a central role in control processes that have a modulatory influence on decision-making and support goal-directed behaviour (Badre & Wagner, 2004; Duncan & Owen, 2000; Johnson, 2001; Petrides, 1996; Shallice & Burgess, 1996). These experiments have included tasks that require future planning and the calculation of long-term values (Balleine & Dickinson, 1998; Basten et al., 2010; Glascher et al., 2010; van den Heuvel et al., 2003; Wallis & Miller, 2003; Wunderlich, Dayan, et al., 2012). Outside of the value domain, the DLPFC is recruited in tasks requiring executive or cognitive control (Badre, 2008; Badre & Wagner, 2004; M. M. Botvinick, Braver, Barch, Carter, & Cohen, 2001; Knight, Grabowecky, & Scabini, 1995; E. E. Smith & Jonides, 1999), and has been associated with the management of working memory (Barbey, Koenigs, & Grafman, 2012; Curtis & D'Esposito, 2003), attentional control (Knight et al., 1995), reasoning and planning (van den Heuvel et al., 2003), and action initiation (Frith, Friston, Liddle, & Frackowiak, 1991). While VLPFC is also implicated in executive control, it is thought to have a much more focused role pertaining to the inhibition of unwanted or prepotent actions (Aron, Robbins, & Poldrack, 2004; Braver et al., 2001). This is particularly notable in Go-NoGo or stop-signal tasks, in which subjects typically have to enact a speeded response on Go trials, but to inhibit responding when a NoGo trial is displayed (e.g. via a visual cue) or a stop signal is presented (e.g. via an auditory tone) (Logan, Cowan, & Davis, 1984).

Despite a growing consensus in neuroimaging that LPFC supports executive or goal-directed decision-making, it is important to note that lesion studies are not always consistent with this. For example, damage to LPFC in humans does not reliably disrupt working memory performance in delayed match-to-sample tasks (D'Esposito & Postle, 1999). Further, while some studies report that patients with LPFC damage show impaired decision-making in the Iowa Gambling task (Fellows & Farah, 2005; Manes et al., 2002) (a task requiring goal-directed inferences (Bechara, Damasio, Damasio, & Anderson, 1994)), other studies report unimpaired performance (Bechara, Damasio, Tranel, & Anderson, 1998). This has led to a degree of controversy regarding the precise role of LPFC in decision-making.

However, it is important to remember that neuroimaging and lesion studies provide complimentary evidence. While neuroimaging relates measures of regional brain activation to behaviour, lesion studies examine whether a particular brain region is *essential* for a given process or component of behaviour. Thus, it is possible that in some cases, fMRI activations in LPFC represent processes that are merely correlated with those supporting goal-directed behaviour (Fellows, 2013). This potential criticism has at least partly been addressed by a recent experiment that has provided a comprehensive mapping of multiple tasks (that measure both cognitive control and decision-making) in a large sample of well-characterized patients with focal brain lesions (including LPFC). Here, Gläscher and colleagues report that LPFC plays an essential role when competing responses need to be inhibited, and is recruited for functions such as error detection and conflict monitoring that are important for adaptive, goal-directed behaviour (Glascher et al., 2012).

In a famous planning task, first designed by Shallice and known as the Tower of London task (Shallice, 1982), a player is presented with two configurations (a start state and a goal state) of three coloured balls arranged in three pins, and the objective for the player is to transform the balls from the start state into the goal state in the least number of moves possible (Figure 2.2A, p. 33). While several variants of the task exist, in all cases, the player has to plan through the correct sequence of moves before initiating any action. It is well-established that the DLPFC (as well as the premotor cortex, supplementary motor area, striatum and parietal cortices) is recruited during the planning phase, with the BOLD response correlating with task difficulty (e.g. an increase in the minimum number of moves) (van den Heuvel et al., 2003) (Figure 2.2B, p. 33). Related to this, tasks that require switches in response or the resolution of conflicting responses, such as in the Stroop (Stroop, 1935) or Eriksen flanker (Eriksen & Eriksen, 1974) tasks, also reliably activate the DLPFC; although these tasks also famously recruit the anterior cingulate cortex (ACC; see 2.2.3) (M. Botvinick, Nystrom, Fissell, Carter, & Cohen, 1999; Kerns et al., 2004).

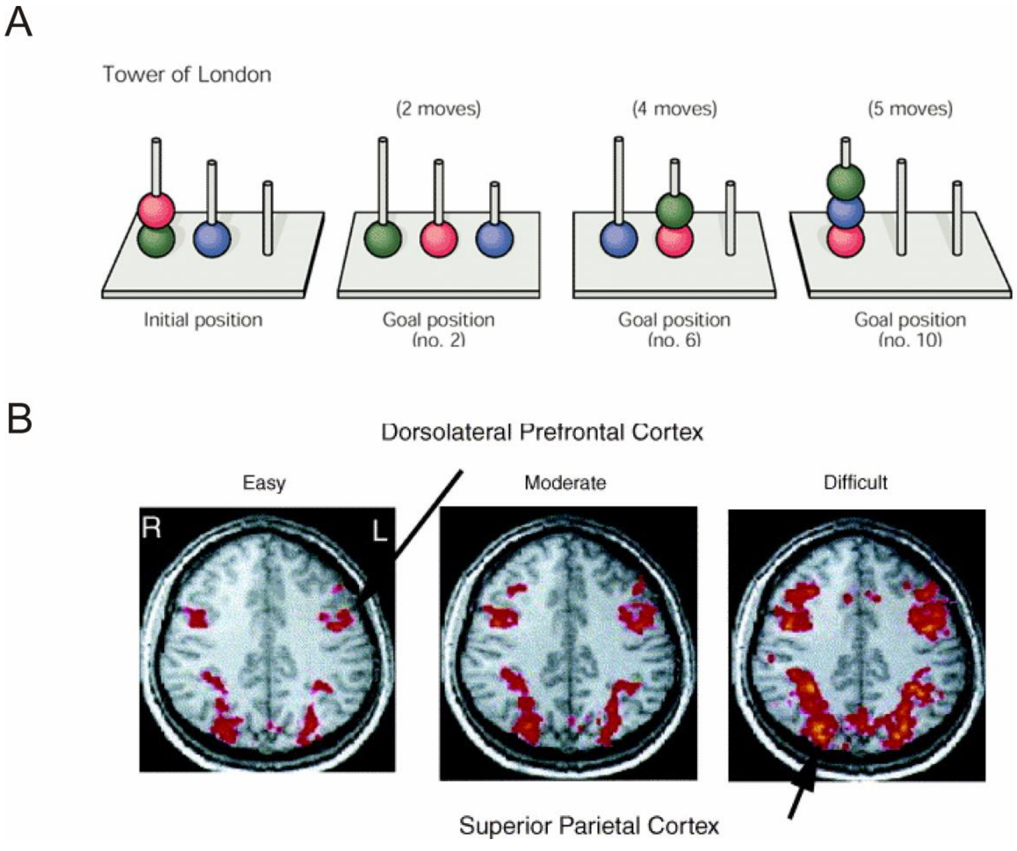


Figure 2.7 A schematic of the Tower of London (ToL) task and the associated neural networks recruited during planning; adapted from (Newman, Carpenter, Varma, & Just, 2003; Saper, Iversen, & Frackowiack, 2000; Shallice, 1982). (**A**) In Shallice's original task a player must plan an action sequence, comprising a pre-allocated number of moves, in order to transverse from an initial position to a goal position. (**B**) Dorsolateral prefrontal cortex and superior parietal cortex are recruited bilaterally during planning, with the strength of activation increasing with task difficulty.

A further role frequently attributed to the LPFC is the representation of abstract task rules (Stokes et al., 2013), such as where subjects are required to respond to colour versus orientation (W. Cai & Leung, 2009). For example, a number of recent studies using multivariate pattern analysis (MVPA; see (Haxby, 2012)) and decoding methods, have shown representations of task rules or contexts in distributed frontoparietal networks, including both the DLPFC and VLPFC (Reverberi, Gorgen, & Haynes, 2012; Waskom, Kumaran, Gordon, Rissman, & Wagner, 2014; Zhang, Kriegeskorte, Carlin, & Rowe, 2013). It is thought that these representations provide a contextual bias on low-level perception, decision-making and action, allowing stimulus-response processing to align with internal goals (Waskom et al., 2014). Further, subjects with prefrontal cortex damage are unable to flexibly adapt to changes in such associative rules (Moore, Schettler, Killiany, Rosene, & Moss, 2009).

While these data suggest that LPFC is engaged during goal-directed choice, many of these paradigms do not allow for attribution of specific computations to the underlying BOLD response, generating uncertainty regarding the precise contribution of LPFC. Recent experiments have attempted to address this using computational modelling and parametric fMRI designs to map specific neural computations relevant for goal-directed choice. For example, it has been shown that the DLPFC (in additional to the parietal cortex) tracks state prediction errors that encode discrepancies between observed state transitions and an agent’s current model of the world, a computation particularly relevant for model-based decision-making (Glascher et al., 2010). In the value domain, several recent experiments suggest that DLPFC supports choice by encoding goal values (Plassmann, O'Doherty, & Rangel, 2010), or by modulating representations of value in other valuation regions, such as the ventromedial prefrontal cortex (vmPFC) and striatum (Diekhof & Gruber, 2010; Hare et al., 2009). Of special note, representations of subjective goal values have been demonstrated using fMRI in both the vmPFC and DLPFC in an economic auction paradigm where subjects bid for the opportunity to eat or avoid foods they liked or disliked respectively, and similar goal values have been identified in the non-human primate DLPFC (Wallis & Miller, 2003).

Both the DLPFC and posterior parietal cortex are engaged during intertemporal choice suggesting these regions form part of a value network that is able to calculate long-term values, and this may be related to the exercise of self-control (Kable & Glimcher, 2007). Interestingly, a recent study has shown that the degree of effective connectivity between DLPFC and vmPFC (a region associated with the computation of stimulus values; see

Ventromedial prefrontal / orbitofrontal cortex, p. 52) was predictive of an individual subject's propensity to discount future rewards in a similar paradigm (Hare, Hakimi, & Rangel, 2014). Other evidence that DLPFC supports self-control comes from a recent study where dieters made choices between food items rated according to healthiness and taste (Hare et al., 2009). The researchers demonstrated that vmPFC tracked goal values independent of the degree of self-control, but while it incorporated a representation of both health and taste in self-controllers, only taste was tracked in non-controllers. Importantly, self-control was associated with increased activity in DLPFC and an enhanced functional connectivity between DLPFC and vmPFC, the latter suggestive of top-down modulation of a representation of value in vmPFC by DLPFC (Hare et al., 2009).

Despite these advances, there are several open questions regarding the role of LPFC in goal-directed choice. While a majority of decision-making paradigms involve one-shot decisions, real life requires making sequences of choices, each conferring an immediate reward or punishment and a complex set of delayed consequences. We know little about what role LPFC plays in estimating these long-term consequences or how this contributes towards adaptive decision-making. Further, it is not clear whether there is a single common value system (Hare et al., 2009; Kable & Glimcher, 2007) that guides choice, or whether separate systems calculate immediate and long-term components respectively (McClure et al., 2004) when choice is sequential. Several lines of evidence point towards multiple value systems, although there are differing accounts of the computations subserved by each system (Balleine, 2005; Dolan & Dayan, 2013; McClure et al., 2004).

2.3.3 Anterior cingulate cortex

It is worth focusing briefly on the anterior cingulate (ACC) region of dPFC and its relation to adaptive decision-making. It has long been known from neurophysiology studies that damage to the ACC results in impaired decision-making, particularly in the ability to adapt to fluctuations in context (Kennerley & Walton, 2011; Kennerley, Walton, Behrens, Buckley, & Rushworth, 2006). For example, ablation of dACC in humans results in an increase in the number of errors when subjects are required to flexibly respond according to a changing set of reward contingencies (Williams, Bush, Rauch, Cosgrove, & Eskandar, 2004). ACC is also of particular interest because of the myriad of roles it has been associated with, including the detection of decision errors (Braver et al., 2001), monitoring decision conflict (Kerns et al., 2004), overriding prepotent responses (Liston, Matalon, Hare, Davidson, & Casey, 2006), and evaluating choice outcomes (Rushworth, Walton, Kennerley, & Bannerman, 2004; Walton, Croxson, Behrens, Kennerley, & Rushworth, 2007). More recently it has been linked to foraging in the context of evaluating alternative choice options (Behrens et al., 2007), and the coding of state prediction errors or surprising events (Ide et al., 2013).

Interestingly, several tasks that recruit the ACC require implicit representations of downstream consequence, and as such it seems plausible that ACC could have a common role in evaluating future outcomes during sequential choice. However, to my knowledge no paradigm has been able to formally address this. Moreover, while it is thought that ACC tracks decision variables that feed into the decision process, and is thus “pre-decisional” (Kerns et al., 2004; Wunderlich et al., 2009), there are also proposals that ACC signals outcome variables that can be used to inform future choice, and is thus “post-decisional” (Blanchard & Hayden, 2014; X. Cai & Padoa-Schioppa, 2012). Recent data from non-human primates suggests a dominant role for the latter, though conflicting evidence has left the issue controversial.

For example, in a recent experiment, Blanchard and Hayden showed that dACC signalled the value of foregone choice options, a post-decisional variable, when monkeys were performing a simple foraging task (Blanchard & Hayden, 2014). Further, lesion experiments suggest that ACC is essential for using extended action-outcome histories to learn the value of actions and optimize future choices (Kennerley et al., 2006). By recording neuronal activity from the PFC in behaving monkeys, Kennerley and colleagues showed that a portion of ACC neurons encoded the probability of reward during decision-making (a likely pre-decisional computation) in addition to the discrepancy between expected and experienced reward during outcome receipt – a post-decisional prediction error signal (see Chapter 1, p. 18-20) (Kennerley et al., 2006; R. S. Sutton, 1988). Recent experiments using fMRI in humans have provided corroborative evidence that ACC indeed encodes a prediction error signal for driving future decisions (Ide et al., 2013). However, other studies have noted that the BOLD response in ACC matches the output of a putative value comparator, and may thus be involved in the decision process itself (Wunderlich et al., 2009). Lastly, the computational role played by ACC with regards to evaluating future consequences, and how this is distinguished from more lateral regions of PFC, remains vague.

2.3.4 Ventromedial prefrontal / orbitofrontal cortex

I have previously discussed that in order to execute decisions that will yield favourable outcomes, the brain needs to assign a value to potential options in the environment. While correlates of goal value have been previously found in LPFC (Kable & Glimcher, 2007; Litt, Plassmann, Shiv, & Rangel, 2011; Plassmann et al., 2007; Sokol-Hessner et al., 2012), this region is more frequently associated with other types of representations, such as the state of the environment (Behrens et al., 2007; Yoshida & Ishii, 2006), the set of current task rules (Moore et al., 2009; Stokes et al., 2013), or the agent’s hierarchical goals (Hare et al., 2009).

In contrast, one of the most coherent findings in the field of value-guided decision-making is that the more ventral portion of PFC, in particular the vmPFC, is especially tuned towards the representation of value in humans (Boorman et al., 2009; Hare et al., 2009; Kable & Glimcher, 2007; Plassmann et al., 2007; Strait et al., 2014). This finding is illustrated in a recent meta-analysis of BOLD response from an array of value-guided fMRI experiments (Bartra et al., 2013), in which the vmPFC (in addition to the striatum) was a region most consistently implicated in value coding. As previously mentioned, vmPFC and medial orbitofrontal cortex (mOFC) are often used interchangeably in fMRI studies, and I use the term vmPFC here to include both regions when discussing human studies.

Figure 2.8 (taken from (Bartra et al., 2013)) (p. 54) shows that overlapping regions of vmPFC and striatum reliably encode subjective value both at the time of making a decision and during receipt of the associated outcome. Importantly, valuations in vmPFC appear to be domain-general, with BOLD tracking subjective value for both primary (e.g. food) (Hare et al., 2009) and secondary (e.g. money) (Kable & Glimcher, 2007) rewards. This has led some to argue that vmPFC and striatum form part of a common currency valuation system (Hare et al., 2009; Kable & Glimcher, 2007), though others postulate that these regions operate on separate value systems that act in parallel (McClure et al., 2004).

Further, while in some studies vmPFC appears to track the value of the chosen option (Kable & Glimcher, 2007; Plassmann et al., 2007) (implicating this region in value representation), others have reported a BOLD response in vmPFC that reflects the difference between chosen and unchosen options (Boorman et al., 2009; Hunt et al., 2012). The latter suggests that vmPFC may act as a final value comparator, thus entering the decision-making hierarchy further downstream. Further still, other evidence suggests that vmPFC encodes the difference in value between attended and unattended choice options, thereby reflecting an attention-modulated value signal (Lim et al., 2011). According to Lim and colleagues, since it is a natural tendency to attend to options that we eventually choose for longer, it can appear as if vmPFC is encoding the difference in value between chosen and unchosen options in experiments that do not control for attention.

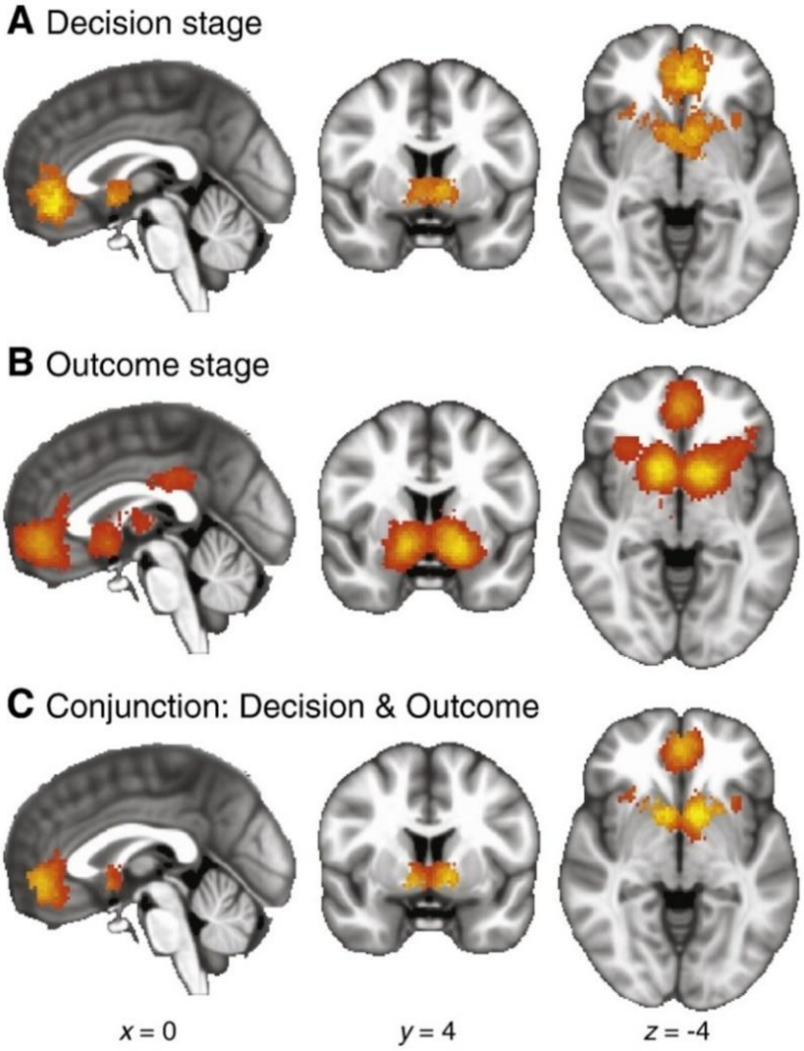


Figure 2.8 Neural representations of subjective value; taken from (Bartra et al., 2013). (**A**) Here, Bartra and colleagues performed a whole-brain meta-analysis of BOLD activation that revealed neural representations of subjective value in vmPFC and striatum at the time of making a choice. (**B**) The same regions of vmPFC and striatum also response to the delivery of a reward at the outcome stage of a trial. (**C**) Conjunction effect of activation maps shown in panels (A) and (B).

While it is clear that vmPFC contributes to behaviour by signalling subjective or “economic” value, it is less clear how this value arises. For example, in sequential environments where decisions can confer both an immediate reward and a complex set of delayed consequences, it remains ambiguous whether vmPFC signals immediate rewards, delayed outcomes, or an integration of both components. Unfortunately, these components are typically correlated and thus indistinguishable in many value-guided decision-making paradigms. A recent study has attempted to address this by asking subjects to make real choices about differing food items (Hare et al., 2009). Here, the BOLD response in vmPFC reflected an integration of taste and health components, suggesting vmPFC accesses long-term, in addition to immediate, components of value; or in more general terms, supports goal-directed behaviour. Yet a value signal that reflects the calculation and integration of both immediate and future consequences has not been previously demonstrated in humans.

Along similar lines, while some argue vmPFC signals value regardless of the associative basis of the information, others have postulated that vmPFC is crucial in contexts where value has to be estimated on the fly through knowledge of the causal structure of the world and the future consequences of actions (a model-based valuation). Recent work conducted in rodents has provided strong evidence for the latter (Jones et al., 2012), though evidence in humans is more sparse.

Evidence from non-human primate electrophysiology studies, in which vmPFC (Brodmann area 14), medial orbitofrontal cortex (mPFC, Brodmann areas 11/13) and lateral orbitofrontal cortex (LOFC, Brodmann areas 47/12) are more easily delineated, suggest that these regions are anatomically and functionally distinct. For example, OFC, unlike vmPFC, receives inputs from sensory systems (Barbas, Ghashghaei, Dombrowski, & Rempel-Clower, 1999; Cavada et al., 2000), while vmPFC, unlike OFC, has dense projections to the nucleus accumbens (Haber et al., 1995) and hypothalamus (Ongur, An, & Price, 1998). Although relatively few studies have recorded from vmPFC, recent evidence suggests partially distinct computations in vmPFC and OFC during value-guided decision-making. For example, while both regions likely encode the subjective value of task events, neurons in OFC may be more sensitive to external factors that relate to value, such as visual cues, while vmPFC may be more sensitive to internal factors that relate to value, such as satiety (Bouret & Richmond, 2010). Further, others have argued for a functional subdivision between ventral vmPFC, in which neurons are more active during appetitive feedback, and dorsal vmPFC, in which neurons are more active during aversive feedback (Monosov & Hikosaka, 2012).

A similar functional subdivision, inspired by human neuroimaging data (for a review see (Kringelbach & Rolls, 2004), has been proposed for mOFC and LOFC, with the former said to specialize in the evaluation of rewards and the latter the evaluation of punishments, but neurophysiology data has largely not supported this theory (Morrison & Salzman, 2009; Rich & Wallis, 2014). Instead, it has been proposed that LOFC neurons encode the value of external stimuli, while mOFC neurons use knowledge of the task structure and environment to make outcome predictions (Rich & Wallis, 2014). This is consistent with other evidence in animals and humans that vmPFC/mOFC is recruited when values are inferred on the fly using a model of action-outcome contingencies (Hampton, Bossaerts, & O'Doherty, 2006a; Jones et al., 2012; Takahashi et al., 2013). Finally, lesion studies suggest that LOFC is required for reward-credit assignment and thus reward-value learning, whereas mOFC is required for value comparison amongst multiple competing alternatives (Noonan et al., 2010).

2.3.5 Striatum

Similar to vmPFC / OFC, the striatum is strongly implicated in value-guided decision-making (Balleine et al., 2007; Balleine & O'Doherty, 2010; Bartra et al., 2013; Brovelli, Nazarian, Meunier, & Boussaoud, 2011; Haber & Knutson, 2010; Kimchi & Laubach, 2009; Roesch, Singh, Brown, Mullins, & Schoenbaum, 2009; Stalnaker, Calhoon, Ogawa, Roesch, &

Schoenbaum, 2010; Yin, Ostlund, et al., 2005). As previously mentioned, the striatum interacts with the cortex via recurrent networks referred to as corticostriatal loops, classically divided into four loops: motivational, executive, visual and motor (Seger, 2008) (see also Figures 2.4 & 2.5). It is well-established that different striatal nuclei are associated with distinct loops and have distinguishable roles (Balleine et al., 2007; Basar et al., 2010; Daw et al., 2011; Yin & Knowlton, 2006).

Here it is thought that the acquisition of reward-related (goal-directed) actions are mediated by converging projections from regions of medial prefrontal cortex (MPC) to the dorsomedial striatum (DM; caudate nucleus in humans), whereas the acquisition of S-O contingencies (habits) are mediated by projections from sensorimotor cortex (SM) to the dorsolateral striatum (DL; putamen in humans) (see *Dorsal striatum*, p. 59). These corticostriatal connections feed back to the cortex via the substantia nigra pars reticulata/internal segment of the globus pallidus (SNr/GPi) and the mediodorsal/posterior (MD/PO) nuclei of the thalamus. A parallel ventral circuit, mediated by the MPC and ventral striatum (VS) drives motivational and Pavlovian influences by feeding into the DM and DL loops (Balleine et al., 2007).

2.3.5.1 Ventral striatum

Much like the vmPFC, the ventral striatum has previously been shown to track the subjective value of choice options (Kable & Glimcher, 2007) (see Figure 2.8, p. 54), and is an integral part of the reward circuit (Haber & Knutson, 2010) (see Figure 2.6, p. 42). However, the ventral striatum is more typically associated with temporal difference learning, and the encoding of reward prediction errors, as evidenced by single neuron recordings in the non-human primate (Schultz et al., 1997) and fMRI paradigms in humans (J. P. O'Doherty, Buchanan, Seymour, & Dolan, 2006). In this framework, the ventral striatum is thought to calculate the difference between expected and received levels of reward, a metric that is important for informing future choice (Rescorla & Wagner, 1972). Typically, trials that contain high value options are also likely to generate a positive prediction error, and thus these two quantities are often highly correlated. This has led to some controversy regarding the true quantity driving ventral striatum responses, though recent accounts claim that prediction errors prevail when they are decorrelated from subjective value (Hare et al., 2008).

While it is clear that the ventral striatum is intimately involved in reinforcement learning, it is less clear what learning system drives activity in this region. For example, prediction errors arising from a model-based system, that require evaluating actions through calling on an internal model of the world, can differ from those arising from a model-free system, where actions that are rewarded are reinforced in a retrospective fashion. While the signal in ventral striatum is classically thought to support model-free learning, a recent study has provided evidence for an integration of both model-based and model-free components of value in this region (Daw et al., 2011).

The ventral striatum is also implicated in value-guided decision-making outside of learning paradigms, particularly in controlling ‘go’ or ‘nogo’ responses. While rewards are typically coupled with the requirement to ‘act’ in many value-guided paradigms, recent experiments in humans have shown that activity in ventral striatum more closely reflects the requirement for action as opposed to the anticipation of wins or losses (Guitart-Masip et al., 2011). This association with action has led to a view that the ventral striatum may be closely involved with impulsivity. For example, it has been shown that the ventral striatum is more active when subjects choose smaller but immediate rewards over larger but delayed rewards in temporal discounting tasks (McClure et al., 2004), and a reduced response in ventral striatum to immediate rewards promotes goal-directed choice (Diekhof & Gruber, 2010). Further, in rodents value representations in ventral striatum have been shown to be insensitive to changes in stimulus-reward contingencies, suggesting this region promotes automatic, stimulus-driven actions as opposed to those that serve goals (Kimchi & Laubach, 2009).

Yet not all the evidence is in accord. For example, lesions of the nucleus accumbens core in rodents induces impulsive choice (Cardinal, Pennicott, Sugathapala, Robbins, & Everitt, 2001), a phenomenon counter to that predicted by this region promoting impulsivity, and ventral striatum activity appears to reflect an integration of value and action requirements in animals (Roesch et al., 2009). Collectively, these data suggest that the ventral striatum may be differentially engaged depending on subtle task differences that probe varying facets of impulsivity.

2.3.5.1 Dorsal striatum

The dorsal striatum is a thought to mediate several important aspects of value-guided decision-making (Balleine et al., 2007). Much like the ventral striatum, the BOLD response in dorsal striatum has been shown to reflect the anticipation of both primary (J. P. O'Doherty, Deichmann, Critchley, & Dolan, 2002) and secondary (Knutson, Adams, Fong, & Hommer, 2001) rewards in humans. In addition the dorsal striatum is thought to be particularly involved in action-contingent learning (Tricomi, Delgado, & Fiez, 2004). However, evidence indicates that the two key sites within dorsal striatum, the caudate nucleus and putamen, play different roles.

The putamen in humans is largely associated with the formation of habits (Yin & Knowlton, 2006) and stimulus-driven responses, possibly through encoding stimulus-response associations (Featherstone & McDonald, 2005). This is consistent with previously discussed evidence from rodents, where lesioning dorsolateral striatum (corresponding to the putamen) disrupts habit formation such that animals remain sensitive to devaluation protocols even after overtraining (Yin et al., 2004). Human studies also implicate the putamen in shifting choice towards acquisition of immediate rewards (Tanaka et al., 2004) and in the tracking of values associated with extensively trained choice (Wunderlich, Dayan, et al., 2012). Collectively, these studies align the putamen more closely with automatic, model-free processing.

In contrast, while the caudate nucleus is associated with the learning of actions and their reward consequences, it is thought to support goal-directed over habitual choice. A number of human fMRI studies now point towards the caudate as a region coding reward prediction errors specifically during goal-directed behaviour (Delgado, Miller, Inati, & Phelps, 2005; J. O'Doherty et al., 2004). Perhaps the most direct example of model-based processing in the caudate comes from a recent experiment showing that the caudate tracks the value of individual branching steps in a decision tree (Wunderlich, Dayan, et al., 2012). The caudate has also been implicated in future reward prediction (Tanaka et al., 2004) and the encoding of both positive and negative action consequences (Tricomi et al., 2004), all suggestive that this region is involved in promoting goal-directed decisions. Evidence from the rodent literature implicating the dorsomedial striatum in goal-directed control is also largely consistent with neuroimaging data from humans. For example, both pre and post-training lesions (Yin, Ostlund, et al., 2005), muscimol-induced inactivation (Yin, Ostlund, et al., 2005), and the infusion of an NMDA antagonist (Yin, Knowlton, et al., 2005) within dorsomedial striatum abolish goal-directed behaviour and render choice insensitive to outcome devaluation.

CHAPTER 3

METHODS

3.1 An introduction to neuroimaging

There are many methods available for measuring brain activity in behaving humans (Bear, Connors, & Paradiso, 2007). The oldest method, electroencephalography (EEG), records electrical activity on the scalp by measuring voltage fluctuations that result from ionic current flows within the neurons of the brain. Data collected via EEG, and its relative magnetoencephalography (MEG), has excellent temporal resolution but poor spatial resolution given the complexities involved with identifying the anatomical source of activity recorded on the scalp. In addition, EEG/MEG is almost always contaminated with artefacts.

An alternative method that has gained great popularity since its emergence over two decades ago is functional magnetic resonance imaging (fMRI). With fMRI, the activity of neurons is not measured directly but rather is inferred through measuring regional changes in the concentration of oxygen within blood vessels - the blood-oxygen-level dependent (BOLD) response - affording a vastly improved spatial resolution. This relies on the fundamental property that when a region of the brain is in use, blood flow to that same region increases to meet metabolic demands, which in turn increases the proportion of oxygenated Hemoglobin and changes the magnetic property of blood. fMRI measures this change in magnetic property. However, these changes in blood flow lag several seconds behind the underlying neuronal activity, resulting in a reduced temporal resolution.

Multiple lines of evidence point towards the implementation of functional localisation and specialisation in the brain. That is, neurons that perform equivalent physiological functions group together into anatomically separable regions (Bear et al., 2007). This is particularly evident in patients with focal brain damage, where lesions to a particular region can induce highly specific and reproducible deficits in cognition (Alvarez & Emory, 2006; Badre, 2008; Bechara, Tranel, & Damasio, 2000; Lavenex, Amaral, & Lavenex, 2006). Further, neurophysiological recordings undertaken in non-human primates have exposed a location-specific mapping of function within domains such as vision (D. L. Adams & Horton, 2003; Takechi et al., 1997). It is this property which makes fMRI a useful tool for exploring brain function.

fMRI allows inferences to be made about the simultaneous activity of the whole brain during a task or cognitive manipulation. One can therefore infer both the magnitude of activity in specific regions and how this activity changes over time. In practice, the analysis of fMRI data involves either comparing the BOLD response in one psychological context versus another, e.g. viewing faces versus scenes, or testing for correlations between BOLD and a given task attribute. The latter is of particular relevance to the study of decision-making, as one can determine which regions of the brain are sensitive to variables fundamental to the underlying neuronal process, such as reward, uncertainty, or subjective value. This provides a much more powerful tool than simply reporting a list of task-related brain activations, as it allows for the attribution of specific computational roles. In this thesis, I employed relatively standard acquisition protocols and analysis pipelines, which will be reviewed in the following sections. However, the reader is also encouraged to consult a number of excellent reviews (Jezzard, Smith, & Matthews, 2003; S. M. Smith, 2004).

3.2 Physics of MRI

Magnetic resonance imaging (MRI) measures an electromagnetic signal from the hydrogen nuclei within water molecules. The positively charged protons in water act as microscopic compass needles that emit a small electromagnetic field, but are randomly oriented in their natural state. The magnet of the MRI scanner generates a strong radiofrequency electromagnetic field that acts to momentarily align the nuclei with the direction of the magnetic field. A second magnetic field, the gradient field, is then applied to induce a higher magnetization level. When the gradient field is removed, the nuclei return to their original orientation which results in the release of an electromagnetic signal detectable by the MRI scanner. In echo-planar imaging, each radiofrequency excitation is followed by a train of gradient fields with different spatial encoding that allows for the rapid acquisition of images.

Functional MRI (fMRI) is used to estimate the brain activity evoked by a particular task through measuring regional changes in oxygen concentration within blood vessels. The process is similar to conventional MRI but uses the change in magnetization between oxygen-rich and oxygen-poor blood as its basic measure. Hemoglobin is an iron-containing molecule, found predominantly in red blood cells, that acts to transport oxygen from the respiratory organs to the rest of the body, to meet the needs of metabolically active tissue. Oxygen binds to the heme component of Hemoglobin in the pulmonary capillaries adjacent to the lungs resulting in Oxyhemoglobin. When oxygen is released into cells, Hemoglobin becomes relatively deoxygenated.

Importantly, external magnetic fields have negligible influences on oxygenated Hemoglobin, but cause local magnetic field variations with deoxygenated Hemoglobin (Figure 3.1, p. 64) (Ogawa, Lee, Kay, & Tank, 1990), as the four outer electrons of the iron electron are unpaired with oxygen. Blood-oxygen level dependent contrast (BOLD), first described by Seiji Ogawa in rat studies (Ogawa et al., 1990), is able to exploit this dissociation to estimate, albeit indirectly, underlying neuronal activity. The usual signal increases reported in BOLD fMRI experiments are due to the fact that neural activation induces a regional increase in cerebral blood flow and glucose utilization that is always larger than the oxygen consumption rate, since oxygen uptake is diffusion-limited. The net effect of neural excitation is thus a seemingly paradoxical drop in the deoxyhemoglobin concentration, which in turn increases the signal strength (Logothetis, 2008).

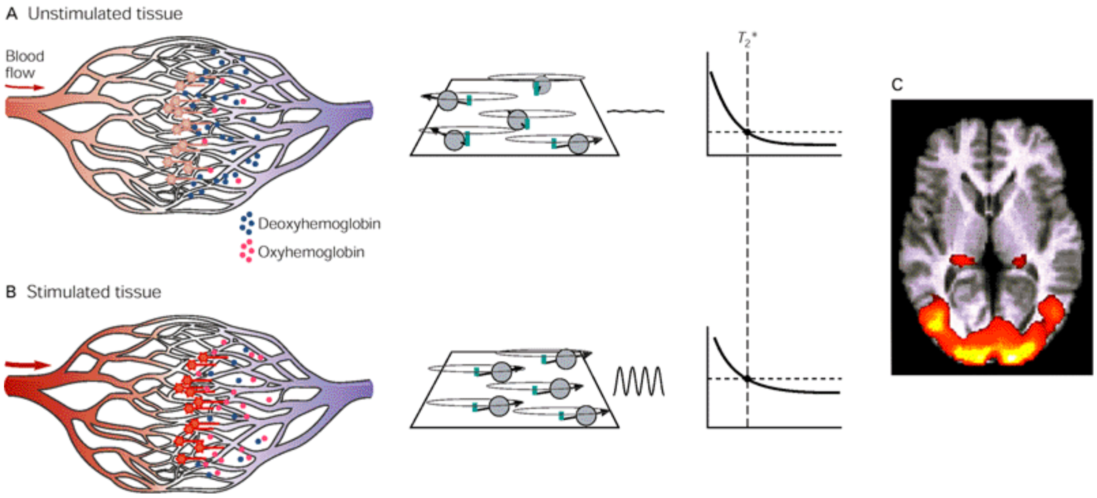


Figure 3.1 Schematic of a change in blood flow in response to a visual stimulus and the associated change in magnetization measured by fMRI; taken from (Saper et al., 2000). (**A**) When neurons in the visual cortex are not stimulated, a relatively large proportion of local Hemoglobin is in the deoxy form. Since deoxyhemoglobin promotes efficient dephasing of the rotating protons, the T_2* curve is steep and the MRI signal is weak. (**B**) Conversely, when neurons in visual cortex are activated, blood flow increases resulting in a heightened proportion of oxygenated relative to deoxygenated Hemoglobin. This results in a slower dephasing of protons and a less steep T_2* curve. (**C**) A heightened BOLD response in visual cortex results from an increase in the relative proportion of oxygenated Hemoglobin following presentation of the visual stimulus.

Following action potentials in the brain, ions are actively pumped across the cell membrane to ensure the appropriate repolarization of the cell. This process requires glucose and oxygen, which is carried via blood, also acting to bring in oxygenated Hemoglobin via red blood cells. A higher rate of firing causes a greater rate of blood flow and a dilation of regional blood vessels. This results in a change in the ratio of oxygenated to deoxygenated Hemoglobin, and a subsequent alteration in the magnetic property of blood. It is this change in magnetic property that is detected during fMRI. A relative decrease in the proportion of deoxyhemoglobin attenuates local susceptibility effects, and thus increased activity results in a higher signal intensity on T2-weighted images (Figure 3.1, p. 64).

fMRI is susceptible to unwanted noise that originates from the scanner, from random brain activity, and from large blood vessels where blood flow is often highly variable due to factors that are not of interest (see also Chapter 7, p. 183). Consequently, fMRI studies require multiple repetitions of the same events to improve the signal-to-noise ratio.

3.3 Analysis of fMRI data

For the purposes of fMRI, the brain is divided into small cubes of volume, typically 2-3mm^3, known as voxels. In order to make inferences about significant task-related effects, the time-series of each individual voxel, that is, how BOLD activation throughout the scanned volume of the brain changes over time, needs to be assessed. In order to ensure these time-series are accurate and free from artefacts, a number of pre-processing steps are performed (for a review see (Strother, 2006)). These also serve to enable analysis across scans and subjects.

3.3.1 Pre-processing

Below I outline the standard pipeline for pre-processing of an fMRI dataset.

3.3.1.1 Bias Correction (for structural scans only)

The use of a 32-channel head coil may result in biases in signal intensity due to inhomogeneities in the magnetic fields of the MRI scanner. This can affect subsequent pre-processing stages such as segmentation. Image intensities are therefore ‘flattened’ following acquisition by means of a multiplicative factor that changes the intensity values of image pixels.

3.3.1.2 Spatial Realignment & Unwarping

The fMRI signal is expressed in 3-dimension space. Thus, any head movements during image acquisition will result in a mismatch of the location of subsequent images in the time-series. Even movements in the order of a few millimetres, such as those caused by swallowing or possibly associated task performance, can contribute significant variance to the fMRI signal, reducing the overall signal-to-noise ratio and decreasing the power of any subsequent analyses. To account for this, spatial realignment is performed by means of a 6-parameter rigid body transformation that minimizes the difference (typically the sum-of-squares) between subsequent images. Head motion estimates can subsequently be analysed as a quality check. EPI images also exhibit substantial signal dropout and spatial distortion in regions where the magnetic field is inhomogenous. By collecting field maps, which measure field inhomogeneity, EPIs can be unwarped by means of a field mapping distortion correction approach, resulting in improved coregistration between EPIs and anatomical images.

3.3.1.3 Coregistration & Spatial Normalisation

In order to ensure activations measured by fMRI are superimposed onto the correct anatomical location, functional images must be coregistered with an anatomical (T1-weighted) scan. For images of different modality (i.e. anatomical versus functional) this is typically done by computing a transformational matrix that matches mutual information, or minimizes differences between images, and applying this to the data of interest. Next, in order to make comparisons between individuals with different brains, and to extrapolate findings to the population as a whole, scans must be normalised to a common template brain. This also allows the reporting of activations within an established standard space. The template adopted in this thesis is the standard template of the Montreal Institute of Neurology (MNI). During normalisation, images are warped so that functionally homologous regions across different subjects are as close together as possible. This involves using a 12-

parameter affine transformation to minimize the sums of squared differences between the template brain and the subject-specific brain, and also the squared number of standard deviations away from the expected parameter values.

3.3.1.4 Spatial Smoothing

Smoothing involves spatially blurring functional images using a 3-dimensional Gaussian kernel. Smoothing can be applied at both the single-subject level and the group-level. In the case of the latter, smoothing the image increases the overlap of activation between subjects. Smoothing also helps to increase signal-to-noise ratio, because the signal from a single voxel in a smoothed image will also contain a signal from neighbouring voxels, reducing the contribution of random noise. Further, smoothing can be set to match the spatial scale of the data to the size of the expected effect. Researchers interested in both cortical and subcortical activations will typically employ an average smoothing kernel of 6-8mm.

3.3.2 Statistical Modelling

The pre-processing stages provide a set of voxel-based time-series of BOLD activation throughout the entire space of the scanned brain. The goal of any fMRI experiment is to relate these dynamic activations to the experimental manipulation in a statistically valid way.

The general approach involves specifying a general linear model (GLM), in which we propose that our observed data (Y) is a function of our experimental manipulation (X), weighted by a parameter 'beta' that governs the size of the 'effect', and some residual error or noise (Friston et al., 1994).

$$Y = \beta X + \text{epsilon}$$

This is the basis of Statistical Parametric Mapping (SPM, Wellcome Trust Centre for Neuroimaging, London, UK) employed in this thesis. In fMRI, Y is the observed BOLD time-series and X is a matrix of explanatory variables, or regressors. The design matrix includes all relevant experimental manipulations that are proposed to modulate brain activity (effects of interest), plus any uninteresting variables that may also contribute to signal variance (effects of no-interest), such as session effects, movement parameters, and physiological regressors (e.g. pulse rate and breathing). Thus, in effect, SPM employs multiple linear regression, as more than one independent variable is considered in the same model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_p X_p + \text{epsilon}$$

where *p* is the number of regressors in the design matrix.

SPM uses a mass univariate approach and standard parametric statistics to test the null hypothesis that the estimated effect size of any individual regressor in the design matrix is zero. Thus, rather than considering variance between groups of voxels, the time-series from each voxel is fit to the GLM in parallel. Effects of interest relate to the influence of a single regressor (after accounting for all other regressors in the design matrix), calculated as the effect size divided by its standard deviation (to give a T-statistic), or to some linear combination of more than one effect with respect to their relative variances (to give an F-statistic). In order to extrapolate inferences to the population-level, one then performs a random-effects analysis to estimate the variance in betas for a given regressor in the design matrix (or contrast map from a within-subject analysis) between subjects.

Note that our observed data, Y, represents BOLD response, which in turn is related to our key interest, neural activity, in some reliable way. Thus, in order to relate any observed effects to the underlying neuronal response, we must model the relationship between BOLD and neuronal activity in our GLM. This relationship is known as the haemodynamic response function (HRF), and is built into the SPM framework. It is known that the peak in BOLD from a burst of neural activity typically has a lag of 5-6 seconds (Logothetis, 2003). Thus, in event-related designs, where we are interested in the neuronal response to single independent events in time (such as the presentation of a stimulus), onset vectors are convolved with the HRF before being inserted into the design matrix. Thus, we can test whether our experimental manipulations are influencing BOLD in a manner predicted under the assumption that they are influencing neural activity.

One potential drawback of using a voxel-based mass univariate approach is the problem of false positives that are likely to arise with multiple comparisons. In fMRI, statistical tests are repeated on over 100, 000 voxels in the brain, and it is likely that several voxels will show a significant effect by chance. A common approach for multiple comparisons is Bonferroni correction, in which the level of statistical significance is equivalent to 1/n times what it would be if only one test was performed, where n is the number of times the test is performed. However, this method is too conservative for fMRI, as it relies on the assumption that each test is independent. Since the signal in neighbouring voxels is often correlated, we use principles of random field theory to construct a more appropriate method for correction (Kilner, Kiebel, & Friston, 2005). Random field theory assumes that the error field conforms to a lattice approximation with a multivariate Gaussian structure, and that these fields have a differentiable and invertible autocorrelation function.

In the absence of a prior hypothesis about which region(s) might respond to a particular experimental manipulation, i.e. if one is interested in exploring activations across the entire volume of brain scanned, then random field theory should be applied to correct for multiple comparisons across the whole brain, and any significant effects are thereafter reported whole-brain corrected. However, if there is a specific interest in how a manipulation may affect activity in a specific region (perhaps informed by previous experiments), then one can predefine functional or anatomical regions of interest (ROI), and correct for comparisons by only taking into account the number of voxels contained within those regions. Finally, an alternative approach is to derive the average effect (beta) across voxels in an ROI, and perform a single statistical test, bypassing the requirement for multiple comparisons and the associated correction.

In practise, fMRI experiments are constructed with a specific design type in mind. The simplest design involves subtraction between two experimental tasks or conditions. For example, if a task involves reacting to a stimulus with an action that is either congruent or incongruent with the displayed stimulus, then subtracting those conditions will identify regions of the brain whose activity is up-(or down)-regulated by congruency. Similarly, one can use multifactorial designs to embed subtractions and allow assessment of how one experimental factor influences another. The most common multifactorial design is a 2 x 2 factorial where the experimenter manipulates two independent factors, each with two levels, e.g. congruent versus incongruent trials, and high reward versus low reward trials. This allows one to assess interactions in addition to main effects.

Perhaps more powerful than subtraction designs are parametric designs. Here, the magnitude of a particular task-relevant quantity is varied over events (or trials), allowing one to test for regions of the brain that are sensitive to that change, or where activity shows a linear correlation with the magnitude of the manipulated quantity. For example, in the context of value-guided decision-making, one can look for brain regions that might track a participant's subjective value for a particular choice option. In parametric designs, the onset regressors in the design matrix are 'parametrically modulated' by the variable of interest, and this variable then becomes an additional regressor in the multiple linear regression. The resulting beta describes the steepness of the slope or correlation between BOLD and the given variable. Thus, if the beta is not significantly different from zero, then there is no effect. Parametric designs are especially useful when trying to identify brain responses that correlate with potentially rich and complex variables derived from computational modelling.

In value-guided decision-making experiments, one might dynamically manipulate a feature of the environment relevant for making-decisions, such as the uncertainty surrounding a reward outcome, and use computational modelling to characterise how participants use this information to make choices. One might also want to test whether the brain tracks this quantity. By modelling a set of key parameters that change according to the complexities of the task environment, one can use parametric designs to assess whether these variables are tracked in the brain.

3.3.3 Computational modelling

In recent years, cognitive neuroscience has seen a rise in the use of quantitative mathematical models to describe, predict and explain peoples' behaviour (Lewandowsky & Farrell, 2011). In general, computational modelling requires the experimenter to conceive a number of possible underlying mechanisms or processes for how a particular set of choices or behaviours emerge. For example, in the context of reinforcement learning, one might conjecture that the expected value of a chosen stimulus is updated according to the difference between the expected outcome (based on previous trials), and the actual outcome, governed by a learning rate. One can then test to see whether the trial-by-trial choices predicted by such a model are a good match to empirical data.

3.3.3.1 Model fitting

Often, computational models have a degree of algorithmic flexibility, in that the parameters governing each algorithm are free to vary across subjects in a manner that maximizes model evidence. A common method for evaluating model evidence (and indeed the method employed in this thesis) is maximum likelihood estimation (MLE). In brief, for a given data point y, maximum likelihood estimation determines the probability or probability density of observing y given a model, M, and a vector of parameter values ϑ. By varying the values in

ϑ, one can characterize how the observed likelihood changes in response to changes in parameter values, providing a measure of likelihood for each possible parameter value. Importantly, this allows one to determine the set of parameters with the highest likelihood.

The most precise method underlying this, named grid search, is to plot the likelihood surface (for each combination of parameter values), and determine those values that correspond to the peak of the surface. However, in practise, grid search can be both inefficient and computationally expensive, particularly when the model in question contains a large number of free parameters. Thus, a number of techniques have been developed that approximate grid search in a far more efficient manner. In this thesis, I employ the simplex method, in which a simplex of parameter values, which starts at a location defined by the experimenter (typically random), attempts to locate the minimum on the error surface (the point of maximum likelihood) by variously reflecting, contracting, or expanding within the parameter space. During a reflection, the point with the greatest discrepancy (worst fit) is flipped to the opposite side, which may then cause the simplex to expand (if it is in a rewarding direction). Conversely, in a contraction, the point with the worst fit moves closer to the centre of the simplex. Parameter estimation is typically performed on each individual subject (under the assumption that each individual is independent and drawn randomly from the population), and can thus be a powerful tool for assessing between-subject, or between-group variability in the associated processes.

In the instance where there is more than one conceivable model, the principles of Bayesian statistics and maximum likelihood (Bayesian information criterion, BIC) can be applied to determine the 'best' model. In BIC, a penalty term is introduced for the number of free parameters within a given model (which protects for overfitting) as follows:

$$BIC = -2 \ln L\,(\theta | y, M) + k \ln N$$

where *L* is the value that minimizes the negative log likelihood of the parameter set given the data (and the model), and *N* is the number of data points on which the likelihood calculation is based.

In summary, computational models allow for detailed interpretations and insights that few other approaches can match, and are particularly relevant in the context of value-guided decision-making.

3.3.3.2 Hierarchical Bayesian procedures

Recently there has been a shift away from conventional (fixed-effects) approaches to model fitting in favour of hierarchical Bayesian (random-effects) methods. The key principle behind this approach is to use the population-level distribution of data to constrain unreliable parameter estimates at the individual level. Here I outline one approach named Expectation-Maximization (E-M) (Huys et al., 2011). Typically, one estimates the maximum-likelihood hyperparameters given the data from a group of *N* subjects:

$$\hat{\vartheta}^{ML} = argmax_{\vartheta}\, p(C_1 \dots C_N|\vartheta) = argmax_{\vartheta} \prod_i p(\,C_i|\vartheta)$$

where

$$p(C_i|\vartheta) = \int d\,\theta_i\, p(C_i|\theta_i)p(\theta_i|\vartheta)$$

where ϑ is a parameter vector, and *C* is a vector of choices for each subject *i*

Thus, on each iteration, the posterior distribution over the group for each parameter is used to specify the prior over the individual parameter fits on the next, *k*th, iteration:

$${\theta_i}^{(k)} = argmax_{\theta}\, p(C_i|\theta_i)p\big(\theta_i\big|\vartheta^{(k-1)}\big)$$

It is often assumed that the likelihood surface is normally distributed around the maximum a posteriori parameter estimate, in which case a Laplace approximation can be applied:

$$p(\theta_i|C_i) \approx N\left({\theta_i}^{(k)}, \Sigma_i^{(k)}\right)$$

where $\sum_i^{(k)}$ is the second moment around ${\theta_i}^{(k)}$, which approximates the variance. In the M-step, the estimated hyperparameters $\vartheta^{(k)}$ of the normal prior distribution, mean μ, and factorized variance, σ^2, are updated as follows:

$$\mu^{(k)} = \frac{1}{N}\sum_i {\theta_i}^{(k)}$$

$$\left(\sigma^{(k)}\right)^2 = \frac{1}{N}\sum_i \left[\left({\theta_i}^{(k)}\right)^2 + \Sigma_i^{(k)}\right] - \left(\mu^{(k)}\right)^2$$

With this method, models are typically compared using integrated BIC (BIC$_{\text{int}}$) which penalises for the number of estimated free parameters:

$$BIC = -2\ln IL\,(\theta|C, M) + k \ln N$$

Note however, that in contrast to conventional BIC, $\ln IL\,(\theta|C, M)$ is a sum over the model evidence at the subject level by integrating over subject-level parameters.

Random-effects model fitting has a distinct advantage over conventional fixed-effects in that the contribution of unreliable subjects to the group mean is effectively down-weighted, and is thus utilized in all experiments reported in this thesis. However, one potential pitfall of this method is that it relies on the assumption that parameter estimates are normally distributed at the group-level.

CHAPTER 4

NEURAL MECHANISMS SUPPORTING ADAPTIVE DECISION-MAKING

Actions can lead to an immediate reward or punishment and a complex set of delayed outcomes. Adaptive choice necessitates the brain track and integrate both of these potential consequences. Here, I designed a sequential task whereby the decision to exploit or forego an available offer was contingent on comparing immediate value and a state-dependent future cost of expending a limited resource. Crucially, the dynamics of the task demanded frequent switches in policy based on an online computation of changing delayed consequences. I found that human subjects choose on the basis of a near-optimal integration of immediate reward and delayed consequences, with the latter computed in a prefrontal network. Within this network, anterior cingulate cortex (ACC) was dynamically coupled to ventromedial prefrontal cortex (vmPFC) when adaptive switches in choice were required. The results suggest a choice architecture whereby interactions between ACC and vmPFC underpin an integration of immediate and delayed components of value to support flexible policy switching that accommodates the potential delayed consequences of an action.

4.1 Introduction

As actions can lead to an immediate reward or punishment and a complex set of delayed consequences, it follows that to ensure the outcome of an action is optimal an agent needs to account for both immediate rewards and delayed consequences, which together constitute long-term expected value. A growing understanding of how hierarchical goals influence value comparison (Hare et al., 2009; Hare, Malmaud, & Rangel, 2011) contrasts with a dearth of knowledge regarding how the brain infers and integrates downstream consequence when evaluating options in a changing environment.

Paradigms requiring calculations of long-term value recruit the prefrontal cortex (Balleine & Dickinson, 1998; Basten et al., 2010; Glascher et al., 2010; Rangel & Hare, 2010; Wallis & Miller, 2003). In particular, the dorsolateral prefrontal cortex (DLPFC) has been linked to task planning (van den Heuvel et al., 2003; Wunderlich, Dayan, et al., 2012), the representation of abstract task rules (Buschman, Denovellis, Diogo, Bullock, & Miller, 2012; Stokes et al., 2013), as well as discounted or goal values (McClure et al., 2004; Plassmann et al., 2010). However, these studies do not address how the brain infers long-term value when decisions are sequential and integrative. It is of interest that several tasks requiring cognitive control implicitly evoke representations of downstream consequence, and as such it seems plausible that these processes could be subserved by a common neural mechanism.

In a typical example, an external cue signals a categorical contingency switch that instantiates a change in action or the inhibition of a prepotent response (Kerns et al., 2004). Although such tasks highlight a fronto-parietal network as being central to control (Badre, 2008; M. M. Botvinick et al., 2001), they are seldom deployed in the value domain, and a focus on isolated choice neglects downstream consequences of decisions. Recent studies have touched on these issues implicating parietal regions and PFC in representing the state-transitions necessary for building a model of the world (Glascher et al., 2010; Wunderlich, Dayan, et al., 2012). It remains unclear what computational role these regions play when action control is reliant a subjective inference about a change in expected value.

Here, I tested whether a context-specific evaluation of action could explain choice in a novel value-guided sequential go/nogo paradigm, whereby an agent tracks time-varying contingencies of a dynamic environment to adapt behaviour in anticipation of future value. Crucially, the dynamics of the task demanded frequent switches in policy based on an online computation of changing delayed consequences. Building on previous studies my paradigm allowed comparisons between policy switches arising either from inference, or by an external cue, that the environment had changed. Thus, by using functional magnetic resonance imaging (fMRI), I could characterize the computations tracked by the brain in a dynamic world. I predicted PFC would compute the downstream consequence of acting by tracking changing aspects of the environment, and interact with regions such as vmPFC and striatum, both strongly implicated in reward (Kable & Glimcher, 2007), to compute an integrated signal of long-term value for guiding choice policy.

4.2 Methods

Subjects

21 adults participated in the experiment (9 male and 12 female; age range 19-28; mean 23.2, SD = 2.3 years). All were healthy, reporting no history of neurological, psychiatric or other current medical problems. Subjects provided written informed consent to partake in the study, which was approved by the local ethics board (University College London, UK).

Training paradigm

In a conditioning phase, performed outside of the scanner, subjects learnt stimulus-reward associations between a set of four differently coloured rectangular cues and their respective monetary values. Each coloured rectangle corresponded to one of four possible value outcomes - 1, 2, 3 or 4 tokens - randomized across individuals. Subjects were instructed that each token would translate into a fixed sum of money at the end of the experiment.

Each trial began with a central fixation cross presented for 1000 ms, followed by presentation of a random pair of coloured boxes, one appearing to the left of the screen and one to the right. Subjects had a 2000 ms time-window to choose between these two boxes via a left or right button press, followed by presentation of the outcome of their choice for 1000 ms. The outcome was revealed as a written message indicating the total number of tokens won.

Subjects were instructed to explore all options until they were confident they had learnt all four associations, after which they should choose the box from the pair with the higher value. Each trial was defined as either correct if the subject chose the more valuable of the two options, and incorrect if they failed to do so. To ensure adequate learning, performance was calculated over six bins of twenty trials, with all subjects reaching a performance criterion of >= 90% by trial 60 onwards. For absolute verification, subjects were asked to verbally communicate the nature of the learnt associations.

Task paradigm

On every trial subjects were presented with a random sequence of trained stimuli (see training paradigm, p. 77), appearing individually and sequentially, with a variable inter-stimulus interval (750 - 1250 ms). The sequence order was pseudo-random and thus unpredictable, with each stimulus having an equal probability of being one of the four possible colours. In addition, the precise number of stimuli to be offered on any trial was uncertain, fluctuating under a uniform distribution between 3 and 7.

Each stimulus constituted an offer with a worth equivalent to its respective token value, for which subjects had 1500 ms to accept or reject via a go or nogo response respectively. A restriction was placed on the number of offers that could be exploited. In high constraint trials (HC), the acceptance budget was between 1 and 3 offers, whilst in low constraint (LC) it ranged between 4 and 6 offers, both varying under a uniform distribution independent of the total number of offers made on the current trial. Subjects were not explicitly told the bounds of the distributions from which the number of offers and total budget were drawn, only that they were uniform. All subjects received 30 training trials (15 per condition) in order to infer these distributions and familiarize themselves with the task attributes.

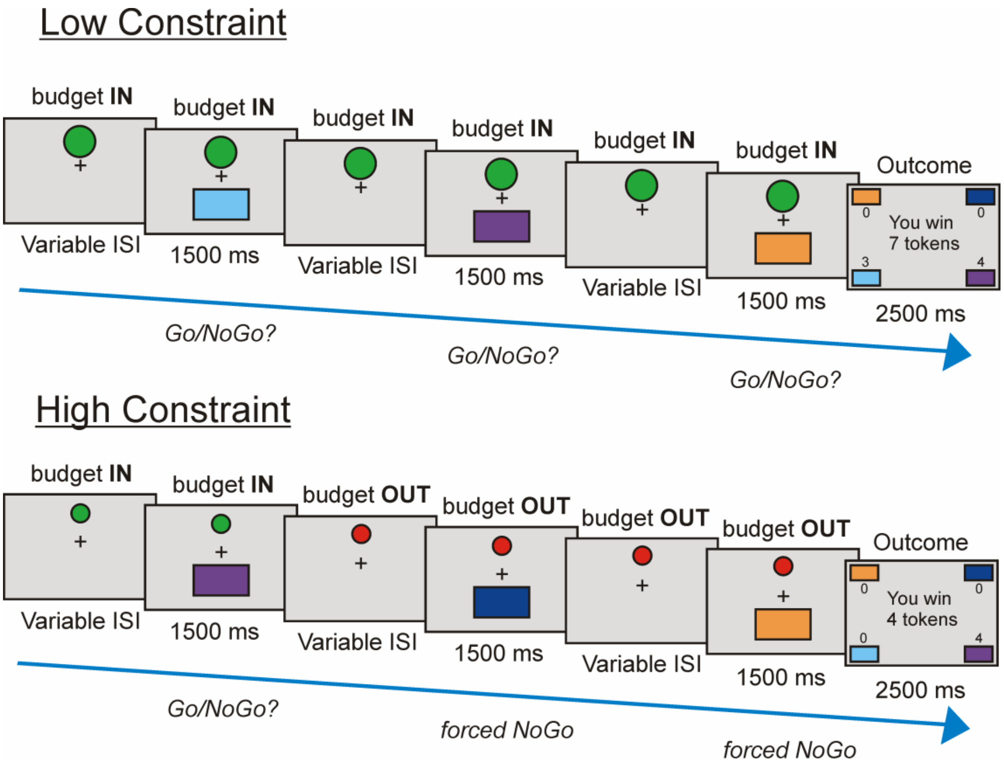


Figure 4.1 Subjects learnt stimulus-value associations, ranging from 1 to 4 tokens, for four collared stimuli. On every trial participants saw a random sequence of these stimuli, varying unpredictably in length between 3 to 7, with each stimulus representing an offer requiring either a go response to win the associated tokens or a nogo response for no token gain (for simplicity, the illustrations span 3 offers). Subjects had a predetermined go budget that placed a restriction on the number of offers that could be accepted. In a low constraint context (LC) subjects could accept between 4-6 offers, but only between 1-3 in a high constraint context (HC), with the exact budget being uncertain. Upon exhausting a go budget, nogo responses were enforced for the remainder of the trial. The context or condition was cued via a large (LC) or small (HC) green circle, whilst a depleted budget was signalled via the green circle turning red.

HC and LC trials were pseudo-randomly interleaved. The trial type was indicated via a small or large green circle, in the top central portion of the screen, for HC and LC respectively. This appeared at trial onset and turned red upon exhaustion of the budget indicating nogo responses were obligatory for the remainder of the trial. After the final offer, an outcome incorporating the total number of tokens won, and corresponding cue-token credit breakdown was revealed for 2500 ms.

120 trials (60 per condition) were completed in the scanner across four sessions. The number of tokens won across sessions was summed and converted to a cash prize.

Behavioural data analysis

Global behaviour

My analysis focused exclusively on choices pertaining to within-budget offers. Accepts (go responses) were obtained as a percentage of the total offer number at each offer value, conditional on HC and LC trials. These measures were entered into a two-way repeated-measures analysis of variance (ANOVA) with factors control (HC/LC) and offer value (1, 2, 3 or 4). The data were analysed in the statistical software package SPSS, version 20.0.

Within-trial modulation of choice

Within a trial, a player transitioned through a number of discrete states dependent on two fluctuating variables, the number of offers already seen and the number of accepts already utilized. To assess whether the probability of accepting a given offer was flat across the entire length of a given trial or fluctuated as a function of these variables, I split trials by offer index (i.e. 1-7) and number of offers already rejected (i.e. 0-6), re-calculating the probability of accepting at every possible permutation (see Figure 4.3, p. 91). For each participant, I summed the number of offers with a given value presented at each possible state within a trial, and then summed the number of accepts at each of those states. Dividing these measures provided a probability of acceptance at every choice point. Thus, for both HC and LC trials and each offer value, I generated a separate probability accept matrix with offer number increasing along the x-dimension, and number of rejects increasing along the ydimension. These matrices were averaged across all participants. For display purposes, I discarded cells with less than a total of 10 data points.

Computational modelling

As I was interested in assaying subjects' strategy for maximizing reward, I evaluated evidence for four competing choice models. Broadly, I conjectured subjects might approach trials with a predetermined decision rule, in effect applying a heuristic uniformly throughout a trial. Alternatively, owing to uncertainty surrounding the number of expected offers and the go-budget (the number of offers they can exploit for reward in a trial), subjects might continually adapt their threshold for accepting offers across a trial. I outline the distinct models below, ordered by increasing complexity, where each model calculated the value of accepting an offer which was then passed through a sigmoid function to determine action probabilities as follows:

$$P_A = \frac{1}{1 + \exp(-\tau \cdot V_A)}$$

where V_A is the expected value of accepting an offer, and τ is a temperature parameter that governs the stochasticity of choices.

Baseline heuristic model

I first specified a baseline heuristic model that calculates the value of accepting (V_A) by comparing the (face) value of every offer to a stationary decision threshold:

$$V_A = R - c_1$$

where R is the (face) value of the current offer and c_1 is a value threshold.

Thus, this model makes choices based solely on the immediate (face) value of an offer with the probability of acceptance fixed throughout a trial.

The model has 3 free parameters: the associated decision threshold for both HC and LC separately, and the steepness of the sigmoid function.

Sliding offer model

I conjectured subjects might track the number of offers seen in a trial and adjust a decision threshold such that an offer is more likely to be accepted if forthcoming offers were scarce. I added a linear slope parameter to the baseline heuristic model that governed the steepness of this decay across a trial, such that:

$$V_A = R - (c_1 - o \cdot c_2)$$

where *R* is the (face) value of the current offer, c_1 is a value threshold, *o* is the current offer index and c_2 governed the steepness of the associated slope.

The model has 5 free parameters: the associated decision threshold and a slope parameter for both HC and LC separately, and the steepness of the sigmoid function.

Sliding budget model

A second variable that subjects could track in order to dynamically adjust their decision threshold is the number of offers already accepted in a trial. Given a limited go budget, a player may be less likely to accept an offer as this resource is exhausted, assuming ample offers. This model linearly increased the decision threshold with every additional offer accepted, but did not take into account the abundance of remaining offers, such that:

$$V_A = R - (c_1 + a \cdot c_2)$$

where R is the (face) value of the current offer, c_1 is a value threshold, a is the number of offers previously accepted and c_2 governed the steepness of the associated slope.

The model has 5 free parameters: the associated decision threshold and a slope parameter for both HC and LC separately, and the steepness of the sigmoid function.

Integrated sliding model

Combining the sliding offer and sliding budget models, subjects could track both the number of offers seen and the number of offers already accepted in a trial, using each source of information to adjust the decision threshold. The threshold should drop linearly with every mounting offer and rise linearly with every mounting go response. I fit separate slope parameters that governed the linear gradient for the number of offers and number of accepts, such that:

$$V_A = R - (c_1 + a \cdot c_2 - o \cdot c_3)$$

where R is the (face) value of the current offer, c_1 is a value threshold, a is the number of offers previously accepted, o is the current offer index, and c_2 and c_3 govern the steepness of the associated slopes.

Interestingly, this 2-factor model predicts the optimal action with a frequency of 87% (based on group mean parameter fits).

The model has 7 free parameters: the associated decision threshold, a slope parameter for the number of offers, a slope parameter for the number of accepts for both HC and LC separately, and a parameter for the steepness of the sigmoid function.

Model comparison

As described previously (Guitart-Masip et al., 2012; Huys et al., 2011) I used a hierarchical Type II Bayesian (or random-effects) procedure using maximum likelihood to fit simple parameterized distributions for higher level statistics of the parameters. Since the values of parameters for each subject are 'hidden', this employs the Expectation-Maximization (EM) procedure. Thus, on each iteration the posterior distribution over the group for each parameter is used to specify the prior over the individual parameter fits on the next iteration. For each parameter I used a single distribution for all participants. Before inference, all parameters were suitably transformed to enforce constraints (log and inverse sigmoid transforms).

Models were compared using the integrated Bayesian Information Criterion (iBIC), where small iBIC values indicate a model that fits the data better after penalizing for the number of parameters. Comparing iBIC values is akin to a likelihood ratio test (Kass & Raftery, 1995).

Reaction time analysis

I conjectured that if subjects were evaluating choice options in light of an action threshold that fluctuated in accordance with the number of offers already seen and accepted/rejected, then reaction times should be faster when the associated threshold is low and a go response is relatively more valuable. To test this, I utilized multiple linear regression to model the dependence of reaction times for all go choices on the corresponding offer values (immediate values) and model thresholds, separately for HC and LC trials. The two regressors were forced to compete for variance so as to explore dissociable contributions to the observed reaction times.

fMRI data acquisition

fMRI was performed on a 3-Tesla Siemens Quattro magnetic resonance scanner (Siemens, Erlangen, Germany) with echo planar imaging (EPI) and 32-channel head coil. Functional data was acquired over four sessions containing 166 volumes with 48 slices (664 volumes total). Acquisition parameters were as follows: matrix = 64 x 74; oblique axial slices angled at -30° in the antero-posterior axis; spatial resolution: 3 x 3 x 3 mm; TR = 3360 ms; TE = 30 ms. The first five volumes were subsequently discarded to allow for steady state magnetization. Field maps were acquired prior to the functional runs (matrix = 64 x 64; 64 slices; spatial resolution = 3 x 3 x 3 mm; gap = 1 mm; short TE = 10 ms; long TE = 12.46 ms; TR = 1020 ms). Anatomical images of each subject's brain were collected using multi-echo 3D FLASH for mapping proton density (PD), T1 and magnetization transfer (MT) at 1mm$_3$ resolution and by T1 weighted inversion recovery prepared EPI (IR-EPI) sequences (spatial resolution: 1 x 1 x 1 mm) with B1 mapping data to correct for the effect of inhomogeneous transmit fields on the T1 maps (3D EPI Transverse partition direction; matrix = 64 x 48; phase direction right to left; 48 partitions; resolution = 4 x 4 x 4 mm).

During scanning peripheral measurements of subject pulse and breathing were made together with scanner slice synchronization pulses using the Spike2 data acquisition system (Cambridge Electronic Design Limited, Cambridge UK). The cardiac pulse signal was measured using an MRI compatible pulse oximeter (Model 8600 F0, Nonin Medical, Inc. Plymouth, MN) attached to the subject's finger. The respiratory signal (thoracic movement) was monitored using a pneumatic belt positioned around the abdomen close to the diaphragm.

fMRI data analysis

Data were analysed using SPM8 (Wellcome Trust Centre for Neuroimaging, UCL, London). Functional data were bias corrected for 32-channel head coil intensity inhomogeneities. Pre-processing involved realignment and unwarpping using individual fieldmaps, co-registration of EPI to T1w images, and spatial normalization to the Montreal Neurology Institute (MNI) space using the segmentation algorithm on the T1w image with a final spatial resolution of 1 x 1 x 1 mm. Finally, data were smoothed with an 8mm FWHM Gaussian kernel. The fMRI time series data were high-pass filtered (cutoff = 128 s) and whitened using an AR(1)-model.

For each subject I used an in-house Matlab toolbox (Hutton et al., 2011) to construct a physiological noise model to account for artefacts that take account of cardiac and respiratory phase as well as changes in respiratory volume. This resulted in a total of 14 regressors which were sampled at a reference slice in each image volume to give a set of values for each time point. The resulting regressors were included as confounds in all first level GLMs.

In order to identify brain areas sensitive to within-trial variations in choice prescribed by my model, I derived an offer-wise go threshold to use as a parametric modulator of offer onsets in all first level GLMs. This model threshold (MT) represented an intercept value that increased linearly with every offer accepted and decreased linearly with every offer seen. The intercept and slopes were based on the mean posterior parameter fits across the group. If the offer value was higher than the MT the preferable decision is accepting, otherwise rejecting is preferred.

Below I outline the GLM constructed for first level analyses. All imaging analyses address time-points when offers are within-budget and the subject has a free choice. Results are reported whole-brain corrected at the cluster level (FWE p =< 0.05) unless otherwise stated.

To explore a main effect of action constraint and value / MT (and their relevant interactions) I split offer onsets according to constraint (HC / LC) and offer (face) value (1, 2, 3 or 4), modelling each in a separate regressor parametrically modulated by MT. This resulted in 16 regressors of interest. The four scanning sessions were concatenated into one, and a binary matrix was included to encode the identity of each session. Additional regressors of no interest included six movement-related covariates (the three rigid-body translations and three rotations resulting from realignment), 14-physiological regressors (6 respiratory, 6 cardiac and 2 change in respiratory/heart rate), the onsets of the go responses (to explain away the effects of action), all offers outside of budget (for which 'nogo' responses were enforced) parametrically modulated by offer value, and outcome onsets parametrically modulated by the relevant number of tokens won. All regressors were modelled as stick functions with duration of zero and convolved with a canonical form of the hemodynamic response function (HRF) combined with time and dispersion derivatives.

At the second level I conducted a random-effects 2 x 4 ANOVA with factors condition (HC / LC) and offer value (1, 2, 3 or 4), using first-level contrast images corresponding to the onset regressors of interest for each participant. This enabled me to explore main effects of condition and value, and their interaction. I generated a second 2 x 4 random-effects ANOVA drawing on first-level contrast images from the 8 MT parametric modulators, to explore an average effect of MT and a MT x value interaction. In order to obtain an average estimate of DLPFC activation in HC compared to LC, parameter estimates for offer values 1-4 were averaged in each condition, and LC was subtracted from HC.

Functional regions of interest

I used a functional regions of interest (f-ROI) approach to extract parameter estimates in a priori regions for a subset of analyses, including correlating neural and behavioural measures, comparing value representations between conditions and exploring functional connectivity patterns. Functional ROIs were derived by identifying significant clusters of activation surrounding peak voxels from the relevant whole-brain mass univariate analysis. Given these clusters often spanned multiple regions, activations were constrained to corresponding anatomical ROIs from the MarsBar toolbox (V. 0.42) for SPM. For the VS, activations were constrained to an anatomical ROI derived from a diffusion tensor imaging connectivity-based parcellation of the right nucleus accumbens (NA) in humans, taken from (Baliki et al., 2013). The ROI consisted of both the core and shell subcomponents of NA and the right region was flipped along the x-dimension in the MarsBar toolbox to obtain a bilateral accumbens mask.

Psycho-physiological interaction

For each subject I defined a volume of interest (VOI) that included all active voxels (at p = 0.2) from a first-level contrast that specified a linear effect of model thresholds across offered value { -2 -1 1 2 } within f-ROIs derived from the same second-level contrast (see Figure 4.6A, black arrows, p. 98). This allowed me to define voxels active on a subject-by-subject basis, but confined to the cluster active at the group-level. I noted that 1 out of 21 subjects had no active voxels when specifying both the ACC and left DLPFC (BA46) as seeds, while 3 out of 21 subjects had no active voxels when specifying dorsal vmPFC as a seed. These subjects were excluded from the corresponding PPI analysis. I used the generalized PPI toolbox for SPM (gPPI; http://www.nitrc.org/projects/gppi) to create a new GLM in which the individual seed time-course was deconvolved to construct a neuronal time-course for multiplication with regressors modelling all task effects, and then reconvolved with the HRF. Thus, the gPPI GLM includes a psychophysiological regressor for all conditions (McLaren, Ries, Xu, & Johnson, 2012). An indicator function for the relevant contrast, the original BOLD eigenvariate, 6 motion and 12 physiological parameters were included as additional regressors.

I first looked for regions in which connectivity with the seed region was modulated by MT, but where this modulation was greater for offers requiring adaptive control (values 1 and 2 in HC, and value 1 in LC > values 3 and 4 in HR, and values 2, 3 and 4 in LC). I also performed a second PPI restricted to offers requiring adaptive choice (values 1 and 2 in HC, and value 1 in LC), to ascertain whether connectivity increased (positive PPI) or decreased (negative PPI) with respect to increases in MT (compared to zero). One-sample t-tests were performed on the relevant contrasts at the second-level.

4.3 Results

Subjects reject lower value offers when a go budget is scarce

Subjects were sensitive to both immediate (face) value and the delayed consequences arising from a budget constraint. Higher value offers were accepted more than lower value offers (a main effect of value: F(1,68, 33.63) = 277.87, MSE = 379.38, $p < 0.001$) and more offers were accepted overall in LC compared to HC (a main effect of constraint: F(1, 20) = 182.70, MSE = 45.69, $p < 0.001$). Importantly, subjects were less willing to accept low value offers in HC compared to LC (a budget constraint x value interaction: F(1,73, 34.67) = 30.41, MSE = 136.19, $p < 0.001$) (see Figure 4.2B, p.90).

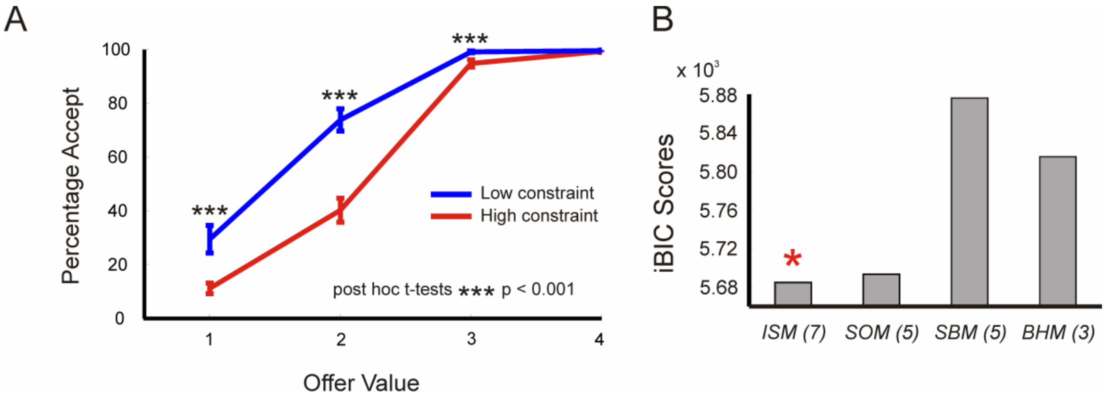


Figure 4.2 (**A**) Plot shows the mean percentage of offers accepted split by token value and condition (HC in red, LC in blue). Subjects were less willing to accept low value offers when the budget was scarce. Post-hoc paired t-tests revealed significant decreases in percentage accept for offer values 1, 2 and 3 in HC compared to LC (all p < 0.001). Vertical lines represent SEM. (**B**) Integrated BIC scores (for the group as a whole) show that a model in which both the number of offers already seen and number of offers already accepted/rejected are used to adjust the threshold for action fits behaviour best. *ISM = Integrated sliding model; SOM = Sliding offers model; SBM = Sliding budget model; BHM = Baseline heuristic model*. The number of free parameters built into each model is indicated in parentheses.

Dynamic versus fixed control

Subjects dynamically adjusted their responses when delayed consequences fluctuated within a trial. These consequences depended on both the number of offers already seen and the number previously accepted/rejected in a trial. Figure 4.3 (p. 91) illustrates that subjects utilized both these components to adjust their responses.

I next quantified this effect by comparing models accounting for the number of previous offers, number of previous accepts, or both (see Methods, p. 81 - 84). I found strong evidence that the integrated sliding model, wherein both components contribute to choice, fitted subject data best at the group level (lowest iBIC score). Although the sliding offer model performs well (in which only the number of offers seen is used to adjust choice), an addition of tracking the number of accepts/rejects improved the maximum likelihood across every subject (Wilcoxon signed rank test, p = 5.96 x 10^{-5}). Consistent with the notion that subjects used a dynamic control strategy, reaction times were faster when action (model) thresholds from the winning model were low (mean beta HC: 192.5, p < 0.0001; mean beta LC: 320.1, p < 0.0001), controlling for the immediate (face) value of the current offer (mean beta HC: -107.3, p < 0.0001; mean beta LC: -92.0, p < 0.0001).

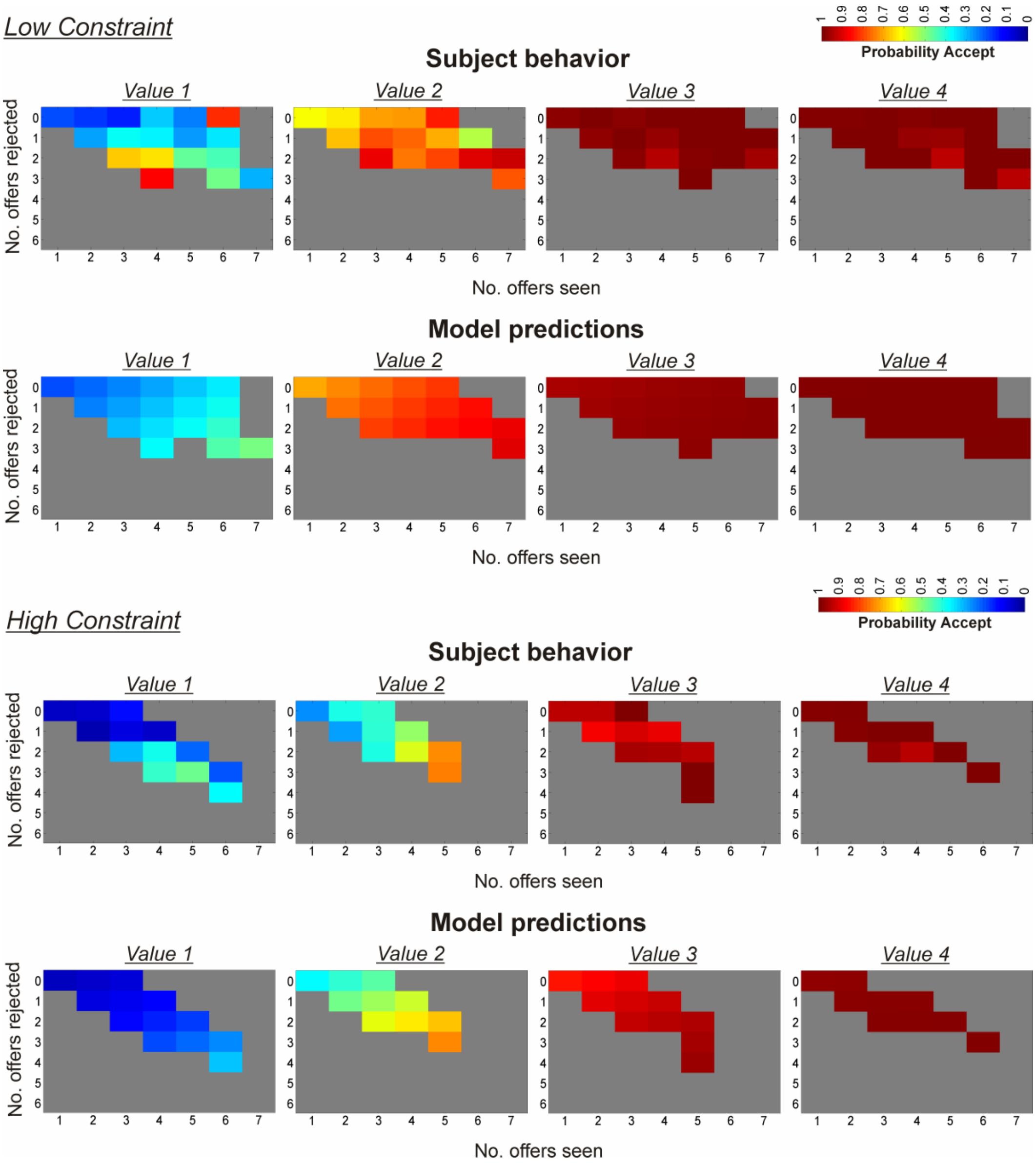


Figure 4.3 Subjects adjust the probability of accepting less desirable offers as a function of the number of offers seen (x-axis) and number of offers already rejected (y-axis). The spectrum runs from blue (probability 0) to red (probability 1).

fMRI neuroimaging

As in other control paradigms (Barber & Carter, 2005; Kerns et al., 2004), I first performed a categorical comparison to identify brain regions more active when the overall demand for control is increased (HC > LC), averaging across offer values (see Methods, *fMRI data analysis*, p. 86 ; see Table 4.1, p. 105). I found greater whole-brain corrected activity in right DLPFC and bilateral superior parietal lobule in HC overall compared to LC (Figure 4.4A, p. 93). These regions are associated with model-based planning (Owen, 1997; van den Heuvel et al., 2003; Wunderlich, Dayan, et al., 2012), task switching and cognitive control (Badre, 2008; M. M. Botvinick et al., 2001; Liston et al., 2006), the resolution of uncertainty (Yoshida & Ishii, 2006) and working memory (Barbey et al., 2012; Curtis & D'Esposito, 2003; Narayanan et al., 2005).

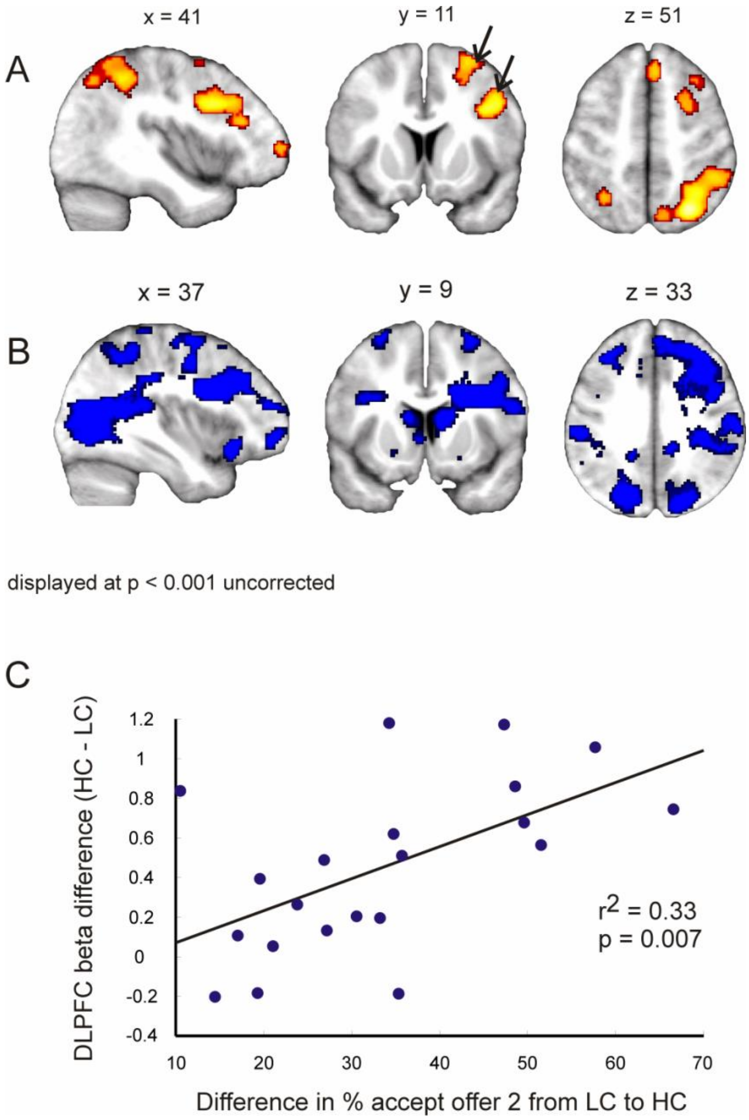


Figure 4.4 Distinct but overlapping fronto-parietal networks are recruited when action constraints increase and when the expected long-term value of an option increases. (**A**) A fronto-parietal network spanning right DLPFC and bilateral parietal cortex was more active in HC compared to LC trials, during offers subject to go/nogo. The black arrows indicate two DLPFC clusters that were combined to form a DLPFC f-ROI responding to HC > LC. (**B**) Model thresholds, denoting the long-term component of expected value, correlated negatively with BOLD in an overlapping fronto-subcortical-parietal network, including ACC, bilateral DLPFC, parietal cortex and striatum. Activity in these regions was highest when the value of conserving a unit of budget (rejecting) was low. (**C**) Subjects with greater right DLPFC recruitment (see panel **A**, black arrows, for DLPFC f-ROI) in HC compared to LC showed a larger adjustment in willingness to accept value 2 offers between conditions ($r^2 = 0.33$, p = 0.007). Each point represents one participant.

I next hypothesized that greater right DLPFC recruitment in HC compared to LC would result in a larger behavioural adjustment between conditions. I focused on value 2 offers for which I observed the largest change in behaviour between HC and LC. I derived an average parameter estimate for a HC > LC contrast in a right DLPFC functional ROI, combining two activated right DLPFC clusters (1078 total voxels; see Figure 4.4A, middle panel, black arrows, p. 93), averaging the betas for the four value regressors, and then subtracting LC from HC. A between-subject correlation revealed a positive association between parameter estimates in right DLPFC for a HC > LC contrast and the change in propensity to accept value 2 offers between HC and LC ($r^2 = 0.33$, $p < 0.007$) (Figure 4.4C, p. 93). Thus, right DLPFC is instrumental in the categorical adjustment of action control in my task.

To identify correlates of value for guiding choice, I tested for a positive average linear effect of offer (face) value across both HC and LC conditions, revealing a value-dependent response in regions that included vmPFC and VS (including nucleus accumbens) (Figure 4.5A, p. 95; see Table 4.1 for all regions, p. 105). Importantly, this value signal was independent of any motor response as go responses were modelled as separate onsets in my GLM. Thus, offer values were tracked in regions involved in value representation (Jenison et al., 2011; Schultz, 2000). Further, as participants' choices were sensitive to action constraint, I anticipated the representation of offer value in vmPFC and VS, two regions widely implicated in value-based choice (De Martino et al., 2013; Guitart-Masip et al., 2012; Hunt et al., 2012), would be modulated accordingly. I derived functional ROIs (see Methods, p. 88) by defining voxels (within whole-brain corrected clusters) in vmPFC (928 voxels; see Figure 4.5B, p. 95) and VS (56 voxels; see Figure 4.5C, p. 95) that showed a linear effect of offer value on average (as above), and then tested for an orthogonal value x condition (HC or LC) interaction. I found a significant interaction in vmPFC ($F(2.38, 47.62) = 5.34$, $MSE = 1.67$, $p = 0.005$) but not in VS (Figure 4.5B, p.95). In LC, vmPFC was more responsive to value 2 than value 1 ($p = 0.02$) and value 3 than value 2 (p = 0.02), whereas in HC, neither value 2 (p = 0.35) nor value 3 (p = 0.38) induced greater BOLD than value 1.

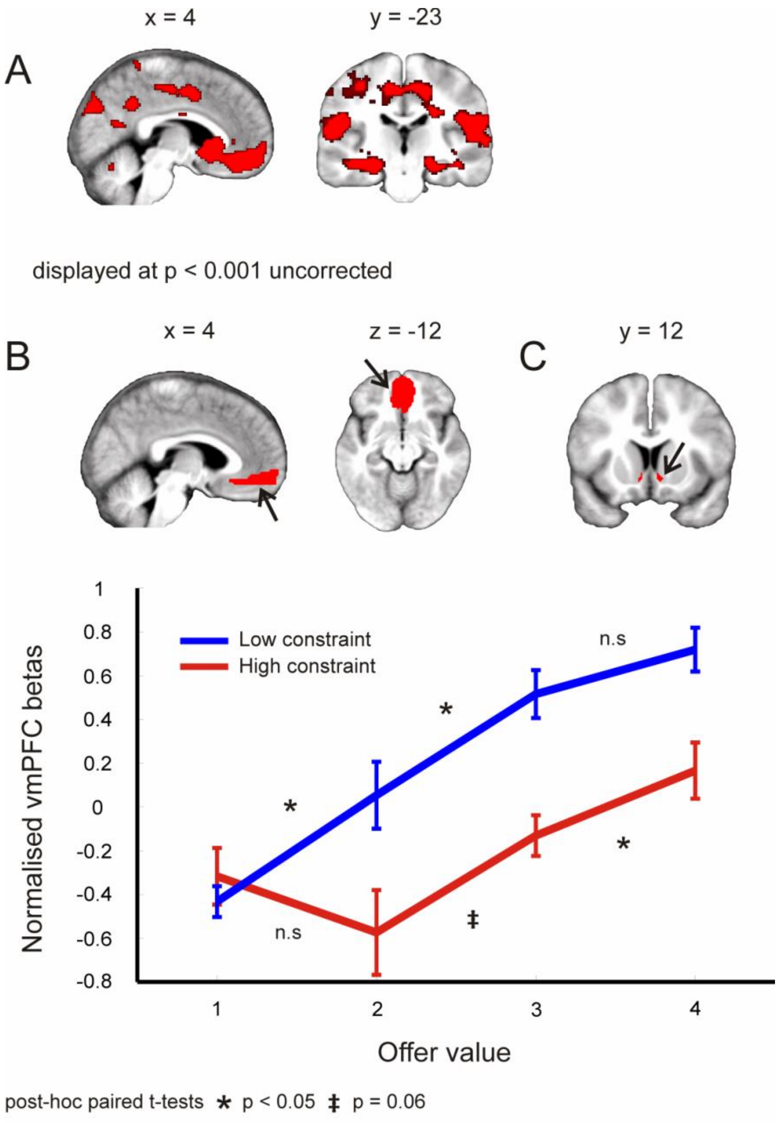


Figure 4.5 Value representations are modulated by context. (**A**) The BOLD signal in vmPFC, VS, right amygdala and precuneus/posterior cingulate covaries with offer value. (**B**) vmPFC tracks value linearly in LC but with a depressed slope for HC. The representation of value 2 offers is particularly degraded, mirroring behavioural data. Vertical lines represent SEM. (**C**) A f-ROI confined to the ventral striatum was used in a constraint (HC/LC) x value (1, 2, 3 or 4) interaction analysis.

Given behavioural and computational evidence that subjects used trial structure to evaluate options, I conjectured within-trial adaptive choice would manifest as a dynamic modulation of value representations in vmPFC, analogous to that observed between HC and LC trials. To test this I constructed a summary measure reflecting a time-varying decision threshold, as prescribed by the winning model, that then provided an offer-wise parametric regressor (see Methods, p. 86). In effect this model threshold (MT) represented the value of carrying one more unit of budget (the number of accepts endowed for a trial) into the next offer, independent of the immediate value of the current offer. The overall value of accepting was thus the difference between offer value and MT. Note however, in contrast to the down-regulation of value 2 offers in HC, the time-variant adaptation in choice prescribed by the winning model require an up-regulation of low value offers when the future benefit of conserving a unit of budget is low.

I first tested for regions where BOLD signal correlated with MTs across both conditions (see Methods, *fMRI data analysis*, p. 86) finding a fronto-subcortical-parietal network was modulated negatively, with no regions modulated positively. This is consistent with BOLD signal being highest when the expected utility of carrying a unit of budget forward was low, and thus a go response was more favourable. This network, that includes ACC, bilateral DLPFC, parietal cortex and striatum (Figure 4.4B, p. 93; see Table 4.1 for all regions, p. 105), is partially overlapping with that seen in the contrast of HC > LC (Figure 4.4A, p. 93), implying similar regions of PFC are recruited when action control is reliant on internal valuations versus external cues. Note that similar networks are engaged during working memory (Barbey et al., 2012; Curtis & D'Esposito, 2003) and in goal-directed and/or cognitive control paradigms (Badre, 2008; Hare et al., 2009; Rushworth et al., 2011; Yoshida & Ishii, 2006).

In my task, the immediate reward gained from accepting value 3 or 4 offers is higher than the maximum MT value and thus these offers should always be accepted. In contrast, the difference between the immediate reward obtainable from value 1 and 2 offers and their corresponding MTs fluctuates about zero, signifying choice policy, consistent with the observed behaviour, should shift in response to trial state. Consequently, I hypothesized that an independent network tracked MTs differentially dependent on offered value. To test this, I looked for brain regions showing a linearly increasing effect of MTs across both conditions. As MTs were tracked negatively this tested an hypothesis they would correlate more strongly with BOLD as offer value decreased. I found clusters in ACC, left DLPFC (BA46), and a dorsal region of vmPFC (BA10) (Figure 4.6A, p. 98; see Table 4.1 for details, p. 105) that were increasingly more responsive to changes in MTs as offered value decreased. The ACC cluster was particularly striking, with post-hoc exploratory one sample t-tests revealing MT representations solely for offers requiring adaptive choice, that is offer value 1 for both conditions (HC: $p = 0.002$, LC: $p = 0.01$) and a trend for offer value 2 for HC alone ($p = 0.09$) (Figure 4.6B, p. 98). Note that I found behavioural evidence of adaptive choice corresponding to these three offers (Figure 4.3, p. 91).

Finally, I used a connectivity analysis to ask whether brain regions tracking MTs for offers requiring policy switches were modulating value representations in vmPFC to instigate adaptive switches in choice. I selected physiological responses from three f-ROI seed regions, showing a linear effect of MTs (reflecting the long-term component of value), that included ACC (739 voxels in group-level ROI), left DLPFC/BA46 (502 voxels in group-level ROI) and dorsal vmPFC/BA10 (179 voxels in group-level ROI) (see Figure 4.6A, black arrows, p. 98). Interestingly, the PFC has previously been implicated in flexible action control, and, in the case of DLPFC, top-down modulation of value signals (Hare et al., 2009; Walton et al., 2007). I performed a PPI to test a hypothesis that coupling would be modulated by fluctuations in MTs, and that this change would be greater for low value offers requiring adaptive choice (values 1 and 2 in HC, and value 1 in LC) than for high value offers (where choice is not dependent on MT). The regions identified by the ensuing PPI correspond to regions whose connectivity with the relevant seed region depends on both the immediate value and MT of the current offer.

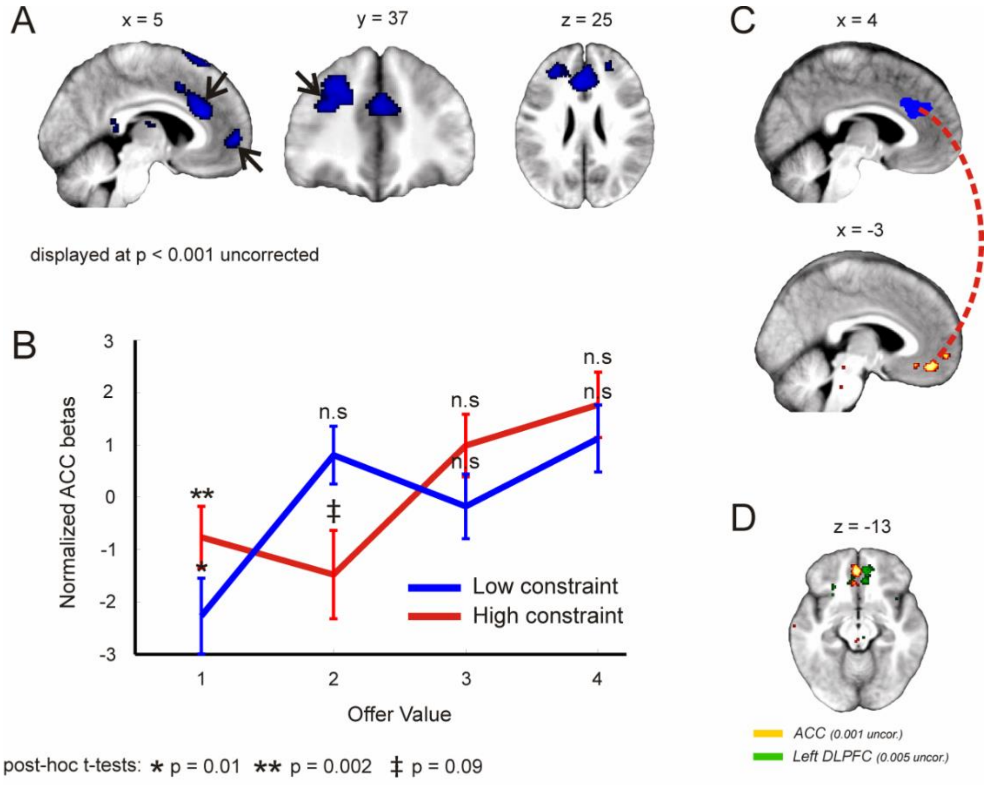


Figure 4.6 Model thresholds are selectively tracked in a prefrontal network. (**A**) BOLD signal in ACC, left DLPFC (BA46) and dorsal VMPFC (BA10) increases as model thresholds decrease (and action is most favourable), only for offers mandating adaptive control. (**B**) Parameter estimates from the ACC cluster shown in panel **A** illustrate model thresholds are tracked for offers requiring adaptive control (value 1 in HC and LC, and a trend for value 2 in HC). Red corresponds to HC; blue to LC. Vertical lines represent SEM. (**C**) A whole-brain voxel-based gPPI analysis revealed ACC is more functionally connected with the vmPFC when actions cost are high and low offers should be rejected. This region of vmPFC overlaps with a cluster that tracks offer value (Figure 4.5A, p. 95) and is sensitive to categorical changes in context (Figure 4.5B, p. 95). (**D**) Comparison of functional connectivity patterns between ACC (yellow; displayed at 0.001 uncorrected) or left DLPFC/BA46 (green; displayed at 0.005 uncorrected) and vmPFC. As with ACC, the left DLPFC demonstrates a functional coupling with vmPFC when accepting an option offering only a small immediate reward is unfavourable, but this effect only emerges at a more liberal threshold.

I found a functional coupling between ACC and vmPFC that was sensitive to fluctuations in MTs, that was larger on average for offers requiring adaptive choice. This effect was significant when using small volume correction for the vmPFC f-ROI that tracked offer value. Given directionality cannot be determined when comparing parametric effects across conditions, I performed a second PPI analysis, now confined to offers requiring adaptive choice, enabling me to assess whether connectivity was positively or negatively modulated by increasing MTs. A vmPFC f-ROI approach revealed that ACC and vmPFC were more functionally coupled when MTs were high (mean ppi = 3.04, p = 0.005), in other words when low value offers need to be rejected. Thus, connectivity between ACC and vmPFC was dependent on both immediate value and MT. Although the left DLPFC did not demonstrate functional coupling with vmPFC that depended on both MT and offer value, qualitatively I observed an effect in vmPFC at a more liberal threshold (p = 0.005 uncorrected). In fact, I did not detect any significant difference in the magnitude of the PPI effect (2-sample t-test, p = 0.68) between ACC and DLPFC when using a vmPFC f-ROI, implying that despite a more prominent contribution of ACC, DLPFC also contributes to the observed connectivity. When dorsal vmPFC was used as a seed, no significant results were observed.

4.4 Discussion

This study addressed the computational implementation of context-specific action control in value-guided choice. I show that subjects incorporate both extrinsic constraints on action and intrinsic fluctuations in opportunity to adaptively switch between a go/nogo response. Mechanistically, a fronto-subcortical-parietal network tracks the downstream consequence of spending a limited action budget, whilst ACC couples to vmPFC to shift the representation of value in favour of long-term profit.

In this task, subjects track the number of offers already seen and number already accepted/rejected in a trial to compute the future value of expending a unit of budget. This model fits behaviour better than simpler candidates in which action is driven solely by immediate reward or where only a restricted set of environmental features is consequential. Of interest, the winning model produces behaviour that closely approximates optimal choice, which relies on back-propagating through a decision tree of all future moves in a trial. Although this strategy is computationally taxing (given the depth of the search tree in this game), subjects could be computing long-term value by recruiting a model-based system that searches through future states 'on the fly' (Dayan, 2008). Alternatively, a player could track aspects of the environment to index stored values, or to update values under a model-free regime. Although my task cannot arbitrate between these possibilities, I note the circuitry that tracks the MTs from the winning model overlaps with that implicated in model-based reinforcement learning (Daw et al., 2011; Glascher et al., 2010; Wunderlich, Dayan, et al., 2012).

Influential accounts of ACC propose a myriad of roles including conflict monitoring (M. M. Botvinick, 2007), error monitoring (Rushworth et al., 2004), overriding pre-potent responses (Kerns et al., 2004), evaluating outcomes (Gehring & Willoughby, 2002) and action-outcome learning for negative feedback (Rushworth et al., 2004). While the task lacked explicit negative feedback, the finding that ACC tracks the MTs necessary for implementing adaptive choice is consistent with the conflict monitoring account, but not with a role in error monitoring, given choices were closely aligned with optimality. Unlike previous paradigms where switches in contingency are explicitly cued (Kerns et al., 2004), I show conflict in ACC can arise endogenously via tracking fluctuations in downstream consequence.

ACC is also implicated in foraging (Kolling et al., 2012) where it is proposed to track the value of alternative choice options during a trade-off between exploration and exploitation. I found ACC activity was highest when exploiting a low value offer was more optimal. However, in my task ACC only tracks MTs corresponding to offers that are routinely rejected. In this light, my findings can be construed as in keeping with the former role. These findings also hint that a conflict monitoring account of ACC can be reinterpreted as reflecting a need to switch behaviour from the current default response, as opposed to encoding a non-specific conflict signal (Shenhav, Botvinick, & Cohen, 2013). Indeed, recent work further supports the notion that ACC assumes a default frame of reference, by adapting choice from the best long-running option (Boorman, Rushworth, & Behrens, 2013).

A number of studies propose ACC expresses a prediction error (Ide et al., 2013),which can be used to update internally-generated models (O'Reilly et al., 2013). This may explain why high-conflict or high-volatility trials, often confounded with surprise, also induce responses in ACC. However, my data indicate that surprise cannot fully account for the ACC activation observed, as stimuli are presented with equal frequency such that surprise does not vary within a trial. Instead, a response to low value offers switches in line with changes in delayed consequence. Thus, in the context of the current study, it is likely that ACC plays a more general role in a strategic adjustment of behaviour that is rooted in processing or initiating atypical stimulus or action requirements, which also includes surprising events.

A dynamic coupling between ACC and vmPFC was seen when MTs dictate action costs are high, with the greatest change in coupling evident in offers where action requirement is most dependent on MT. One interpretation is that ACC suppresses the representation of low value offers in vmPFC when the future value of conserving a unit of budget is high and the optimal decision is to reject. Conversely, when MTs are low, decoupling between ACC and vmPFC may reflect a disinhibition of value signals relating to previously unfavourable offers. This contrasts with other suggestions that ACC signals a need for control but plays no causal role in conflict resolution (Kerns et al., 2004), or that dissociable decision variables are computed in vmPFC and ACC that compete for behavioural output (Boorman et al., 2013). Since ACC activity in the current task is not sensitive to changes in MTs corresponding to high value offers, it is unlikely to represent an unrelated correlate of trial time or WM content.

In contrast to the selectivity implemented by ACC, I found that MTs were tracked indiscriminately within an extensive fronto-subcortical-parietal network. Though planned choice has only recently been studied in a value domain, a finding that this network tracks computations related to future value is consistent with previous work from the model-based reinforcement learning literature (Daw et al., 2005; Glascher et al., 2010; Wunderlich, Dayan, et al., 2012). Interestingly, recent evidence suggests PFC neurons can adapt their tuning profiles to accommodate changes in behavioural context (Stokes et al., 2013), a mechanism that could underlie a network-level implementation of the adaptive responses observed in my task. I note this fronto-parietal network also encompasses regions implicated in executive control (Barber & Carter, 2005; Hare et al., 2009; Wallis & Miller, 2003), exploratory behaviour (Daw et al., 2006; Yoshida & Ishii, 2006), intertemporal choice (McClure et al., 2004) and WM (Barbey et al., 2012; Curtis & D'Esposito, 2003).

One limitation of the current task is that it cannot characterize a neural correlate of the fully integrated value derived from my computational model (the difference between the current offer and the associated MT) because this is correlated with the immediate value of the offer. However, the observed fronto-subcortical-parietal activity may reflect a value comparison between offer value and MT. As MTs decrease the difference in value between go and nogo shifts in favour of a go response, whilst when MTs increase they approach the average worth of the offer value range (2.5), making the decision to accept or reject harder. Alternatively, given that MTs trended downwards as trials progressed (although not exclusively, as they are also a function of the current budget), they are anti-correlated with WM demand, following the contents of trial history become harder to maintain (and update) through time.

Since I found activity in this fronto-subcortical-parietal network tracked MTs across all offers, this profile may reflect a WM signature. Interestingly, it has been shown that goal-directed choice is dependent on WM (Otto, Gershman, et al., 2013). In this regard, there is considerable debate as to whether delay-period DLPFC activity, classically interpreted as a correlate of WM, reflects the pure maintenance of information, or instead if WM is merely an emergent properly of executive and attentional functions implemented in DLPFC (Postle, 2006).

My paradigm also incorporated high (HC) and low action constraint (LC) environments, and in the former subjects reject lower value options to increase the probability of capitalizing from larger later rewards. I found categorically switching from LC to HC correlated with the fMRI signal in a similar fronto-parietal network. Within this network, the more DLPFC was recruited in HC compared to LC, the more a subject would modulate their behavioural response to value 2 offers between conditions. In addition, I found widespread correlates of offer value in regions previously linked to value computations, including vmPFC (Hare et al., 2009), VS (Guitart-Masip et al., 2012), posterior cingulate/precuneus (Litt et al., 2011) and amygdala (Jenison et al., 2011). Importantly, value representations were altered in HC in vmPFC, a key value-coding region.

Interestingly, a comparable fronto-parietal network is reliably up-regulated in conditions requiring cognitive control or overcoming response conflict in task switching paradigms (Badre, 2008; Kerns et al., 2004; Mansouri, Tanaka, & Buckley, 2009; Pochon, Riis, Sanfey, Nystrom, & Cohen, 2008). This likeness suggests participants may be engaging cognitive control mechanisms to appropriately reject appetitive, though relatively less valuable, offers in light of increasing environmental demands in HC trials. In this framework, my data corroborate previous ideas of interplay between PFC and value regions, suggestive of a scheme whereby value signals are modulated directly to achieve adaptive choice (Diekhof &

Gruber, 2010; Hare et al., 2009). However, as with previous control paradigms, I note that a categorical difference in activity profiles between conditions does not pose any properties that allow attribution of specific computational roles.

Table 4.1 Results for all second-level contrasts (whole-brain corrected at the cluster-level, FWE p =< 0.05).

Contrast	Name of Region	Cluster FWE p value	MNI Coordinates			Statistics	
			x	y	z	t value	Z score
HC > LC	Right Parietal	< 0.001	28	-64	48	5.70	5.39
	Right DLPFC	< 0.001	40	12	30	4.75	4.56
	Left Parietal	0.002	-28	-56	46	4.53	4.37
LC > HC	Right V1	< 0.001	8	-76	6	7.38	6.77
	Left V1		-8	-84	-8	6.08	5.72
	Left	< 0.001	-28	-28	-12	5.80	5.48
	Left Parietal	< 0.001	-50	-24	24	5.77	4.46
	Left Insula		-40	-6	-2	4.59	4.42
	vmPFC	< 0.001	-6	44	-10	5.06	4.84
	Left Precuneus	< 0.001	-10	-54	-12	4.56	4.39
	Right Parietal	< 0.001	44	-34	24	4.49	4.33
	Mid Cingulate	0.001	14	-20	46	4.08	3.96
Linear effect offer value	vmPFC	< 0.001	4	52	14	6.31	5.91
	Bilateral		-4	14	-8	5.70	5.39
	Left Mid Temporal	< 0.001	-52	-58	20	5.83	5.50
	Left Parietal		-54	-26	22	5.62	5.33
	Left Sup Temporal		-58	-18	10	4.46	4.30
	Left Mid Occipital		-42	-74	32	4.15	4.02
	Left	< 0.001	-28	-32	-14	5.67	5.37
	Right Lingual		14	-44	2	5.45	5.18
	Right Cuneus	< 0.001	16	-82	32	5.65	5.35
	Right M1	0.004	56	-8	44	5.38	5.12
	Left M1	0.001	-44	-16	58	4.91	4.71
	Right Hippocampus	0.008	24	-18	-16	4.89	4.69
Negative offer value	Right Insula	< 0.001	30	22	-10	5.44	5.17
	ACC	< 0.001	6	24	48	5.26	5.02
	Left Insula	0.009	-34	18	-4	5.04	4.82
	Right Parietal	0.001	36	-50	50	4.47	4.31
Negative model thresholds	Left Caudate	< 0.001	-12	-6	18	8.23	7.42
	Right Sup Parietal		24	-56	52	7.95	7.21
	Right IFGpt		32	18	28	7.92	7.19
	Right Thalamus		12	-10	18	7.77	7.08
	Left Mid Occipital		-38	-72	10	7.49	6.86
	Right Lingual		20	-74	4	7.16	6.60
	Right DLPFC		32	8	26	6.99	6.46
	Right Frontal Mid		34	52	-6	4.87	4.67
	Left Putamen	< 0.001	-18	17	-6	4.25	4.11
	Left M1	< 0.001	-42	-2	52	6.40	5.99
	Left DLPFC		-26	6	64	4.94	4.73
	Right IDGpo	< 0.001	34	26	-10	6.06	5.70
	Right Caudate		10	20	-8	4.27	4.13
Linear effect model thresholds	ACC	< 0.001	-6	28	22	5.29	5.04
	Left DLPFC (BA46)	< 0.001	-32	46	18	4.69	4.51
	Dorsal vmPFC	0.037	-8	56	2	4.50	4.34

CHAPTER 5

ARBITRATION BETWEEN CONTROLLED AND IMPULSIVE CHOICE

The impulse to act for immediate reward often conflicts with more deliberate evaluations that support long-term benefit. The neural architecture that negotiates this conflict remains unclear. One account proposes a single neural circuit that evaluates both immediate and delayed outcomes, while another outlines separate impulsive and patient systems that compete for behavioural control. Here I designed a task in which a complex pay-out structure divorces the immediate value of acting from the overall long-term value, within the same outcome modality. Using model-based fMRI in humans, I demonstrate separate neural representations of immediate and long-term value, with the former tracked in anterior caudate (AC) and the latter in ventromedial prefrontal cortex (vmPFC). Crucially, when subjects' choices were compatible with long-run consequences, value signals in AC were down-weighted and those in vmPFC were enhanced, while the opposite occurred when choice was impulsive. Thus, my data implicate a trade-off in value representation between AC and vmPFC as underlying controlled versus impulsive choice.

5.1 Introduction

Everyday occurrences often involve negotiating immediate temptations whose consumption might jeopardize long-term goals. A common instance is where the prospect of a large immediate reward is coupled with a harmful yet delayed consequence, such as enjoying a cigarette that can imperil long-term health. Behavioural findings suggest that in this context the desire for an hedonic payoff competes with the intent to act with foresight (Baumeister, Bratslavsky, Muraven, & Tice, 1998; Hare et al., 2009; Hofmann, Friese, & Strack, 2009), demanding self-control.

A longstanding notion in psychology is that resisting temptation involves a competition between two competing systems (Hofmann et al., 2009; Hofmann & Van Dillen, 2012). In support of this idea, several experiments have found evidence for a trade-off between separate neural systems that preferentially activate when choice is driven by immediate and delayed rewards respectively (McClure et al., 2004; Tanaka et al., 2004). These systems are thought to guide choice by encoding value on opposing time-scales, though it is unclear whether their selective involvement reflects the tracking of other decision components.

An alternative perspective, particularly within neuroeconomics, suggests choice is driven by a single system that represents both immediate and delayed decision outcomes. In dietary choice paradigms, where individuals choose between foods that vary along a scale of healthiness and tastiness (Hare et al., 2009; Hare et al., 2011), neuroimaging supports a role for the ventromedial prefrontal cortex (vmPFC) in integrating both components of value (Hare et al., 2009; Rangel, 2013). This is reinforced by other evidence that a common vmPFC-striatal circuit tracks the subjective value of choice options (Kable & Glimcher, 2007). The divergence between these two perspectives remains largely unresolved.

Here, I designed a novel paradigm that required subjects to accept or reject offers with known immediate value, presented sequentially within a trial. The probability of receiving large or small offers depended on past actions, such that an early acceptance of a large immediately available offer harmed long-term earnings by diminishing the opportunity for future rewards. Thus, maximizing long-run earnings sometimes required rejecting seemingly attractive offers associated with a high immediate payoff. In contrast to previous paradigms, long-run consequences were fully defined within a single outcome modality based on knowledge of the formal structure of the task. In this way I was able to decorrelate immediate from long-term value across offers, where the latter includes the delayed consequences of acting. I used model-based functional magnetic resonance imaging (fMRI) to investigate the neural representation of each value component and linked this to a disposition for controlled versus impulsive action.

5.2 Methods

Subjects

23 adults participated in the experiment (9 male and 14 female; age range 18-26; mean 21.2, SD = 2.33 years). All were healthy, reporting no history of neurological, psychiatric or other current medical problems. Subjects provided written informed consent to partake in the study, which was approved by the local ethics board (University College London, UK).

Training paradigm

In a conditioning phase, performed outside of the scanner, subjects learnt stimulus-reward associations between a set of three differently coloured rectangular cues and their respective reward values. Each coloured rectangle corresponded to one of three possible outcomes involving receipt of 3, 5, or 7 tokens, randomized across individuals. Subjects were instructed that each token would translate into a fixed sum of money at the end of the experiment. Each trial began with a central fixation cross presented for 1000 ms, followed by presentation of a random pair of coloured cues, one appearing to the left one to the right of the screen. Subjects had a 2000 ms time-window to choose between these two boxes via a left or right button press, followed by presentation of the outcome of their choice for 1000 ms. The outcome was a written message indicating the total number of tokens won. Subjects were instructed to explore all options until they were confident they had learnt all three associations, after which they should choose the box from the pair with the higher value. Each trial was defined as either correct if the subject chose the more valuable of the two options, and incorrect if the less valuable option was chosen. To ensure adequate learning, performance was calculated over six bins of twenty trials, with all subjects reaching a performance criterion of >= 90% by trial 60 onwards. Subjects were asked to verbally communicate the nature of the learnt associations.

Task paradigm

On every trial subjects were presented with a random sequence of trained stimuli (see *training paradigm*), appearing individually and sequentially, with a variable inter-stimulus interval (750 - 1250 ms). Each stimulus, presented for 1500 ms, constituted an offer requiring either a go response to win the relevant number of tokens or a nogo response which lead the player to forego monetary gain. However, a restriction was placed on the number of offers that could be exploited for reward on any given trial. Specifically, subjects were instructed that they could receive between 7-9 offers out of which between 4-6 could be accepted. The precise offer number and acceptance budget were drawn randomly and independently on every trial under a uniform distribution, and thus every combination was equally likely. A green circle on the top central portion of the screen turned red to indicate that a player had exhausted their go budget, after which nogo responses were enforced for any remaining offers.

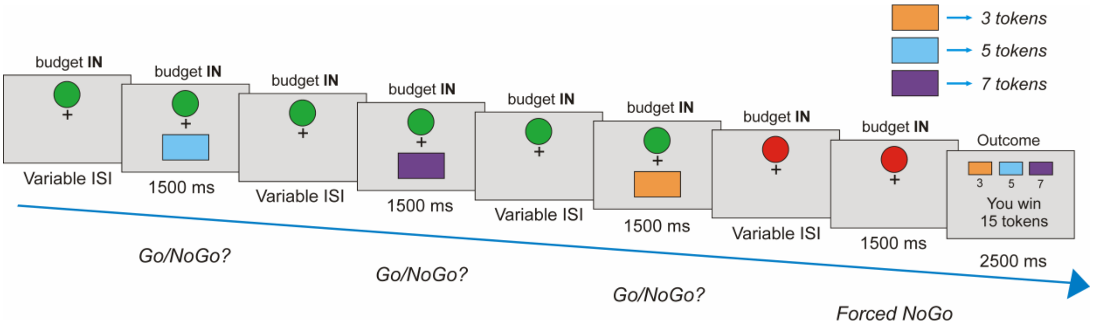


Figure 5.1 In pre-scanning training (not shown), subjects learnt to associate three distinct colour stimuli with a token value of 3, 5 or 7, with each token won translated into a cash prize at the end of the experiment. In the actual experiment proper (shown above), a player was presented with a sequence of stimuli, each constituting an individual offer. These offers required a go response to win or a nogo response to forego a gain. Crucially, a restriction was placed on the number of offers that could be exploited per trial sequence, such that on every trial a player could receive an overall amount of 7-9 offers but where only 4-6 (go budget) could be accepted, with every combination being equally likely. A green circle at the top central portion of the screen turned red to indicate a player had exhausted their go budget, after which they passively observed the remaining sequence of outstanding offers. At trial onset, each offer had an equal probability of being the colour associated with 3, 5 or 7 tokens {0.33 0.33 0.33, respectively}. With the exception of the first offer, if a player accepted a value 7 offer before rejecting at least three previous offers, the distribution would shift in favour of value 3 offers for the remainder of the sequence {0.9 0.05 0.05}. Likewise, if a player accepted a value 5 offer before rejecting at least three previous offers, the distribution would modestly shift in favour of value 3 offers {0.5 0.25 0.25}. The current distribution was updated based on the most recent action. Thus, an optimal player had to track the immediate reward environment as well as calculate overall (long-term) value by taking account of how an immediate go response might impact on future reward abundance, entailing often rejecting an offer associated with a large immediate reward.

Importantly, the value of each offer was probabilistic and governed by a set of explicitly instructed contingencies. At trial onset, each offer had an independent and equal probability of being worth 3, 5 or 7 tokens {0.33 0.33 0.33 (for 3, 5 and 7 respectively)}. Excluding the first offer, if a player accepted a value 7 offer before rejecting three or more previous offers the distribution would shift such that every future offer would have a probability distribution greatly in favour of value 3 {0.9 0.05 0.05}. Similarly, excluding the first offer, if a player accepted a value 5 offer before rejecting three or more previous offers the distribution would shift such that every future offer would have a probability distribution moderately in favour of value 3 {0.5 0.25 0.25}. The probability distribution was updated according to the choice made on the most recent offer. Thus a player had to consider both the immediate and long-term consequences of a go response in order to maximize payoff across a trial. Following the last offer, an outcome displaying the total number of offers won was presented on the screen for 2500 ms.

All subjects received 1 block (36 trials) of training outside the scanner in order to familiarize themselves with the task attributes and to diminish learning in the scanner. Subsequently, 108 trials were completed in the scanner across three sessions of 36 trials. The number of tokens won across sessions was summed and converted to a cash prize.

Due to the complex nature of the task, subjects were probed to ensure they had currently understood the nature of the contingencies that linked actions to switches in the distribution of offers, prior to scanning. Specifically I constructed a written set of hypothetical trials, where for each trial subjects were ask to indicate their belief in the current offer distribution given a history of specific offers and actions. For example, “What is the probability of the next offer being worth 5 tokens given that a value 7 offer was accepted at the third index and no offers had previously been rejected?”. One subject failed to demonstrate correct knowledge of the task and was excluded from participating in the scanning portion of the experiment. This participant thus is not reflected in the remaining 23 participating subjects.

Behavioural data analysis

Within-trial modulation of choice

Within a trial, a player transitions through a number of discrete states dependent on three fluctuating variables, the number of offers already seen, the number of accepts already expended and the current offer distribution. To assess how the probability of accepting a given offer fluctuated as a function of these variables, I split trials by offer index (i.e. 1-9), the number of offers already rejected (i.e. 0-8), and the current offer distribution, calculating the probability of accepting at every possible permutation (Figure 5.3A, p. 124). Note that here I only display behaviour corresponding to offers where the probability distribution is equal given that choice under this contingency is most relevant to the questions of interest. The probability of accepting at every state was averaged across all participants. For display purposes, I discarded cells with less than a total of 15 data points.

Robust logistic regression

In order to confirm my hypothesis that both immediate and long-term value show independent effects on choice, I used a robust logistic regression to model the dependence of a go/nogo response (across all choice data) on immediate and long-term value in a model in which both regressors competed for variance. The algorithm implemented used iteratively reweighted least squares with a logistic weighting function. I performed one-sample t-tests on the resulting beta coefficients across subjects. A positive beta implies subjects are more likely to go when value is high.

Computational modelling

A major interest here is the extent to which subjects' utilize estimates of immediate and long-term value to guide choice. I used computational modelling to evaluate evidence that choice was guided purely by immediate value, purely by (the optimal) long-term value, or by a corresponding trade-off. Each model calculated the value of accepting an offer which was passed through a sigmoid function (σ) to determine action probabilities as follows:

$$P_A = \frac{1}{1 + \exp(-\tau \cdot V_A)}$$

where V_A is the expected value of accepting an offer, and τ is a temperature parameter that governs the stochasticity of choices.

Immediate reward model

I conjectured subjects might choose on the basis of immediate value, disregarding the downstream consequences associated with prematurely accepting high (face) value offers, whereby

$$V_A = IR - c_1$$

where *IR* is the face value of the current offer and c_1 represents a value intercept.

IR, c_1 and τ (the temperature parameter of the associated sigmoid function) were fit by maximum likelihood estimation (see *Model fitting & comparison*, p. 116).

Optimal model

I built a model that calculated the optimal decision at each offer, where optimal is defined as maximizing expectation of total reward delivery in the trial. The model assumes correct knowledge of the structure of the task. The current state of the task was defined by three belief distributions: **O**, over *o*, the number of offers remaining, **A**, over *a*, the number of accepts remaining, and **M**, the probability distribution governing the value of the forthcoming offer. The expected value of being in a state was:

$$SV(\boldsymbol{O},\boldsymbol{A}) = \sum_{r=\{3,5,7\}} \mathbf{M}_m(r) \cdot \max\{P(o > 1) \cdot SV(\boldsymbol{O}',\boldsymbol{A}), r + P(o > 1) \cdot P(a > 1) \cdot SV(\boldsymbol{O}',\boldsymbol{A}')\}$$

where ***O'*** is defined by

$$P(\boldsymbol{O}' = o) = \frac{P(O = o + 1)}{\sum_o P(O = o + 1)}$$

and **A'** is defined analogously. Thus going from ***O*** to ***O'*** or ***A*** to ***A'*** updates the probability distribution such that it remains uniform but shifts to the left. Note that calculating the recursive *SV* function was effectively a search through a tree of all possible moves. The recursion ends when *P*(*o* > 1) or *P*(*a* > 1) are 0, and *SV* is not evaluated.

M is defined by three discrete probability distributions as follows:

$$M_0 \ = \ \{0.33\ 0.33\ 0.33\}$$

$$M_1 \ = \ \{0.50\ 0.25\ 0.25\}$$

$$M_2 \ = \{0.90\ 0.05\ 0.05\}$$

At trial onsets, *m* = 0, and is updated according to the following rules:

If we are on the first offer, or 3 offers have previously been rejected, *m* doesn’t change.

Otherwise, if a value 5 offer is accepted, *m* = 1, and if a value 7 offer is accepted, *m* = 2

At each offer the model calculated the value of rejecting,

$$V_R = P(o > 1) \cdot SV(\boldsymbol{O}',\boldsymbol{A})$$

and the future value of accepting,

$$V_{AF} = P(o > 1) \cdot P(a > 1) \cdot SV(\boldsymbol{O'}, \boldsymbol{A'})$$

The expected value difference between accepting and rejecting, *EV*, was calculated as,

$$EV = V_{AF} + IR - V_R$$

where *IR* represents the (face) value of the current offer.

EV was passed through a sigmoid function to determine P_A, the probability of a go response (see above).

Tradeoff model

Given evidence that both immediate and long-term value had dissociable influences on choice, I hypothesized choice might involve a trade-off between two value systems. Accordingly, I specified a model whereby immediate and long-term value both contributed independently to the calculation of expected value (TV, tradeoff value), whereby the associated trade-off was captured by a single parameter that governed the weight placed on either value as follows:

$$TV = (EV \cdot c_1) + (IR - c_2) \cdot (1 - c_1)$$

where *EV* is the expected, or long-term value, derived from the optimal model (see above), *IR* is the (face) value of the current offer, c_1 governs the nature of the trade-off, and c_2 represents a value intercept.

In addition, it seemed reasonable to assume that subjects might trade-off immediate and long-term value differently depending on the face value of the current offer. I therefore specified a second trade-off model in which a separate trade-off parameter governed the weight placed on immediate and long-term value for each face value (3, 5 and 7).

Model fitting & comparison

As described in previous reports (Guitart-Masip et al., 2012; Huys et al., 2011) I used a hierarchical Type II Bayesian (or random-effects) procedure using maximum likelihood to fit simple parameterized distributions for higher level statistics of the parameters (see also *Hierarchical Bayesian procedures*, Chapter 3, p. 73). Since the values of parameters for each subject are 'hidden', this employs the Expectation-Maximization (EM) procedure. Thus on each iteration the posterior distribution over the group for each parameter is used to specify the prior over the individual parameter fits on the next iteration. For each parameter I used a single distribution for all participants. Before inference, all parameters were suitably transformed to enforce constraints (log and inverse sigmoid transforms).

Models were then compared using the integrated Bayesian Information Criterion (iBIC), where small iBIC values indicate a model that fits the data better after penalizing for the number of parameters. Comparing iBIC values is akin to a likelihood ratio test (Kass & Raftery, 1995).

fMRI data acquisition

fMRI was performed on a 3-Tesla Siemens Quattro magnetic resonance scanner (Siemens, Erlangen, Germany) with echo planar imaging (EPI) and 32-channel head coil. Functional data was acquired over three sessions containing 280 volumes with 48 slices (664 volumes total). Acquisition parameters were as follows: matrix = 64 x 74; oblique axial slices angled at -30° in the antero-posterior axis; spatial resolution: 3 x 3 x 3 mm; TR = 3360 ms; TE = 30 ms. The first five volumes were subsequently discarded to allow for steady state magnetization. Field maps were acquired prior to the functional runs (matrix = 64 x 64; 64 slices; spatial resolution = 3 x 3 x 3 mm; gap = 1 mm; short TE = 10 ms; long TE = 12.46 ms; TR = 1020 ms) to correct for geometric distortions. In addition, for each participant an anatomical T1-weighted image (spatial resolution: 1 x 1 x 1 mm) was acquired for co-registration of the EPIs.

During scanning peripheral measurements of subject pulse and breathing were made together with scanner slice synchronization pulses using the Spike2 data acquisition system (Cambridge Electronic Design Limited, Cambridge UK). The cardiac pulse signal was measured using an MRI compatible pulse oximeter (Model 8600 F0, Nonin Medical, Inc. Plymouth, MN) attached to the subject's finger. The respiratory signal (thoracic movement) was monitored using a pneumatic belt positioned around the abdomen close to the diaphragm.

fMRI data analysis

Data were pre-processed and analysed using SPM8 (Wellcome Trust Centre for Neuroimaging, UCL, London). Functional data were bias corrected for 32-channel head coil intensity inhomogeneities, realigned to the first volume, unwarpped using individual fieldmaps, co-registered to T1w images, spatially normalized to the Montreal Neurology Institute (MNI) space (using the segmentation algorithm on the T1w image with a final spatial resolution of 1 x 1 x 1 mm) and smoothed with an 8mm FWHM Gaussian kernel. The fMRI time series data were high-pass filtered (cutoff = 128 s) and whitened using an AR(1)-model.

For each subject I computed a statistical model by applying a canonical hemodynamic response function (HRF) combined with time and dispersion derivatives. Using an in-house Matlab toolbox (Hutton et al., 2011) I constructed a physiological noise model to account for artefacts that take account of cardiac and respiratory phase as well as changes in respiratory volume. This resulted in a total of 14 regressors which were sampled at a reference slice in each image volume to give a set of values for each time point. The resulting regressors were included as confounds in my GLM at the first level (see below).

GLM 1

In order to investigate regions tracking immediate or long-term value, I designed a GLM that allowed me to explore the BOLD response to a subset of offers for which immediate and long-term value were most decorrelated, corresponding to offers between index 2 and 3 within a trial (see Figure 5.3A, middle panel, yellow boxes, p. 124). I split these offers contingent on their face value, such that each value (3, 5 and 7) was modelled as a separate regressor. Although these offers were selected from neighbouring states (meaning that for any given (face) value, the long-term value of a go response was similar), each onset regressor was parametrically modulated by long-term value (from the optimal model) so as to account for variance associated with a difference in current state. Additional regressors included the onsets of all within-budget offers outside of the yellow box in Figure 5.3A (p. 124), parametrically modulated by both immediate and long-term value, all out-of-budget offers (for which nogo responses were enforced), parametrically modulated by immediate value, the onset of go responses (button presses) across the entire experiment, so as to explain away motor-related activity, and the onset of trial outcomes (parametrically modulated by tokens won). Regressors of no interest included 6 movement-related covariates (the 3 rigid-body translations and 3 rotations resulting from realignment) and 14 physiological regressors (6 respiratory, 6 cardiac and 2 change in respiratory/heart rate). All regressors were modelled as stick functions with duration of zero and convolved with a canonical form of the hemodynamic response function (HRF) combined with time and dispersion derivatives.

To explore the BOLD response to the onset of value 3, 5 and 7 offers when immediate and long-term value were decorrelated, I conducted a random-effects one-way ANOVA at the second level, with a single factor (face value) and 3 levels (3, 5, 7), containing individual subject first-level contrast images corresponding to the first three onset regressors from my GLM. I constructed functional ROIs (fROIs) from clusters that survived small volume correction for a prior volume of interest (see *Anatomical volume of interest*, p. 120) using the MarsBar toolbox (v. 0.42) for SPM. I extracted mean parameter estimates from each fROI for the three onset regressors of interest and performed post-hoc paired t-tests to explore differences in BOLD response between offer values. For display purposes, onset parameter estimates were normalized (mean centred). In addition, I specified the contrast {0 -1 1} corresponding to the onset of value 3, 5 and 7 offers to explore regions that covaried with the demand for control. The latter was performed as a whole-brain analysis.

In order to test whether the BOLD response to value 7 offers in my four ROIs was related to choice, I correlated mean parameter estimates (corresponding to the value 7 onset regressor) extracted from each ROI, with the trade-off parameter captured by our model fitting procedure. This resulted in four independent correlations.

GLM 2

In order to quantify the extent to which the BOLD response was modulated by immediate and long-term value, I built a second GLM where I concatenated the first three regressors from *GLM 1* (onsets of 3, 5 and 7-token offers) into a single regressor, and added a parametric modulator for immediate value, which was forced to compete for variance (and was thus not orthogonalized) with an overall (long-term) value modulator. All other regressors remained identical to those specified in *GLM 1*.

I extracted mean parameter estimates from each ROI for the immediate and long-term value parametric modulators. I conducted Grubb's test to probe for extreme values so as to remove subjects who were significant outliers at a threshold of $p < 0.05$, and performed one sample t-tests at the second level on the resultant betas across subjects.

GLM 3

In order to look for difference in value coding during correct and incorrect responses, I split the regressor corresponding to the onset of value 7 offers from *GLM 1* into correct (nogo) responses and incorrect (go) responses. All other regressors remained equivalent.

I conducted a random-effects one-way ANOVA at the second level, with a single factor (accuracy) and 2 levels (correct, incorrect), containing individual subject first-level contrast images corresponding to the go 7 and nogo 7 onset regressors. I extracted parameter estimates from each ROI and performed post-hoc paired t-tests to explore a main effect of response accuracy. I noted that only 15 out of 23 subjects had enough variance in their ability to respond accurately across trials, and thus the above analysis was restricted to these individuals.

Anatomical volume of interest

I also constructed an anatomical volume of interest (VOI) that included individual valuation regions of a prior interest for the purposes of small volume correction, effectively reducing the number of voxel-wise comparisons. This consisted of the entire vmPFC, caudate nucleus, putamen and ventral striatum (nucleus accumbens) (see Figure 5.4, p. 127). The vmPFC and dorsal striatum were defined as anatomical ROIs from the MarsBar toolbox (v. 0.42) for SPM. For the ventral striatum I used a group-average ROI derived from a diffusion tensor imaging connectivity-based parcellation of the right nucleus accumbens in humans, taken from (Baliki et al., 2013). This ROI consisted of both the core and shell subcomponents of nucleus accumbens. The right region was flipped along the x-dimension in the MarsBar toolbox to obtain bilateral accumbens.

Functional regions of interest

I defined functional regions of interest (fROIs) from clusters that survived small volume correction for a pre-defined VOI (see above) when testing for regions tracking IR or EV using GLM 1 (see *GLM 1*, p. 118). For the anterior caudate, I excluded voxels that fell outside of an anatomical ROI for bilateral caudate from the MarsBar toolbox. These fROIs were used for all remaining fMRI analyses. All ROI analyses were performed using the MarsBar toolbox (v. 0.42) for SPM.

5.3 Results

On every trial, subjects received between 7-9 offers, but an imposition of a limited "budget" meant they could only accept between 4-6 offers. Importantly, I penalized acceptance of the largest (7-token) and second largest (5-token) offers early in a trial by impoverishing remaining offers in that trial, where the penalty scaled with the face value of the current offer (see Figure 5.1 p. 110; see *Task paradigm* p. 109). Here, immediate value equates to the face value of each offer (3, 5 or 7 tokens), whereas long-term value represents the total expected utility from accepting. Thus, long-term value includes the face value, the cost of expending a unit of budget, and the cost of changing the future probability of reward. In some cases, total earnings could be maximized by rejecting 7-token but not 5-token offers. This is because the penalty associated with an accept response can be greater than the immediate payoff for 7-token, but not 5-token offers. In other words, the long-term value of a 7-token offer can sometimes be negative while nonetheless yielding the highest immediate payoff. Hence, immediate value was decorrelated from long-term value across offers, despite the former being a component of the latter.

Given the complexity behind the rules governing how actions shaped future offers, subjects were probed prior to scanning to ensure they correctly understood the contingencies of the task (see Methods, p. 111). In brief, each subject was shown a series of hypothetical trials where they had to predict the probability of a forthcoming offer being a specific value, given a preceding sequence of offers and actions. All subjects demonstrated correct understanding of the task and were fully aware of the contingencies linking actions to states following careful instruction. In addition, in order to minimize effects of learning and uncertainty during scanning, subjects played one block (36 trials) of the task prior to performing the experiment in the scanner.

Although self-control is multi-faceted, one important aspect is the ability to override one's impulses or prepotent responses (Gailliot & Baumeister, 2007). In my task, the requirement for this form of self-control is greatest near the start of a trial, where accepting a large immediate offer has detrimental future consequences (see Figure 5.3A, middle panel, yellow boxes, p. 124). Interestingly, subjects were faster to accept 7-token offers compared to 5-token ($p < 0.001$) or 3-token ($p < 0.001$) offers across a trial, suggesting of a prepotent tendency to reap large immediate rewards (see Figure 5.2, p. 123). In this part of a trial, I found that subjects under-chose 3-token offers and over-chose 5 and 7-token offers, as compared to an optimal model (Figure 5.3A, p. 124).

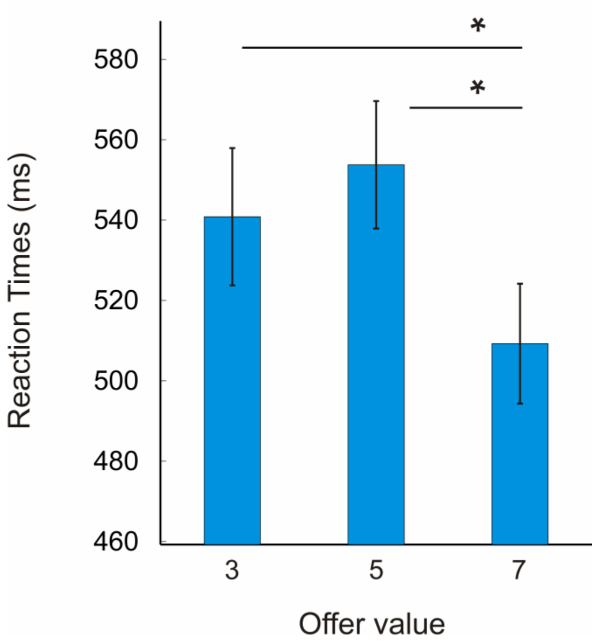


Figure 5.2 Considering all ‘go’ responses in a trial, subjects were faster to accept token-value 7 offers compared to value 3 or 5 offers, suggestive of a prepotent attraction to 7-token cues. Vertical lines represent SEM. * indicates $p < 0.05$ (paired t-tests).

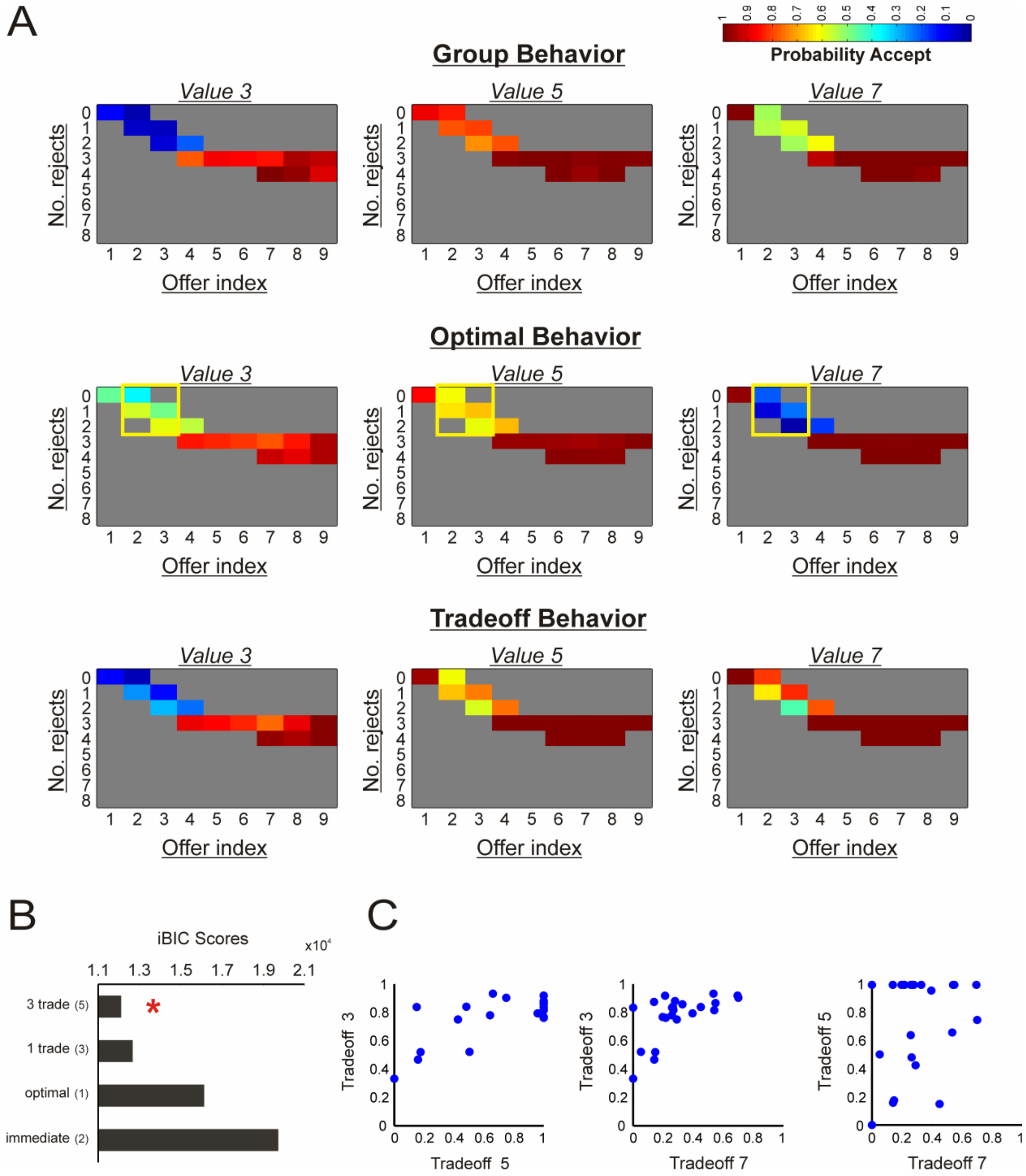


Figure 5.3 (**A**) Plotted above are subjects' mean probability of offer acceptance as a function of the number of offers already seen (ranging from 1-9) and number of offers already rejected (ranging from 0-8) in a trial, split by offer value (3, 5, 7) (top panel). The spectrum runs from blue (p=0) to red (p=1). Compared to an optimal model in which choice is dictated by correctly inferring long-term value (middle panel), subjects under-accept value 3 offers and over-accept value 7 offers at the start of trials (top panel; based on group mean data, n=23). This discrepancy is rectified by a model in which immediate and long-term value trade-off for behavioural control (lower panel). Note that the lower panel illustrates choice predicted by the trade-off model based on mean group parameter fits (n=23). Yellow boxes in the middle panel demonstrate offers for which immediate and long-term value are

maximally decoupled, and those for which all fMRI analyses are centred on. (**B**) Model comparison showed that a model in which each offer value (3, 5, 7) is assigned a separate parameter that governs how much weight is placed on immediate versus long-term value in the associated trade-off fits behaviour better than alternatives, indicated by its lowest iBIC score (3 trade). These alternatives included a model in which a single parameter governs the trade-off (1 trade), a model dependent on optimally inferring long-term value (optimal), and a model driven purely by immediate value (immediate). The number of free parameters is indicated in brackets for each model. (**C**) Pair-wise scatter plots show individually fit trade-off parameters (c_1, see Methods, p. 115) from the winning model for 3 versus 5-token offers, 3 versus 7-token offers, and 5 versus 7-token offers. A trade-off value closer to 0 indicates behaviour is predominantly driven by immediate value, while a value closer to 1 indicates behaviour is predominantly driven by long-term value. Each circle represents one participant.

This pattern of choice is consistent with subjects being mindful of the future consequences of their actions, but nevertheless being over-susceptible to an influence of a current offer's face value. I therefore predicted that both immediate and long-term value (see Methods, p. 113, for an explanation of how these are calculated) would independently influence behaviour. Using a logistic regression I indeed found that immediate (mean b = 0.047; p = 0.001) and long-term value (mean b = 0.113; P < 0.0001) were significant predictors of choice, implying behaviour was neither exclusively optimal nor impulsive, but incorporated features of both traits.

Given evidence that immediate and long-term value exert a differential impact on action selection, I conjectured that a model encompassing a trade-off between each valuation would capture choice behaviour. I used Bayesian model comparison to evaluate whether group behaviour was driven exclusively by immediate value, by long-term value, or by a trade-off between the two (see Methods, p. 113 - 116). While subjects varied in their ability to prioritize long-term value in the face of high-token offers, a model in which each offer value (3, 5, 7) was assigned an independent trade-off parameter captured group-level choice best (Figure 5.3B, p. 124). Hence, while subjects were considerate of future consequence, immediate rewards were (on average) overweighted across the experiment. The finding that subjects weight immediate and long-term value differently depending on face value is intuitive, as long-term value deviates from immediate value to a greater degree for some offers compared to others, and is thus sometimes harder to track. Indeed, the best-fitting trade-off parameters, which provide a measure of how strongly each player weighted immediate relative to long-term value for the three offers, strongly endorse this account (Figure 5.3C, p. 124).

Since immediate and long-term value exert distinct influences on choice I conjectured these quantities would have dissociable representations in value sensitive brain regions. To test this, I used fMRI and implemented a GLM (see Methods, *GLM 1*, p. 118) in which each offer value (3, 5, 7) was modelled separately, but focusing on a subset of offers where immediate and long-term value were maximally dissociable within any given trial (see Figure 5.3A, middle panel, yellow boxes, p. 124). In this set of offers, optimal behaviour mandated strongly rejecting 7-token offers, strongly accepting 5-token offers, and weakly accepting 3-token offers. Thus, regions representing long-term (overall) value should display a BOLD signal profile that is attenuated for 7-token offers, boosted for 5-token offers and modestly boosted for 3-token offers. In contrast, regions that track immediate rewards should show a BOLD signal profile that increases linearly as a function of face value. Importantly, I modelled go responses as an independent regressor in all GLMs, and this spanned button presses across the entire experiment, including those corresponding to offers outside of the yellow box in Figure 5.3A. Thus, any variance in activity attributed to cue onsets is independent from the generation of a motor response per se.

Given an a priori interest in responses within valuation regions, I generated a volume of interest (VOI; see Figure 5.4, p. 127) that included the ventromedial prefrontal cortex (vmPFC) (Balleine & Dickinson, 1998; Hare et al., 2009; Wunderlich, Dayan, et al., 2012), ventral striatum (Baliki et al., 2013; Guitart-Masip et al., 2012), caudate nucleus (Tricomi et al., 2004) and putamen (Brovelli et al., 2011) to constrain the search space and reduce the number of statistical comparisons. I used anatomical ROIs from the MarsBar toolbox (v. 0.42) for SPM and from previous research (see Methods, *Anatomical volume of interest*, p. 120).

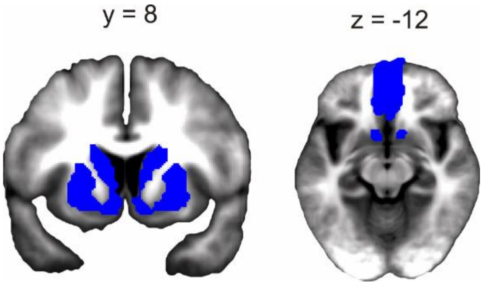


Figure 5.4 Volume of interest consisting of valuation regions of a priori interest used for small volume correction. Regions include vmPFC, bilateral caudate, bilateral putamen and bilateral ventral striatum (accumbens).

When testing for regions that track long-term value (a contrast of {0 1 -1} for 3, 5 and 7-token offers) I identified two clusters that survived small volume correction (SVC) for the VOI in vmPFC, including a ventral (Figure 5.5A, p. 128) and more lateral portion (Figure 5.5C, p. 128). Although it is difficult to distinguish between small cortical subdivisions along the medial prefrontal cortex in imaging studies (Haber & Knutson, 2010), it is possible that the more lateral portion of vmPFC is in the orbitofrontal cortex (OFC). In fact, the peak voxel in both vmPFC clusters (Figure 5.5 A and C) falls within Brodmann area 11. Thus, I use the term vmPFC in a broad manner to include the medial OFC. I also note that the vmPFC itself does not have a universally agreed upon demarcation in humans.

When testing for regions that track immediate value (a contrast of {-1 0 1} for 3, 5 and 7-token offers) I identified activation in both left (Figure 5.5B, p. 128) and right (Figure 5.5D, p. 128) anterior caudate nucleus that likewise survived SVC for the VOI. These clusters were then used to define functional regions of interest (fROIs) in vmPFC and anterior caudate for further analysis, which correspond to the regions displayed in Figure 5.5 (below).

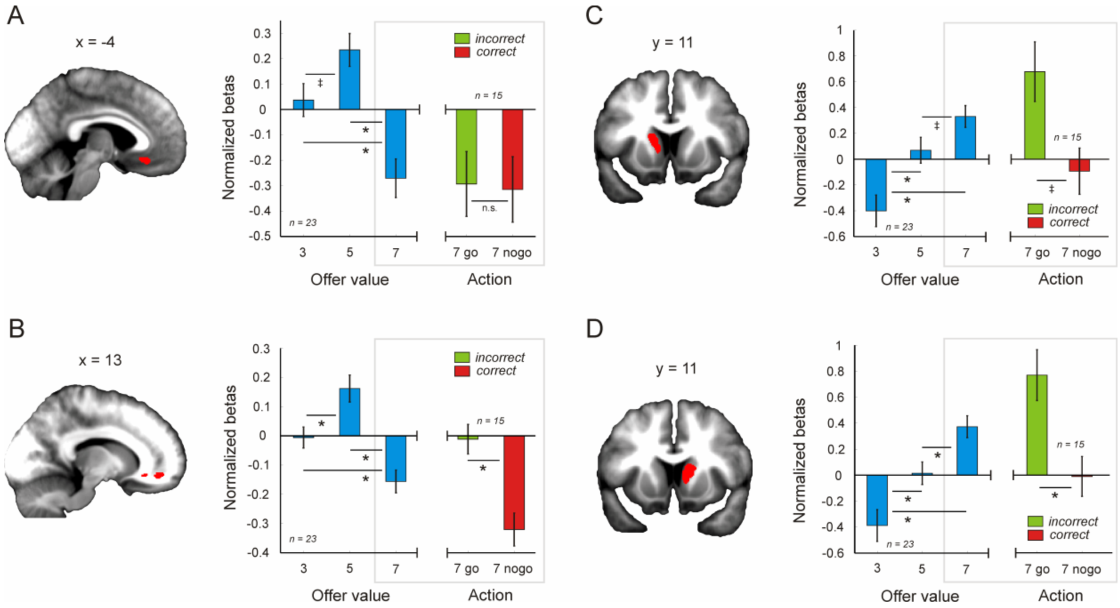


Figure 5.5 (**A**) Ventral vmPFC showed greater activation in response to 5-token compared to 3-token offers, but a deactivation in response to 7-token relative to 3 and 5-token offers, consistent with 7-token offers having a negative overall (long-term) value (see yellow boxes, Figure 5.3, panel **A**, p. 124). (**B**) BOLD in lateral vmPFC / OFC also reflected a representation of long-term (optimal) value. In trials where 7-token offers were impulsively accepted (7 go) compared to rejected (7 nogo), the representation of long-term value (for 7-token offers) was attenuated (less negative beta). (**C**) By contrast, anterior caudate exhibited a linearly increasing response profile to the presentation of 3, 5 and 7-token offers, consistent with this region showing preferential sensitivity to immediate value. Panel C shows the response in left anterior caudate. In trials where 7-token offers were impulsively accepted (7 go) compared to rejected (7 nogo), the representation of immediate value (for 7-token offers) was boosted in this region (more positive beta). (**D**) Right anterior caudate also tracks immediate value in this task, with BOLD response for 7-token offers being higher when these offers were impulsively accepted compared to when they were rejected. Vertical lines represent SEM. * indicates p =< 0.05; ‡ indicates p = 0.07; n.s. indicates not significant (paired t-tests).

To quantify the extent to which each fROI was preferentially driven by immediate versus long-term value, I constructed a second GLM that allowed me to regress both values against the BOLD signal within the same model, by collapsing offers into a single regressor and using immediate and long-term value as parametric modulators. Note these regressors, for which the average correlation was $r^2 = 0.24$, were not orthogonalized in my GLM and were forced to compete for variance (see Methods, *GLM 2*, p. 119). This analysis again showed that BOLD response in both vmPFC fROIs was driven by long-term value ($p = 0.004$, $p < 0.001$) and was not explained by immediate value ($p = 0.371$, $p = 0.795$). By contrast, BOLD activity in anterior caudate was driven predominantly by immediate value ($p < 0.001$, $p < 0.001$), though long-term value also contributed to signal variance ($p = 0.038$; $p = 0.024$) suggesting it represented mixed value components.

It has been proposed that self-control involves a conflict between competing value systems (Hofmann et al., 2009; McClure et al., 2004; Tanaka et al., 2004), and this idea gains support from evidence that the brain draws on multiple systems when making decisions (Balleine, 2005; Daw et al., 2005; Dolan & Dayan, 2013). However, an alternative suggestion is that choice is governed by a common value system embedded in vmPFC (Hare et al., 2009) or a vmPFC-striatal network (Kable & Glimcher, 2007). A finding here that distinct representations of immediate and long-term value are tracked in the brain fits better with the idea of two competing value systems. However, I note that long-term value in my task includes both immediate and delayed components of value. Thus, my data is consistent with the notion that both value components are integrated within vmPFC (Economides, Guitart-Masip, Kurth-Nelson, & Dolan, 2014; Hare et al., 2009). Importantly, if the separate encoding of immediate and long-term value is linked to the observed trade-off between these values during choice, I would expect between-subject variability in self-control to correlate with the strength with which long-term value was represented relative to immediate value. Specifically, a stronger representation of long-term relative to immediate value should track greater self-control. Indeed one might also expect that representations of immediate and long-term value would be altered in trials where subjects (incorrectly) accepted a 7-token offer compared to when subjects (correctly) resisted the temptation.

To test the first prediction, I correlated parameter estimates for the onset of 7-token offers with the trade-off parameter which captures the weighting placed on immediate versus long-term value (for 7-token offers), for each of the four fROIs. The parameter estimates were derived from GLM 1 (see Figure 5.5, p. 128) and correspond to offers early in a trial where accepting a 7-token offer is detrimental overall despite yielding a large immediate reward. The weighting parameter effectively provides a measure of self-control for each individual player, although my task cannot distinguish whether subjects that over-accept 7-token offers do so because they overweight immediate value, or alternatively because they underweight the future consequences of accepting a high value offer (and thus miscalculate long-term value). In vmPFC, a higher BOLD activation in response to 7-token offers was linked to impulsively accepting (trade-off parameter fit closer to 0), while a lower BOLD activation was linked to foregoing the option (trade-off parameter fit closer to 1). This correlation was seen in ventral but not lateral vmPFC (Figure 5.6, p. 131).

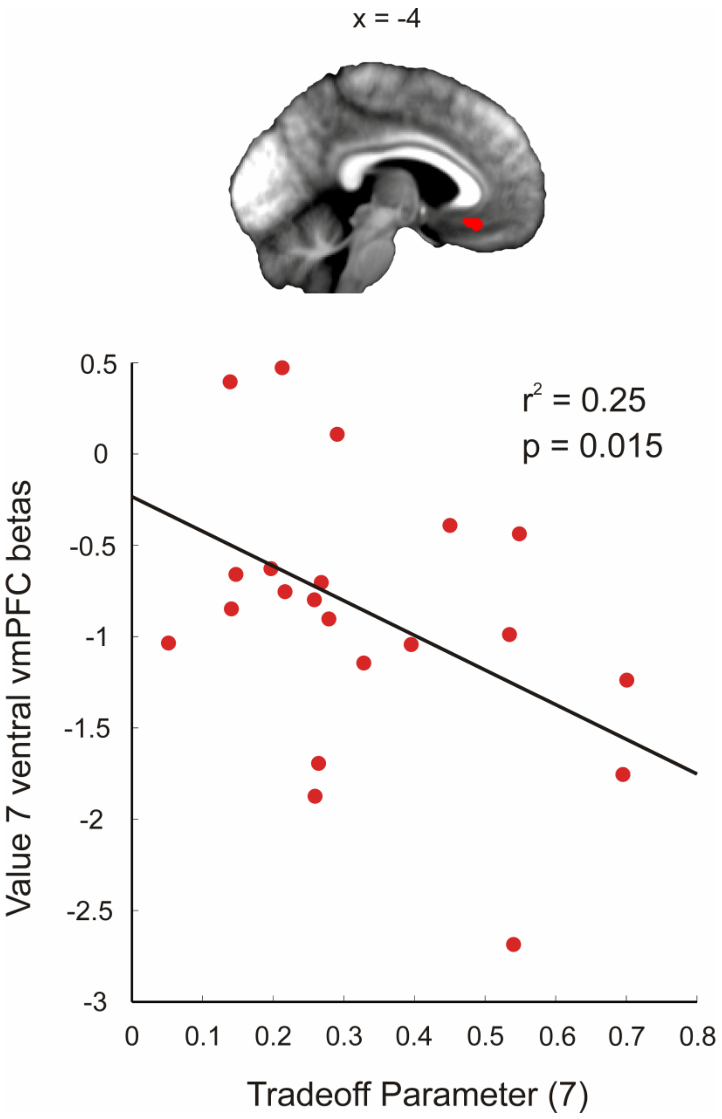


Figure 5.6 When confronted with an offer associated with a high immediate value but low long-term value, between-subject variability in ventral vmPFC BOLD response to 7-token offers was tightly coupled with choice (r^2 = 0.25, p = 0.015). The higher the signal in vmPFC, the more choice was driven by immediate value (more positive beta, trade-off parameter closer to 0). In contrast, the lower the signal in vmPFC, the more choice was driven by long-term value (more negative beta, trade-off parameter closer to 1).

Here, a more negative beta implies a greater weighting on future consequence and a value representation that resembles long-term value, while a more positive beta implies a greater weighting on face value and a value representation that favours immediate rewards. To my surprise, there was no significant correlation in either the left or right caudate fROIs. Thus, while the BOLD response in anterior caudate was similar in both self-controlled and impulsive players (on average), value representations in the most ventral and medial region of vmPFC were tied to each individual's capacity for self-control.

In addition to observing variability in self-control between subjects, players were also highly variable in their ability to exercise control across trials. To test the prediction that trial-by-trial switches between controlled and impulsive choice is linked to a change in the representation of immediate or long-term value, I constructed a new GLM (see Methods, *GLM 3*, p. 120) where I split 7-token offers contingent upon whether they were (incorrectly) accepted or (correctly) resisted. This analysis once again focused on the subset of offers that fall inside the yellow box in Figure 5.3A (p. 124). Note that although a difference in BOLD between go and nogo at the time of cue onset could reflect a modulation of value representation, it could also be driven by the execution of a motor response in one condition and not the other. To control for this motor confound, I regressed out button presses using a motor regressor that included a large proportion of button presses from outside of the yellow box in Figure 5.3A. However, I cannot fully exclude the possibility that any difference observed might be driven by the anticipation of an upcoming action.

Bearing in mind this caveat, I found that a BOLD response to a 7-token offer was on average less negative in lateral but not ventral vmPFC, and more positive in bilateral anterior caudate when subjects chose to incorrectly accept compared to correctly reject (Figure 5.5, p. 128). Thus, impulsive responses were accompanied by a weaker representation of long-term value within lateral vmPFC and an enhanced representation of immediate value in bilateral caudate, while optimal choices followed the reverse pattern. This profile implies that the representational fidelity of one aspect of a value computation may be promoted at the expense of the other.

Previous studies show that self-control recruits the lateral prefrontal cortex (LPFC) with evidence suggesting the ventrolateral prefrontal cortex (VLPFC) acts to initiate inhibitory control (Aron et al., 2004) or in other cases that the ateroventral prefrontal cortex (Diekhof & Gruber, 2010) or dorsolateral prefrontal cortex (DLPFC) (Hare et al., 2009) modulates the representation of value within valuation regions. While my primary interest with imaging was to identify the neural representations of immediate and long-term value, I conjectured that activity in LPFC might scale with the demand for control, and that this in turn may contribute towards the observed representations of value. Within a subset of offers at the start of each trial (yellow box in Figure 5.3A, p. 124), 7-token offers require amplified self-control relative to 3 and 5-token offers, as the immediate value of accepting a 7-token offer here is most decorrelated from the overall long-term value. Thus, the BOLD response in regions enacting 'control' should be enhanced in response to 7-token offers, diminished in response to 5-token offers, and modestly enhanced in response to 3-token offers. Note this is the opposite profile to that observed in vmPFC that encodes long-term (overall) value (see Figure 5.5A/C, p. 124).

I tested for this in a contrast ({0 -1 1} for 3, 5 and 7-token offers) using GLM 1 where I identified activation in a frontal network including anterior cingulate cortex and right inferior frontal gyrus that survived whole-brain correction (see Table 5.1 for all areas, p. 134). Thus, activity in these regions did not scale with value but instead with the demand for control (Figure 5.6, p. 134). Here I also note these regions are strongly implicated in cognitive control (Kerns et al., 2004), response inhibition (Aron et al., 2004) and self-regulated choice (Hare et al., 2009).

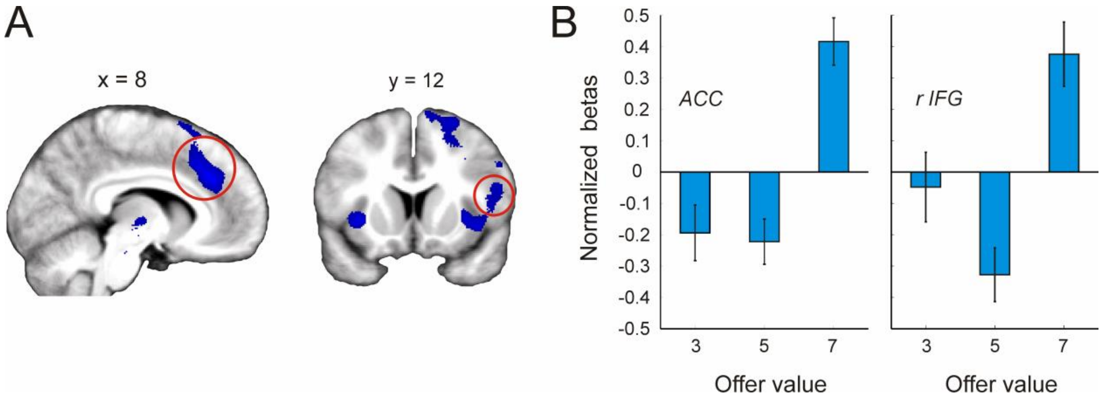


Figure 5.6 (**A**) BOLD response within a frontal network including ACC, rIFG and bilateral insula cortex, was enhanced for 7-token offers compared to 3 and 5-token offers, and thus scaled with the demand for control. (**B**) The betas for clusters in the ACC and rIFG (circled in red) are plotted for illustration. Vertical lines represent SEM. * indicates p =< 0.05; n.s. indicates not significant (paired t-tests); see also Table 5.1 (below).

Name of Region	Cluster FWE p value	MNI Coordinates			Statistics	
		x	y	z	t value	Z score
Anterior Cingulate	< 0.001	10	34	30	5.33	4.77
R Supplementary Motor Area		8	16	66	3.69	3.42
L Insula	0.013	-33	15	0	5.09	4.49
R Insula	< 0.001	32	20	2	4.91	4.36
R Inferior Frontal Gyrus		54	12	19	3.78	3.50
R Parietal	< 0.001	58	-43	34	4.23	3.85

Table 5.1 Regions where BOLD covaried with the demand for action control ({ 0 -1 1 } for offers of token-value 3, 5 and 7 respectively) from *GLM 1*.

5.4 Discussion

Both vmPFC and striatum are implicated in computing value for action selection (Brovelli et al., 2011; FitzGerald, Friston, & Dolan, 2012; Guitart-Masip et al., 2012; Tricomi et al., 2004; Wunderlich, Dayan, et al., 2012), and these regions are differentially activated when individuals choose immediate versus delayed rewards (McClure et al., 2004). Whether this distinction arises from divergent computational roles has remained unclear. Here, I used a computational formalization to address how vmPFC and striatum arbitrate between immediate and long-term value where these are dissociable and can motivate differing actions. Further, by contrasting incorrect and correct decisions I could map the computational mechanisms that contribute towards impulsive or controlled choice respectively.

Previous studies have proposed that choice utilizes a common value system based in vmPFC (Hare et al., 2009), or in a vmPFC-striatal loop (Kable & Glimcher, 2007). Consistent with this, I identified a value representation in vmPFC that takes into account the immediate and delayed consequences of actions. However, in contrast to the common value framework, I identified a separate representation of immediate value in anterior caudate that likely impacts action selection in parallel, and in a fashion that often opposes a course of action endorsed by vmPFC. In this scheme, failures of self-control stem from a degraded representation of long-term value in lateral vmPFC and a concurrent enhancement of immediate value within anterior caudate. Analogously, successful control is not only dependent on an accurate representation of long-term value in lateral vmPFC, but also an attenuation of immediate value in anterior caudate.

There are several possible explanations for the discrepancy between my finding that the brain represents dual values and previous accounts that it uses a single value system. In the

Hare choice paradigm (Hare et al., 2009), subjects chose between a reference food item and alternatives that varied in healthiness and tastiness. The authors then asked whether the BOLD response significantly correlated with taste or health ratings in subjects who demonstrated either high or low capacity for self-control. However, this analysis was confined to the vmPFC, and it is possible that activity in anterior caudate may have tracked taste ratings in a manner similar to the immediate value representations that I observed in my data. Further, while tastiness and healthiness map onto different outcome modalities, my task considers immediate and long-term value attributes within a single modality. A second prominent study closely aligned with the single value account utilized an intertemporal choice paradigm to probe preference for rewards at differing time-scales (Kable & Glimcher, 2007). Here, subjects had to choose between an immediately available sum of money and a larger but delayed alternative. Similar to results reported here, Kable and Glimcher found that vmPFC (amongst other regions) computes the subjective value of the chosen option. However, since the immediate reward was kept constant in their design, it remains unknown whether this value is tracked separately in the brain.

Another important consideration is that unlike the previous studies, my task did not require a choice between two options presented simultaneously. Rather, subjects were required to flexibly approach or avoid an option with both immediate and delayed consequences, spanning both action and valence (Guitart-Masip et al., 2012). This action dependency was adopted so as to more closely resemble natural settings, where self-control often involves arbitration between approach and avoidance, and where the value of choice options often change dynamically. Given that the striatum is heavily implicated in both action and value processing (Guitart-Masip et al., 2014; Rothwell, 2011; Samejima, Ueda, Doya, & Kimura, 2005), and that the distinction between these roles is not clearly defined, anterior caudate may in fact integrate value with a propensity to act during go/nogo judgments (Guitart-Masip et al., 2011; Guitart-Masip et al., 2012; Roesch et al., 2009). In turn, this contribution may be absent in self-control tasks that do not pair the prepotent choice (accepting a large immediate reward) with a prepotent action (the execution of a 'go' response). Other evidence that task modality can impact value coding comes from a recent finding that switching the frame of reference used for decision-making alters patterns of value coding in the brain (Hunt, Woolrich, Rushworth, & Behrens, 2013).

In humans, activity in vmPFC has been shown to include a representation of healthiness in individuals who resist temptation for unhealthy foods (Hare et al., 2009), a finding complimented by evidence that vmPFC acts to integrate multiple components of value (Wunderlich, Dayan, et al., 2012). Further, in rodents, the orbitofrontal cortex has been shown to compute values based on anticipation of latent outcomes (Jones et al., 2012), while patients with bilateral vmPFC lesions demonstrate reduced sensitivity to future consequence and increased reliance on immediate rewards (Bechara et al., 2000). However, to the best of my knowledge, no previous study has demonstrated a value signal in human vmPFC that reflects an overall (long-term) value that is decoupled from immediate rewards. This points to the likelihood that vmPFC draws on contextual information to calculate an overall expectation of value (Hampton et al., 2006b; Jones et al., 2012; McDannald, Lucantonio, Burke, Niv, & Schoenbaum, 2011; Takahashi et al., 2013), while other valuation regions may only be privy to immediate outcomes.

I found value coding in a more ventral region of vmPFC is dependent on subjects' baseline ability to appropriately adjust a prepotent response, raising an important question regarding the underlying mechanism. One conjecture is that this region lacks access to representations required for inferring long-term value in impulsive players. This may be related to a weaker functional connectivity between this region of vmPFC and more dorsal prefrontal cortex regions associated with goal-directed control (Hare et al., 2009; Hare et al., 2014). By contrast, value coding in a more lateral region of vmPFC was predictive of upcoming choice in a context requiring self-control. While I can only speculate as to the functional differences between these regions, one possibility is that the ventral portion encodes long-term value regardless of context, whereas the more lateral portion integrates long-term value with additional components that contribute to the action selection process, and is thus more representative of upcoming choice. Interestingly, a recent study has identified a similar pattern of differential reward processing within subregions of vmPFC in non-human primates (Monosov & Hikosaka, 2012).

My finding that anterior caudate predominantly tracks immediate value is surprising given previous accounts that this region represents the utility of actions by differentiating between positive and negative consequences (Tricomi et al., 2004), or computing values for planned choice (Wunderlich, Dayan, et al., 2012) and future reward prediction (Tanaka et al., 2004). A long-line of animal research has implicated the dorsomedial striatum (the caudate homologue in rodents) in representing the consequences of an animal’s actions, with lesions to this region impairing the acquisition of R-O contingencies (Yin, Ostlund, et al., 2005). Yet, much of the animal literature relies on devaluation paradigms that utilize immediate outcomes (Balleine & O'Doherty, 2010). Similarly, experiments in humans have implicated anterior caudate in outcome devaluation (Valentin et al., 2007) and in tracking contingencies between actions and outcomes (Tanaka, Balleine, & O'Doherty, 2008), yet often do not require valuations that integrate immediate and long-term consequences. Thus, one possibility is that both vmPFC and anterior caudate support goals by representing outcomes (Valentin et al., 2007), while vmPFC predominantly receives the input required to calculate long-term value. An alternative interpretation, given a finding that at least some component of the anterior caudate response is explained by long-term value, is that this region contains populations of neurons tuned to either immediate or long-term value respectively.

Although I used a model-based tree search to define overall value for the purposes of my analysis, my task cannot differentiate between model-based versus alternate choice strategies. For example, the use of heuristics may be more probable given the complexity of the tree search. Further, subjects' probability of accepting an offer between offer index 2 and 3 in a trial (see Figure 5.3A, yellow boxes, p. 124) is somewhat uniform, and this choice pattern is not well-captured by the winning model. Yet my key interest lay in exploring the behavioural and neural consequences of dissociating immediate from overall (long-term) value, and the trade-off model provides corroborative evidence that subjects take both quantities into account. An important follow-up question is whether long-term value is calculated online by projecting into the future, or whether it is cached and retrieved in a model-free framework following a sufficient number of trials.

The data from this study have a number of implications. A comorbidity between impulsivity and selected psychiatric disorders is well-documented (Moeller, Barratt, Dougherty, Schmitz, & Swann, 2001), raising an interesting question as to the relationship between the biological substrates of these disorders and the dissociable value representations I identify. The current task might provide a novel avenue for probing this, including assessing the impact of both behavioural and pharmacological interventions. Finally, given a strong association between affective state and the capacity for self-control, the dual-value framework outlined could be useful for evaluating the impact of emotion, mood, stress, and other state-dependent factors on the representation of immediate and long-term value, and the resulting impact on decision-making in these contexts.

CHAPTER 6

THE EFFECTS OF TASK TRAINING ON MODEL-BASED REASONING

The brain has been suggested to employ multiple distinct strategies for solving problems using habitual (model-free) or goal-directed (model-based) algorithms. Hitherto, model-based reasoning has been identified with slow, serial, executive processes, and model-free with fast, parallel, automatic processes. In a task that engages both model-based and model-free systems, increasing cognitive load with a challenging concurrent task reduces the expression of model-based behaviour, consistent with the idea that a shared, limited pool of cognitive resources is used for model-based calculations and the concurrent task. Here, however, I show that this impairment in model-based reasoning under load is eliminated when subjects receive prior primary task training, whether or not the training is under load. Thus, task familiarity permits model-based reasoning even under substantial cognitive load. These data suggest a shift in the mechanism by which model-based calculations are implemented with increasing task exposure and also imply that model-based reasoning can be dissociated from serial executive functions.

6.1 Introduction

A wealth of experimental data shows that the brain makes use of at least two distinct decision strategies. One system prospectively reasons about action-outcome contingencies, while the other retrospectively links rewards to actions (Balleine & O'Doherty, 2010; Daw et al., 2005; Dolan & Dayan, 2013; Loewenstein, 1996). The interplay between these two choice strategies has substantial practical implications. For example, over-reliance on habits could lead to inflexible decision-making in addiction (Everitt & Robbins, 2005) and compulsion (Voon et al., 2014).

A compelling computational account of these two mechanisms draws on reinforcement learning (RL) theory (Daw et al., 2005). In Daw and colleagues' framework, model-free RL exploits temporal difference mechanisms (R. S. B. Sutton, A. G., 1998) closely associated with striatal dopamine signals (Montague, Dayan, & Sejnowski, 1996) to learn a preference for actions through direct reinforcement (Dayan & Niv, 2008). Model-based RL, on the other hand, prospectively evaluates actions by mapping the contingencies between actions and future states (Daw et al., 2005; Dayan, 2008; Dayan & Niv, 2008). This renders the model-based system more flexible, but at a heightened computational cost.

Contemporary theories posit that model-based reasoning engages limited-resource executive functions (Donald & Tim, 1986) associated with regions of prefrontal cortex (PFC), in particular the dorsolateral prefrontal, ventromedial prefrontal and anterior cingulate cortices (Alvarez & Emory, 2006; Barbey et al., 2012; M. M. Botvinick et al., 2001; Glascher et al., 2010; Jones et al., 2012; Kennerley et al., 2006; S. W. Lee et al., 2014; Owen, 1997; Valentin et al., 2007; Wunderlich, Dayan, et al., 2012). Further evidence for this view comes from the observations that model-based reasoning is impaired by increasing cognitive load (Otto, Gershman, et al., 2013), by disrupting dorsolateral prefrontal cortex function (Smittenaar et al., 2013) and by acute stress (Otto, Raio, Chiang, Phelps, & Daw, 2013), with the degree of impairment often interacting with baseline working memory capacity.

However, studies of model-based decision-making often utilize tasks in which the stimuli, contingencies and other task parameters are novel to the subject. Thus, one possibility is that reliance on limited-resource executive functions is not an intrinsic property of model-based reasoning, but rather a characteristic of reasoning with an unfamiliar model. This is consistent with the everyday experience that practice lets us perform increasingly complex tasks with less demand for exclusive attention, and may be important for the human ability to progressively acquire more complex behaviour. Nevertheless, the effect of training on model-based and model-free decision making remains unexplored.

Here, I used a two-step decision-task that engages and measures both model-free and model-based reasoning (Daw et al., 2011; Otto et al., 2013). Using simple behavioural analyses as well as more sophisticated computational modelling, I quantified the degree to which model-free and model-based reasoning were manifest in choice, both before and after task training, and with or without cognitive load. I hypothesized that a shift in the neural mechanism for model-based calculations, as a result of task training, could lead to a reduction in the detrimental effect of cognitive load on model-based reasoning.

6.2 Methods

Subjects

35 adult participants formed a group (referred to as the 'high load group') which received training both with and without cognitive load, of which 22 were included in the final analysis (7 male and 15 female; age range 18-34; mean 21.5, SD = 3.71 years).

30 adult participants formed a second independent group (referred to as the 'low load group') for which cognitive load was omitted from training on days 1 and 2, of which 23 were included in the final analysis (9 male and 14 female; age range 18-26; mean 21.2, SD = 3.61 years).

Inclusion criteria: In line with (Otto, Gershman, et al., 2013) I excluded 11 subjects from the 'high load group' and 5 subjects from the 'low load group' whose accuracy on the Stroop task during dual-task trials was < 70% on any given day so as to ensure participants were in fact attempting to perform both tasks simultaneously. In addition I excluded 2 participants from the 'high load group' and 1 participant from the 'low load group' who chose the same first-stage fractal on > 90% of trials (on any given day), irrespective of events on the previous trial. Finally I excluded 1 participant from group two whose probability of repeating a first-stage action following a common-rewarded transition on the previous trial was < 0.25 on day 1 of training.

General design

Subjects in the 'high load group' performed alternating blocks of two-step (128 trials) and dual-task (64 trials) trials until two blocks of each trial type were completed (256 two-step trials, 128 dual-task trials in total). This protocol was repeated across three consecutive days. Subjects received 20 practice trials of each trial type at the start of day 1. Subjects in the 'low load group' performed 256 trials of the two-step task on each of two consecutive days. On day 3, they performed alternating blocks of two-step (128 trials) and dual-task (64 trials) trials until two blocks of each trial type were completed (256 two-step trials, 128 dual-task trials in total). Thus, day 3 was identical in both group protocols. Subjects received 20 practice trials of two-step task at the start of day 1, and 20 practice trials of dual-task at the start of day 3.

Task

Subjects performed a two-step decision task based on (Daw et al., 2011) and equivalent to that used in (Otto, Gershman, et al., 2013). At the first stage, a player was presented with two fractal images presented side-by-side on a grey background and had 2000 ms to select one via a left or right button press. After a response was made the selected fractal was highlighted for the remainder of the choice period with a yellow boarder. Each first stage fractal lead to one of two second stage fractal pairs with a probability of 70% (common transition) and to the other with a probability of 30% (uncommon transition). Following the transition, one of two second stage pairs of fractals was displayed on a green or blue background in accordance with whether a common or uncommon transition had occurred. In addition, the chosen first-stage fractal was minimized and moved to the top central portion of the screen. The player again had 2000 ms to select a fractal via a left or right button press, and the selected action was highlighted for the remainder of the response period. Finally, an outcome was presented in the form of a golden coin (to indicate a monetary gain) or a ‘0’ (to indicate no monetary gain), followed by an inter-trial interval (fixation cross). The position of each fractal (left versus right) was counter-balanced across trials for first and second-stage pairs.

Dual-task trials followed the same procedure, except that subjects had to additionally perform a numerical Stroop task (Waldron & Ashby, 2001). At the beginning of the first-stage, two digits were presented, one above each choice fractal, for 200 ms, and then covered by a white mask for a further 200 ms. After second-stage choice feedback, either the word ‘SIZE’ or ‘VALUE’ appeared alone in the centre of the screen on a grey background. The player had 1000 ms to indicate with a left or right button press which digit of the two that appeared at the first-stage choice was larger in size or value, respectively. In accordance with (Otto, Gershman, et al., 2013) and (Waldron & Ashby, 2001), the numerically larger number was physically smaller on 85% of trials. Thus, subjects had to hold incidental information in working memory whilst performing the two-step task. Following their response, feedback in the form of the word ‘CORRECT’ or ‘INCORRECT’ was presented a further 1000 ms. If participants failed to respond during the Stroop task probe, a red “X” appeared for 1000 ms. Trial lengths were equated across two-step and dual task trials (7200 ms per trial).

The reward probabilities associated with second-stage fractals were governed by independently drifting Gaussian random walks (SD = 0.025). I generated a pool of fifteen random walks for which reward probabilities did not exceed ~0.75 or fall below ~0.25. For each subject, three walks were selected at random from the pool for use on each successive day of training. Thus, walks were continuous between blocks of two-step and dual task trials.

Logistic regression

In keeping with previous studies (Daw et al., 2011; Otto, Gershman, et al., 2013; Smittenaar et al., 2013; Wunderlich, Smittenaar, & Dolan, 2012), I first probed model-based versus model-free reasoning by analysing stay-switch behaviour at the first-stage of each trial. Model-free reinforcement learning predicts that first stage choices should be repeated if they lead to a reward (a main effect of reward), regardless of whether a common or uncommon transition is experienced on the previous trial. By contrast, a model-based learner is more likely to switch their choice at the first stage if a reward follows from an uncommon transition on the previous trials (a reward x transition interaction). This is because the model-based system can infer that the rewarding second-stage fractal can be accessed with a higher probability by choosing the alternate first-stage fractal. In short, by evaluating the dependence of switch-stay choice on the reward and transition status from the preceding trial (and their interaction), one can approximate the strength with which model-free and model-based reasoning are manifest in choice.

I performed a random-effects logistic regression, implemented in the Matlab software package (MathWorks), in which the dependent variable was the first-stage choices in the current trial (coded as 0 for stay, 1 for switch), and the explanatory variables included the reward and transition type on the previous trial (coded as 1 or -1), and their interaction. Blocks of trials from the same day of training were concatenated, and trials where subjects failed to respond at either the first or second-stage were excluded from the analysis. When analysing data across all days, I included a variable for the day of training, in addition to all possible interactions (see Table 6.1 for all variables, p. 160). Here, my key interest lay in the

2-way interaction between reward and transition, and whether this interaction changes with training in the dual-task condition (a 3-way reward x transition x day interaction). Since these two regressors were highly correlated, I orthogonalized the latter with respect to the former, using a Gram-Schmidt process (Bjorck, 1994). Thus, any significant 3-way interaction represents a proportion of variance unaccounted for by a simple 2-way effect. One-sample t-tests were performed on all coefficients across subjects. When analysing dual-task trials from the 'high load group', I performed an additional logistic regression where I included Stroop task performance on the previous trial (coded as 1 for correct, -1 for incorrect), and all possible interactions (see Table 6.2 for all variables, p. 161), as additional predictors. Here my main interest was whether errors on the Stroop task would interfere with subjects' ability to use reward and transition events on the previous trial to make a model-based choice on the following trial (a Stroop performance x reward x transition interaction). As before, 3 and 4-way interactions were orthogonalized with respect to the simpler 2 or 3-way effect.

In line with other recent studies that have used the two-step task, I also considered model-free and model-based influences on choice in the current trial, with respect to events that occurred up to 3 trials in the past (Smittenaar, Prichard, FitzGerald, Diedrichsen, & Dolan, 2014). Here, the dependent variable on trial *t* was 1 when stimulus A was chosen and 0 when stimulus B was chosen at the first-stage. Each regressor then described whether events on trial t_{-1}, t_{-2} and t_{-3} would increase (coded as +1) or decrease (coded as -1) the probability of choosing A according to a model-free or a model-based system (6 regressors in total). Importantly, if a trial involved a common transition, both systems make identical predictions. However, opposing predictions emerge following uncommon transitions. I implemented a random-effects logistic regression in Matlab (MathWorks) and performed one-sample t-tests on the resulting coefficient estimates for the 6 regressors, separately for trained (day 3) versus un-trained (day 1), and high load (dual-task) versus low load (two-step) (see Figure 6.5, p. 170).

Computational modelling

Based on (Daw et al., 2011), the task was modelled as consisting of three states (s_A for the first-stage fractal pair; s_B and s_C for the second-stage fractal pairs) where two possible actions (a_A, a_B) can be taken from each state. The goal of each RL algorithm is to learn a state-action value function $Q_{(s,a)}$that maps each state-action pair to its expected future value. In each trial *t,* the first and second-stage states are indicated as $s_{1,t}$ and $s_{2,t}$ respectively, while first and second-stage choices (actions) are indicated as $a_{1,t}$ and $a_{2,t}$. Since there is no reward at the first stage , $r_{1,t}$ is always zero, while $r_{2,t}$ can be zero or one.

Model-free

The model-free algorithm was temporal difference Q-learning (R. S. B. Sutton, A. G., 1998) in which the value of a given state is assumed to be equivalent to the expected reward from taking the best available action from that state. At each stage *i* of each trial *t*, the value of the chosen state-action pair was updated according to:

$$Q_{TD}\left(s_{i,t}, a_{i,t}\right) = Q_{TD}\left(s_{i,t}, a_{i,t}\right) + \ \alpha\delta_{i,t}$$

where δ, the reward prediction error (RPE), is defined as

$$\delta_{i,t} = r_{i,t} + \gamma \ \max_a\left[Q_{TD}(s_{i+1,t}, a)\right] - \ Q_{TD}(s_{i,t}, a_{i,t})$$

where α is a learning rate fit for each subject and γ is a discount factor that trades off the importance of sooner versus later rewards (fixed at 1).

Note that for the first stage choice, $r_{i,t}$ is always zero and δ is instead driven by the second-stage value.

After outcome delivery, the second stage RPE is used to update the first-stage action $Q_{TD}(s_{1,t}, a_{1,t})$ according to the eligibility trace λ, which assigns credit to the first-stage action without the need for an additional step.

$$Q_{TD}(s_{1,t}, a_{1,t}) = Q_{TD}(s_{1,t}, a_{1,t}) + \alpha\lambda\delta_{2,t}$$

Thus, in the event that λ=0, choice is driven by the estimated value of the second-stage state on the previous trial. Consistent with previous studies (Daw et al., 2011; Otto, Gershman, et al., 2013), this model assumes that eligibility traces are cleared between trials.

Model-based

A model-based RL algorithm involves learning a set of contingencies between actions and states (a state-transition function), estimating a reward value for each state, and then combining the two by iterative expectation. Here, since first-stage transitions are probabilistic, a player must map action-state pairs to a probability distribution over subsequent states.

One can approximate subjects' estimate of the transition probabilities by assuming they believe one of two alternatives:

$$P(s_B|\, s_A\,, a_A) = 0.7, P(s_C|\, s_A\,, a_A) = 0.3, P(s_C|\, s_A\,, a_B) = 0.7, P(s_B|\, s_A\,, a_B) = 0.3$$

or

$$P(s_B|\, s_A\,, a_A) = 0.3, P(s_C|\, s_A\,, a_A) = 0.7, P(s_C|\, s_A\,, a_B) = 0.3, P(s_B|\, s_A\,, a_B) = 0.7$$

based on the number of previous transitions from s_A to s_B given a_A and from s_A to s_C given a_B (or vice versa). A previous study has shown this scheme settles on the true transition matrix after the first few trials and fits subjects' choices better than implementing a traditional trial-by-trial learning algorithm (Daw et al., 2011). Therefore I assume the true transition probabilities are learnt during practice trials and are known by the start of the first experimental block.

Since the second-stage action is the only choice associated with immediate reward, and is the final step in a trial, an agent can learn the value of the second-stage state in a manner equivalent to temporal difference Q-learning (as above). Thus, $Q_{TD}(s_{2,t}, a_{2,t})$ is simply an estimate of the immediate reward $r_{2,t}$, and the model-based algorithm converges with model-free learning at this stage.

By combining the transition function with the second-stage values I can define the values of the two first-level actions (using Bellman's equation) as follows:

$$Q_{MB}(s_A, a_j) = P(s_B | s_A, a_j) \max_a [Q_{TD}(s_B, a)] + P(s_C | s_A, a_j) \max_a [Q_{TD}(s_C, a)]$$

where these are computed on every trial based on the updated second-stage Q-values.

Hybrid model

For the hybrid model I consider contributions from both model-free and model-based RL. First-stage action values were defined as the weighted sum of values from the algorithms described above as follows:

$$Q_{HM}(s_A, a_j) = wQ_{MB}(s_A, a_j) + (1 - w)Q_{TD}(s_A, a_j)$$

where ***w*** is a weighting parameter.

When fitting data across all sessions, I included a slope parameter ***s*** that allowed ***w*** to shift across days:

$$w_D = w[\exp(s(Day - 2))]$$

and used $\boldsymbol{w_D}$ as the new weighting parameter.

At the second-stage, all three models (model-free, model-based, hybrid) converge.

Action selection

For each model, values were converted to action probabilities using a sigmoid (softmax) function:

$$P(a_{A,t}) = \varepsilon + \frac{1 - 2\varepsilon}{1 + \exp(-\beta[Q(s_{i,t}, a_{A,t}) - Q(s_{i,t}, a_{B,t})])}$$

where ***ε*** is a lapse rate, such that when $\boldsymbol{\varepsilon} > 0$ the boundaries of the sigmoid function are compressed and deviations from the model are less harshly punished, and ***β*** is an inverse temperature parameter that governs the stochasticity of choice options.

Model sets

When fitting data from individual days, I considered a hybrid RL model that included a single learning rate (***α***) and softmax temperature (***β***), a weighting parameter that governs the balance between model-free/model-based control (***w***), and a lapse rate (***ε***). The eligibility trace (***λ***) was fixed at 1. Model-free and model-based algorithms were nested versions of the hybrid model where ***w*** was set to 0 and 1 respectively.

When fitting data across all days, I considered a family of (nested) hybrid RL models in which specific parameters were omitted or included as fixed versus free parameters. More complex models included separate RL parameters for first and second stage choices, an eligibility trace, and a slope parameter that permitted the weighting between model-free and model-based control to shift across days. See Table 6.4, p. 167, for the full model set.

Model comparison

As described previously in this thesis I used a hierarchical Type II Bayesian (or random effects) procedure using maximum likelihood to fit simple parameterized distributions for higher level statistics of the parameters. Since the values of parameters for each subject are 'hidden', this employs the Expectation-Maximization (EM) procedure. Thus on each iteration the posterior distribution over the group for each parameter is used to specify the prior over the individual parameter fits on the next iteration. For each parameter I used a single distribution for all participants. Before inference, all parameters were suitably transformed to enforce constraints (log and inverse sigmoid transforms).

The model fitting routine follows that previously described by Huys and colleagues (Huys et al., 2011). Each model yielded a parameter vector, θ_i, for each subject, i. Before inference, all parameters were suitably transformed to enforce constraints (log and inverse sigmoid transforms). Model fitting at the individual level aimed to find the maximum a posteriori estimate of θ_i, given a vector of each subject's choices, C_i:

$$\theta_i = argmax_\theta \; p(C_i|\theta_i)p(\theta_i|\vartheta)$$

I used a hierarchical (random effects) model-fitting approach, with the assumption that parameter estimates were normally distributed at the group level, where ϑ are the parameters of the empirical normal prior distribution (hyperparameters) on θ . The hierarchical approach allows the population-level distribution of data to constrain unreliable parameter estimates at the individual level. I estimated the maximum-likelihood hyperparameters, given the data from all N subjects:

$$\hat{\vartheta}^{ML} = argmax_{\vartheta}\, p(C_1 \dots C_N|\vartheta) = argmax_{\vartheta} \prod_i p(\, C_i|\vartheta)$$

where:

$$p(C_i|\vartheta) = \int d\,\theta_i\, p(C_i|\theta_i)p(\theta_i|\vartheta)$$

The intractable integral above was estimated by Expectation-Maximization (EM). The E-step at the kth iteration sought the maximum a posteriori parameter estimates for each subject (given an estimate of the empirical prior from the preceding iteration, achieved by unconstrained nonlinear optimization in Matlab, Mathworks, MA, USA):

$${\theta_i}^{(k)} = argmax_{\theta}\, p(C_i|\theta_i)p(\theta_i|\vartheta^{(k-1)})$$

I used a Laplace approximation, which assumes that the likelihood surface is normally distributed around the maximum a posteriori parameter estimate:

$$p(\theta_i|C_i) \approx N\left({\theta_i}^{(k)}, \Sigma_i^{(k)}\right)$$

where $\sum_i^{(k)}$ is the second moment around ${\theta_i}^{(k)}$, which approximates the variance. In the M-step, the estimated hyperparameters $\vartheta^{(k)}$ of the normal prior distribution, mean μ, and factorized variance, σ^2, were updated as follows:

$$\mu^{(k)} = \frac{1}{N}\sum_i {\theta_i}^{(k)}$$

$$\left(\sigma^{(k)}\right)^2 = \frac{1}{N}\sum_i \left[\left({\theta_i}^{(k)}\right)^2 + \Sigma_i^{(k)}\right] - \left(\mu^{(k)}\right)^2$$

I compared models by Bayesian model evidence, $p(C_1 \dots C_N|M)$, approximated as BIC_{int} :

$$-\frac{1}{2}BIC_{int} = \log p(C_1 \dots C_N|\hat{\vartheta}^{ML}) - \frac{1}{2}|M|\log(|C_1 \dots C_N|)$$

where $|C_1 \dots C_N|$ is the total number of choices made by all subjects, and $|M|$ is number of hyperparameters fitted. Notably here, by distinction from conventional BIC, $\log p(C_1 \dots C_N|\hat{\vartheta}^{ML})$ is a sum over the model evidence at the subject level by integrating over subject-level parameters:

$$\log p(C_1 \dots C_N|\hat{\vartheta}^{ML}) = \sum_i \log \int d\theta \;\, p(C_i|\theta)\, p(\theta|\hat{\vartheta}^{ML}) \;\approx\; \sum_i \log \frac{1}{K} \sum_{k=1}^{K} p\left(C_i|\theta^k\right)$$

The right hand expression approximates the integral by summing over K samples, drawn from the empirical prior, $p(\theta|\hat{\vartheta}^{ML})$. Thus the individual-level parameters intervene between the data and the group-level inference, but are averaged out when comparing models.

6.3 Results

I employed the two-step task introduced by Daw and colleagues (Daw et al., 2011), and used in many recent studies (Otto, Gershman, et al., 2013; Otto, Raio, et al., 2013; Skatova, Chan, & Daw, 2013; Smittenaar et al., 2013; Smittenaar et al., 2014; Voon et al., 2014; Wunderlich, Smittenaar, et al., 2012), which measures separate model-free and model-based influences on choice. Each trial of the task consists of two stages, where each stage involves a two-alternative forced choice between a pair of adjacent fractal images (Figure 6.1, p. 155). Choice at the first-stage always involves the same two fractals, whereas choice at the second-stage involves one of two distinct pairs of fractals. The first-stage choice is causally related to a transition to the second-stage, where each first-stage fractal is predominantly associated (with a 70% probability) with one of the second-stage pairs. The transitions with 70%

probability I call "common"; those with 30% "uncommon". In turn, each second-stage fractal is associated with a probabilistic reward. Importantly, these reward probabilities are different for each second-stage fractal, and fluctuate independently across a session. Thus, subjects have to make trial-by-trial adjustments in choice so as to maximize the current probability of reward.

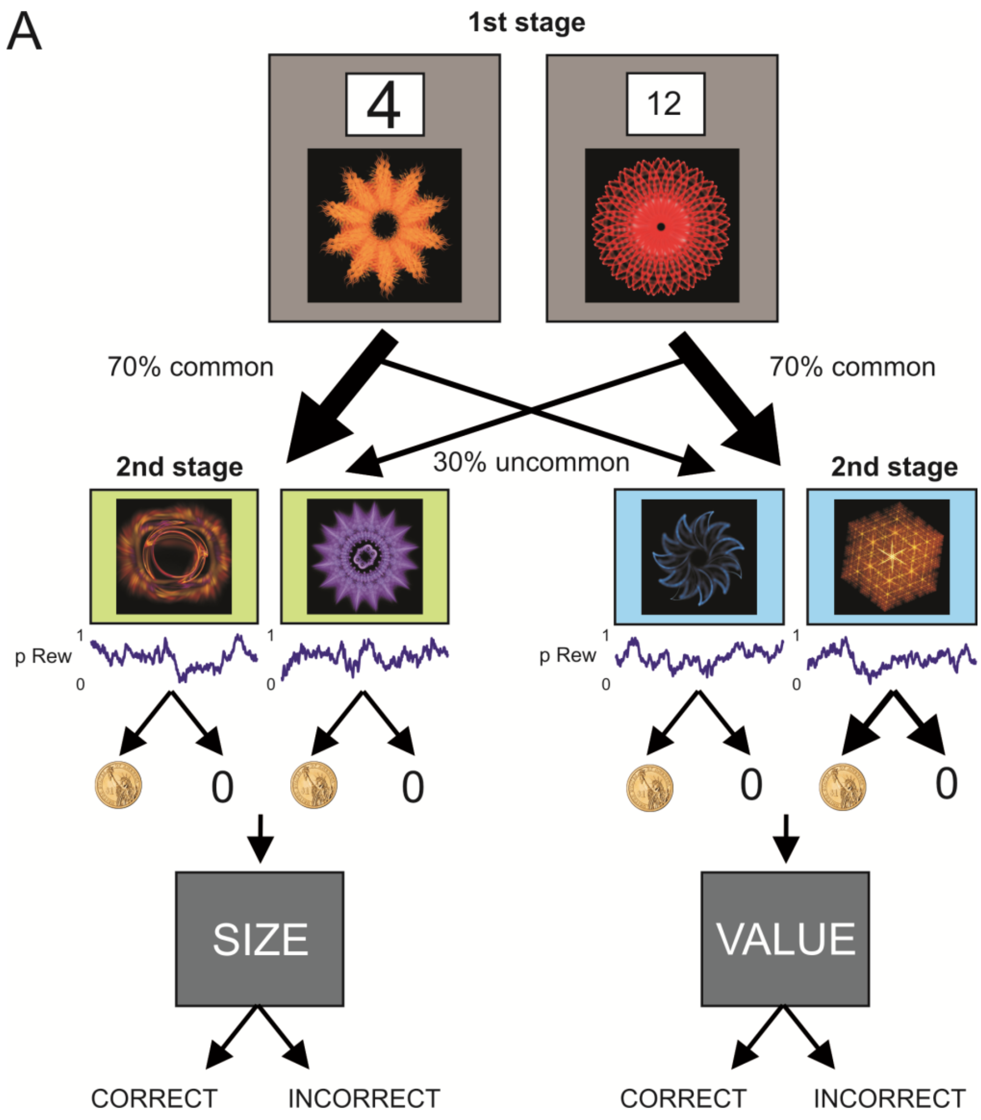


Figure 6.1 On every trial of the two-step task, a choice between a pair of fractals (first-stage) led probabilistically to a second pair of fractals (second-stage), of which one fractal had to again be chosen. The second-stage choice followed either a reward (gold coin) or no reward (0), according to the reward probability associated with the chosen second-stage fractal, which fluctuated over time. Importantly, each first-stage choice led predominantly (on 70% of occasions) to one of the two second-stage pairs, and this transition structure could be exploited by the player. On dual-task trials only (displayed in this figure), two different numbers of physically different sizes were displayed briefly above each fractal at the start (first-stage) of the trial. After receiving second-stage reward feedback, either the word 'SIZE' or 'VALUE' was presented on the screen, and the player had to indicate whether the number that was larger in size, or value, respectively, had previously appeared on the left or right side of the screen. Correct responses were incentivized via a small monetary gain while

incorrect responses were unrewarded. The requirement to retain information pertaining to the numerical task whilst solving the two-step task bore a heightened cognitive load.

Model-free and model-based decision strategies make different predictions about how choice depends on transitions and rewards from previous trials. I used a variety of analysis methods to estimate the relative contribution of model-free and model-based strategies when subjects performed the two-step task alone or in combination with a demanding concurrent task. I also tested whether the effect of the concurrent task changed with practice. I trained subjects on the two-step task for 3 consecutive days with short periods of concurrent task at various periods throughout the training. An initial group of 22 healthy subjects, referred to as the 'high load group', experienced the high load condition on each day of training. This allowed me to characterize choice under load across the entire training period. In contrast, a second group of 23 healthy subjects, referred to as the 'low load group', experienced the high load condition only on day 3. This allowed me to determine how training on the two-step task alone would impact choice under load, and thus provided a more conservative test of my hypothesis.

Switch-stay choice strategy

As with previous studies utilizing the two-step task (Daw et al., 2011; Otto, Gershman, et al., 2013; Otto, Raio, et al., 2013; Skatova et al., 2013; Smittenaar et al., 2013; Smittenaar et al., 2014; Voon et al., 2014; Wunderlich, Smittenaar, et al., 2012), I first examined pairs of successive choices in isolation. Specifically, one can estimate the contribution of model-based versus model-free reasoning by means of a two-factor analysis of the effect of the previous trial's reward and transition type on the first-stage choice (switch versus stay) in the current trial (Daw et al., 2011). Here, a model-free strategy predicts that a player should repeat first-stage choices that lead to rewards at the second-stage, regardless of the transition between states (a main effect of reward on the probability of repeating the previous action). By contrast, a model-based strategy predicts that a player should switch their first-stage choice if a reward is delivered after an uncommon transition, generating a higher probability of reaching a rewarding second-stage state (a cross-over reward x transition interaction in the probability of repeating the previous action). Note these predictions stem from the assumption that subjects make choices based only on events occurring on the immediately preceding trial, and thus provide only an approximate measure of the balance between model-free and model-based reasoning.

I first analysed data from the 'high load group'. Consistent with previous research (Daw et al., 2011; Otto, Gershman, et al., 2013; Smittenaar et al., 2013; Wunderlich, Smittenaar, et al., 2012), I found that subjects' first-stage choices on the two-step task were indicative of a mixture of both model-based and model-free reasoning on day 1 of training (Figure 6.2A, p. 158). This choice strategy remained stable across days, with a logistic regression revealing both a main of effect of reward (all $p < 0.01$) and a reward x transition interaction (all $p < 0.005$) on all 3 days. By contrast, both model-based and model-free reasoning were disrupted on day 1 of dual-task training (Figure 6.2A, p. 158). Here, first-stage choices did not reveal a main effect of reward, and although I identified a reward x transition interaction ($p < 0.05$), it was not characterized by a full cross-over as predicted by a model-based strategy. Importantly however, subjects' behaviour in the dual-task condition shifted across days. I identified a reward x transition interaction ($p < 0.003$), but no main effect of reward, on days 2 and 3 of dual-task performance. Thus, on these days, subjects' choices were consistent with model-based control despite being under heavy cognitive load.

When considering data from the 'low load group' I again observed a choice pattern indicative of both model-free and model-based reasoning across all 3 days of two-step training (main effect of reward, all $p < 0.05$; reward x transition interaction, all $p < 0.05$), but no change in behaviour across days (reward x transition x day, $p = 0.40$) (Figure 6.2B, p. 158). By contrast, dual-task performance on day 3 exhibited a reward x transition interaction (p < 0.001) but no main effect of reward (p = 0.59). These analyses suggest that task training renders model-based reasoning resistant to load, whether training occurs in the presence or absence of load.

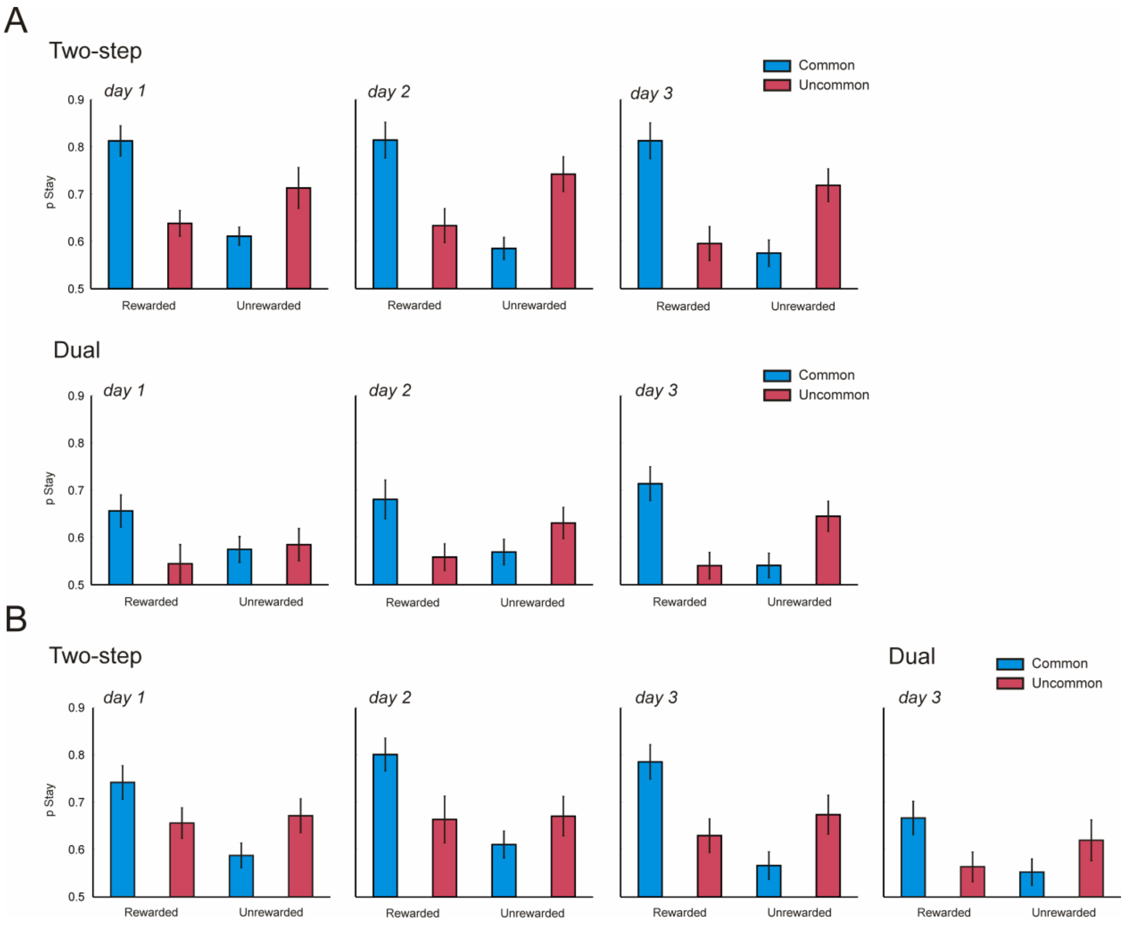


Figure 6.2 Bars plots show the average probability with which subjects chose to repeat their first-stage action on the subsequent trial as a function of the transition (common vs. uncommon) and outcome (rewarded vs. unrewarded) on the previous trial (switch-stay choice pairs). Data are divided according to trial type (two-step vs. dual-task), training period (days 1-3) and subject group (panel A, ‘high load group’; panel B, ‘low load group’). (**A**) A heightened cognitive load during dual-task trials disrupts stay-switch behaviour associated with a model-free or model-based choice strategy given no prior training on the two-step task (day 1). This deficit is largely recovered following training (day 3). (**B**) Training on the two-step task alone (i.e. without dual-task trials) permits a degree of model-based reasoning under load (dual-task condition). Errors bars represent SEM.

For completeness, I repeated the logistic regression concatenating data across all 3 days and including a regressor for the day of training (and all possible interactions; see Table 6.1, p. 160). I identified a significant 3-way reward x transition x day interaction for dual-task trials ($p < 0.05$), but not two-step trials ($p = 0.41$), consistent with the notion that training altered subjects' performance under high but not low cognitive load (see Table 6.1, p. 160).

Finally, I performed a separate logistic regression for dual-task trials where I included regressors relating to Stroop task performance on the previous trial (see Table 6.2, p. 161, for a full list of regressors). Here, I again found a 3-way reward x transition x day interaction, reflective of an increase in the influence of a model-based strategy across days, but this effect only trended towards significance ($p = 0.07$), likely owing to an increase in correlation between regressors. Interestingly, I found a main effect of Stroop performance ($p < 0.05$), indicating that subjects were more likely to repeat their first-stage choice on the next trial if they performed the Stroop task correctly on the current trial. In addition, I found a Stroop performance x reward x transition interaction ($p < 0.01$), indicating that subjects were less likely to switch their first-stage choice on the next trial following a rewarded uncommon transition on the current trial if they also made a Stroop error on the current trial. Thus, negative feedback on the Stroop task interfered with subjects' ability to utilize a model-based strategy on the next trial. One possible explanation is that when subjects make a Stroop error, they allocate more attentional resources to the Stroop task cues at the start of the following trial, hindering choice on the two-step task. Another possibility is that errors on the Stroop task disrupt credit assignment (the appropriate updating of action values) on the current trial such that subjects make a less optimal choice on the following trial.

Regressor	*Two-step*						*Dual*		
	High load group			*Low load group*					
	Coef	*SE*	*p*	*Coef*	*SE*	*p*	*Coef*	*SE*	*p*
intercept	-1.123	0.199	**< 0.001**	-0.828	0.167	**< 0.001**	-0.397	0.163	**0.0235**
rew	-0.179	0.033	**< 0.001**	-0.259	0.052	**0.0001**	-0.058	0.030	0.0691
trans	-0.139	0.084	0.1136	0.062	0.081	0.4510	-0.084	0.073	0.2622
day	0.056	0.066	0.406	-0.055	0.054	0.3219	-0.057	0.055	0.3124
rew x trans	-0.540	0.100	**< 0.001**	-0.300	0.102	**0.0081**	-0.278	0.052	**< 0.001**
rew x day	0.018	0.035	0.6105	-0.085	0.048	0.0911	-0.024	0.036	0.5164
trans x day	0.006	0.035	0.876	-0.062	0.050	0.2234	-0.002	0.036	0.9676
rew x trans x day	-0.019	0.037	0.606	-0.023	0.051	0.6599	-0.110	0.052	**0.0480**

Table 6.1 Table shows the group-level output of a logistic regression on first-stage switch-stay behaviour, separately for two-step ('high load group' and 'low load group') and dual-task trials, from data concatenated across all 3 training sessions. Note that 'reward x day' was orthogonalized with respect to reward, and in turn 'reward x transition x day' was orthogonalized with respect to 'reward x transition'. These regressors thus account for variance unexplained by the 2-way effect (see Methods, p. 145 - 146). Bold-face denotes p < 0.05 uncorrected for multiple comparisons. *rew = reward*; *trans = transition*; *Coef = regression coefficient*.

Regressor	*Dual*		
	Coef	*SE*	*p*
intercept	-0.263	0.352	0.135
rew	-0.084	0.126	**0.020**
trans	-0.109	0.126	**0.012**
day	-0.057	0.157	0.368
stroop acc	-0.091	0.179	**0.026**
rew x trans	-0.249	0.127	**< 0.001**
rew x day	-0.025	0.157	0.512
trans x day	0.004	0.156	0.926
rew x trans x day	-0.105	0.157	0.075
rew x stroop acc	0.039	0.193	0.364
rew x trans x stroop acc	-0.107	0.191	**0.006**
rew x trans x stroop acc x day	-0.047	0.231	0.420

Table 6.2 Table shows the group-level output of a logistic regression on first-stage switch-stay behaviour, for 'high load group' data and dual-task trials only, where data is concatenated across all 3 training sessions. Note that 'reward x day' was orthogonalized with respect to reward, and in turn 'reward x transition x day' was orthogonalized with respect to 'reward x transition'. Similarly, 'reward x transition x stroop accuracy' was orthogonalized with respect to 'rew x stroop accuracy' and so on. These regressors thus account for variance unexplained by the simpler 2, 3 or 4-way effect (see Methods, p. 145 - 146). Bold-face denotes p < 0.05 uncorrected for multiple comparisons. *rew = reward; trans = transition; stroop acc = Stroop task accuracy; Coef = regression coefficient.*

Computational modelling

One limitation of a switch-stay regression analysis is that it attempts to explain choice on the current trial by events occurring on the immediately preceding trial. Thus, 'switching' after an uncommon rewarded trial is always deemed model-based, whilst 'staying' is deemed model-free. However, a model-based player might repeat a first-stage choice after an uncommon rewarded trial if the expected utility of the previous first-stage action is high. Reinforcement learning models typically account for this by assuming a decaying influence of all previous trials. I therefore used computational modelling to corroborate the switch-stay regression analysis.

First, I used Bayesian model comparison to validate that choice in the two-step task reflects a hybrid of both model-free and model-based valuations (Daw et al., 2011). The hybrid model makes choices according to a weighted combination of model-free and model-based action values (with the weight governed by the parameter ***w***), where choice is purely model-free or model-based when ***w*** = 0 or 1 respectively. I then fit the hybrid model to the 'high load group' and the 'low load group', separately for two-step and dual-task trials, treating each day of training as a discrete set of data. This allowed me to assess the impact of training on the two-step task, by comparing the value of ***w*** under load on day 1 in the 'high load group', and day 3 in the 'low load group' (a between-group comparison). Specifically, this provided a conservative test of my hypothesis that training on the two-step task alone would permit increased model-based choice under load. I also evaluated the effects of training under load over time, by comparing the value of ***w*** during step-task and dual-task trials in the 'high load group' across each day of training (a within-group comparison). Finally, in order to validate my findings within a fully Bayesian framework, I performed a second model comparison where I concatenated data from the full training period and tested model variants in which the value of ***w*** could change across days.

Between-group comparison

I fit a hybrid RL model, in addition to reduced (nested) versions that describe pure model-free and model-based choice respectively, to ‘high load group’ data from day 1 of training, and to ‘low load group’ data from day 3 of training, separately for two-step (low load) and dual-task (high load) trials. Using Bayesian model comparison, I found that the hybrid model provided a better fit to subject data in both groups and both trial types, as indicated by a lower iBIC score (see Table 6.3, below). Importantly however, ***w*** was significantly higher in the two-step condition compared to the dual-task condition ($p < 0.01$) in the ‘high load group’ on day 1, consistent with previous evidence that model-based reasoning is impaired under high cognitive load in untrained subjects (Otto, Gershman, et al., 2013) (Figure 6.3A, p. 164). Conversely, I found no difference in the value of ***w*** between two-step and dual-task trials when fitting ‘low load group’ data from day 3 of training (Figure 6.3B, p. 164). This suggests that prior training on the two-step task permitted a strong degree of model-based reasoning under load, despite subjects having no prior experience with performing a task under load.

Models	*iBIC Two-step (x 10^4)*		*iBIC dual (x 10^3)*		*No. Parameters*
	High load group (day 1)	*Low load group (day 3)*	*High load group (day 1)*	*Low load group (day 3)*	
α β ε (model-free)	1.3831	1.4605	7.2260	7.8552	3
α β ε (model-based)	1.3688	1.4589	7.3056	7.8471	3
α β ε w (hybrid)	**1.3491**	**1.4195**	**7.1863**	**7.8143**	4

Table 6.3 Results of a Bayesian model comparison that accounts for differences in model complexity. The hybrid model, which incorporates influences from both model-free and model-based control, fit subject data better than pure model-free and model-based RL algorithms across both trial types (two-step versus dual-task) and both groups (‘high load group’ day1, ‘low load group’ day 3). Bold-face denotes the winning model (lowest iBIC score) for each condition.

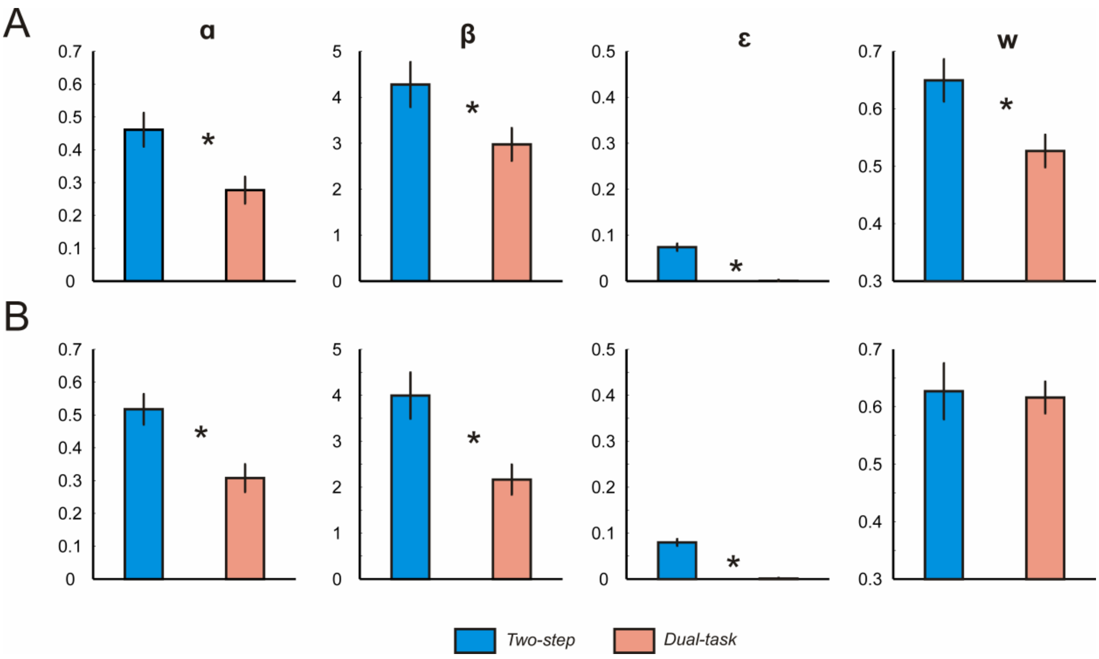


Figure 6.3 (**A**) Mean best-fitting parameters from the hybrid model for 'high load group' data from day 1 (i.e. before task training). The weighting parameter ***w***, which represents a measure of model-based (***w*** = 1) relative to model-free (***w*** = 0) control, was lower in the dual-task (high load) condition compared to the two-step (low load) condition. (**B**) By contrast, ***w*** did not differ between dual-task and two-step trials in the 'low load group' on day 3, who had received prior training on the two-step task over two consecutive days. Vertical bars represent SEM. * denotes p < 0.05.

Within-group comparison

Next, I fit the hybrid model to data from days 2 and 3 of training in the 'high load group', separately for two-step and dual-task trials. This allowed me to characterize the temporal dynamics of a shift in the balance of model-free or model-based control with increasing task exposure (i.e. across all 3 days). Here, I was principally interested in whether I would observe an abrupt switch in subjects' strategy at the start of a given training day, or alternatively, whether a gradual shift in behavioural control would emerge across days (as suggested by subjects' switch-stay choices; see Figure 6.2, p. 158). To test these hypotheses, I performed paired t-tests on parameter estimates from Bayesian model inference. In the two-step task, I found evidence for a moderate shift towards more model-based choice, as indexed by higher ***w*** values, on days 2 and 3 of training compared to day 1 (both p < 0.01) (Figure 6.4A, below). During dual-task trials, I found a more pronounced shift towards model-based choice, with an approximately linear increase in the value of ***w*** across days (all p < 0.001) (Figure 6.4B, below). Thus, consistent with the regression analysis, training increased the relative contribution of model-based reasoning to a greater degree during high load trials, suggesting that the addition of load is necessary to expose training-induced changes in behaviour in the two-step task.

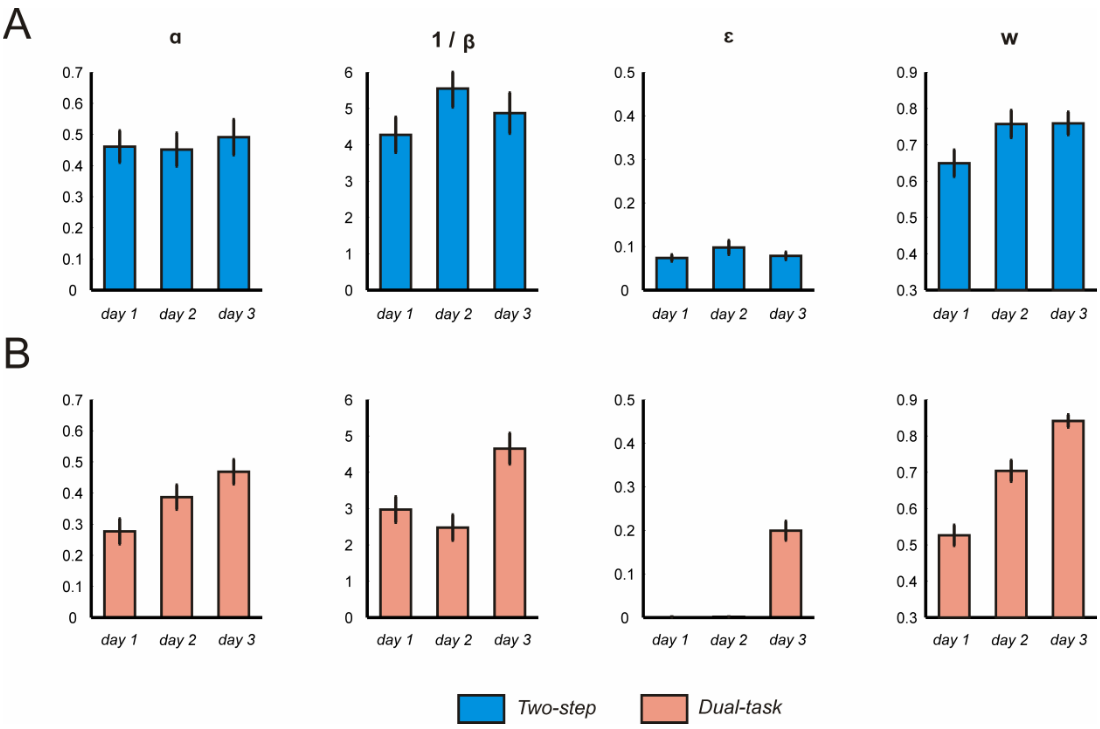


Figure 6.4 Mean best-fitting parameters from the hybrid model for 'high load group' data from days 1-3 of training. The weighting parameter ***w*** represents a measure of model-based (***w*** = 1) relative to model-free (***w*** = 0) control. (**A**) At the group level, model parameters remained relatively stable across two-step trials, indicating that performance in the absence of load was not largely affected by training. (**B**) By contrast, performance under load shifted across days, with higher learning rates and higher ***w*** values as task exposure increased. Vertical bars represent SEM.

Multi-day model comparison

To corroborate the finding that ***w*** changes with training within a fully Bayesian framework, I fit a full hybrid RL model (in addition to various nested alternatives) to 'high load group' data across all 3 days (combined), separately for two-step and dual-task trials. Importantly, I tested model variants in which ***w*** could shift across days, governed by a slope parameter ***σ***, allowing the balance between model-free and model-based control to vary with each consecutive day of training. Bayesian model comparison revealed that the slope parameter ***σ*** was supported for the dual-task condition but not for the two-step condition, with the latter result replicating in both the 'high load group' and the 'low load group' (see Tables 6.4 & 6.5, p. 167). Thus, training influenced the balance between model-free and model-based control across each day of training in dual-task trials but not in two-step trials (however, ***w*** was higher on days 2 and 3 compared to day 1 of training during two-step blocks, a subtlety not captured by the slope model). Importantly, the value of ***σ*** was negative at the group-level, indicating a higher degree of model-based control on day 3 compared to day 1 (see Table 6.5, p. 167). Thus, subjects' ability to perform model-based reasoning gradually became immune to cognitive load when training included both the two-step and dual-task conditions, both within a fully Bayesian framework, and when fitting behaviour from each day individually.

Models	*iBIC Two-step (x 10^4)*		*iBIC dual (x 10^4)*	*No. Parameters*
	High load group	*Low load group*		
$\alpha\ \beta\ \varepsilon\ w$	3.9598	4.2879	2.1602	4
$\alpha\ \beta\ \varepsilon\ w\ \lambda$	3.9544	4.2880	2.1598	5
$\alpha\ \beta\ \varepsilon\ w\ \sigma$	3.9703	4.2917	2.1598	5
$\alpha\ \beta\ \varepsilon\ w\ \lambda\ \sigma$	3.9563	4.2912	2.1608	6
$\alpha^1\ \alpha^2\ \beta\ \varepsilon\ w$	3.9592	4.2877	2.1498	5
$\alpha^1\ \alpha^2\ \beta\ \varepsilon\ w\ \lambda$	3.9494	4.2887	2.1507	6
$\alpha^1\ \alpha^2\ \beta\ \varepsilon\ w\ \sigma$	3.9612	4.2907	2.1490	6
$\alpha^1\ \alpha^2\ \beta\ \varepsilon\ w\ \lambda\ \sigma$	3.9459	4.2906	2.1494	7
$\alpha\ \beta^1\ \beta^2\ \varepsilon\ w$	3.9153	4.2556	2.1494	5
$\alpha\ \beta^1\ \beta^2\ \varepsilon\ w\ \lambda$	*3.9072*	**4.2494**	2.1472	6
$\alpha\ \beta^1\ \beta^2\ \varepsilon\ w\ \sigma$	3.9214	4.2612	2.1476	6
$\alpha\ \beta^1\ \beta^2\ \varepsilon\ w\ \lambda\ \sigma$	3.9120	4.2541	2.1476	7
$\alpha^1\ \alpha^2\ \beta^1\ \beta^2\ \varepsilon\ w$	3.9134	4.2586	2.1449	6
$\alpha^1\ \alpha^2\ \beta^1\ \beta^2\ \varepsilon\ w\ \sigma$	3.9196	4.2632	**2.1433**	7
$\alpha^1\ \alpha^2\ \beta^1\ \beta^2\ \varepsilon\ w\ \lambda$	**3.9055**	*4.2501*	*2.1442*	7
$\alpha^1\ \alpha^2\ \beta^1\ \beta^2\ \varepsilon\ w\ \lambda\ \sigma$	3.9111	4.2569	2.1462	8

Table 6.4 Results of a Bayesian model comparison that accounts for differences in model complexity. More complex model variants include those that have separate RL parameters for first and second stage choices, eligibility traces, and a parameter for capturing shifts in model-free versus model-based control across days. Bold-face denotes the winning model (lowest iBIC score) for each condition.

Condition	α^1	α^2	β^1	β^2	ε	w	λ	σ
High load group: two-step	0.442	0.405	6.682	2.734	4.97 x10^{-5}	0.723	0.546	0^*
High load group: dual-task	0.097	0.355	3.981	2.250	0.021	0.899	1^*	-0.306
Low load group: two-step	0.510		4.752	2.560	6.66 x 10^{-5}	0.566	0.628	0^*

Table 6.5 Best-fitting parameter estimates shown separately for each group and condition (two-step versus dual-task), using data concatenated across all 3 days of training. Values represent mean parameter fits across all subjects. * represents fixed parameter values.

Other learning parameters

In addition to differences in the value of ***w*** between two-step and dual-task trials, I also found differences in a number of other learning parameters. When fitting data from the 'high load group' on day 1, and the 'low load group' on day 3, I found subjects were less considerate of the most recent reward information (as indexed by a lower learning rate) and chose more stochastically (as indicated by a lower inverse temperature) during dual-task trials compared to two-step trials (all $p < 0.05$) (Figure 6.3, p. 164). I identified similar differences when fitting data across all training days consecutively (Table 6.5, p. 167). However, when subjects were able to practice the dual-task condition on each day ('high load group'), both the learning rate and inverse temperature under load increased across days (all $p < 0.001$ comparing day 1 to day 3) (Figure 6.4B, p. 165).

3-Back Logistic Regression

Computational modelling characterizes subject behaviour in the two-step task by integrating over a history of choices. In this modelling, I quantified the relative degree of model-free and model-based control by fitting ***w*** to subjects' choices. However, this approach also relies on fitting several other model parameters that may exhibit a degree of shared variance. This has the potential to complicate interpretation when the true value of more than one parameter differs between two conditions.

I therefore employed a second logistic regression, to capture the main effects of the model, in a manner that more closely approximates a modelling approach. Here, rather than consider pairs of isolated choices, I quantified the degree to which choice on the current trial reflected a model-free and model-based influence relative to events occurring on the preceding 3 trials (see Methods, *Logistic regression*, p. 146) (Smittenaar et al., 2014). For example, if a player received a reward following an uncommon transition 3 trials in the past, a model-free system would be more likely to choose the same first-stage fractal on the current trial, whereas a model-based system would be more likely to choose the opposite first-stage fractal on the current trial.

During two-step trials, I identified a significant model-free and model-based influence on choice extending up to 3 trials in the past (all $p < 0.05$), consistent with a notion that subjects utilized a hybrid of both systems (Figure 6.5A, p. 170). However, I found a reduction in model-based control in the dual-task condition compared to the two-step condition in the 'high load group' on day 1, an effect that spread up to 2 trials in the past (pared t-tests, all $p < 0.05$). Importantly, this difference was reduced following task training (on day 3), independent of whether training included ('high load group', Figure 6.5A, p. 170) or excluded ('low load group', Figure 6.6, p. 171) the high load condition. To help visualize these effects, I derived single indices of model-free and model-based learning by summing the coefficients that correspond to an influence of events on 1, 2 or 3 trials in the past respectively (see Figure 6.5B, p. 170). Further, to my surprise, I was unable to identify a significant model-free influence in either group in the high load condition. However, model-free coefficients were not significantly different when comparing the two-step and dual-task conditions (with paired t-tests). Thus, I do not wish to draw strong inferences from this subtle dissimilarity. In summary, these results replicate my computational modelling in a format that is slightly less powerful but is also freer from parametric assumptions.

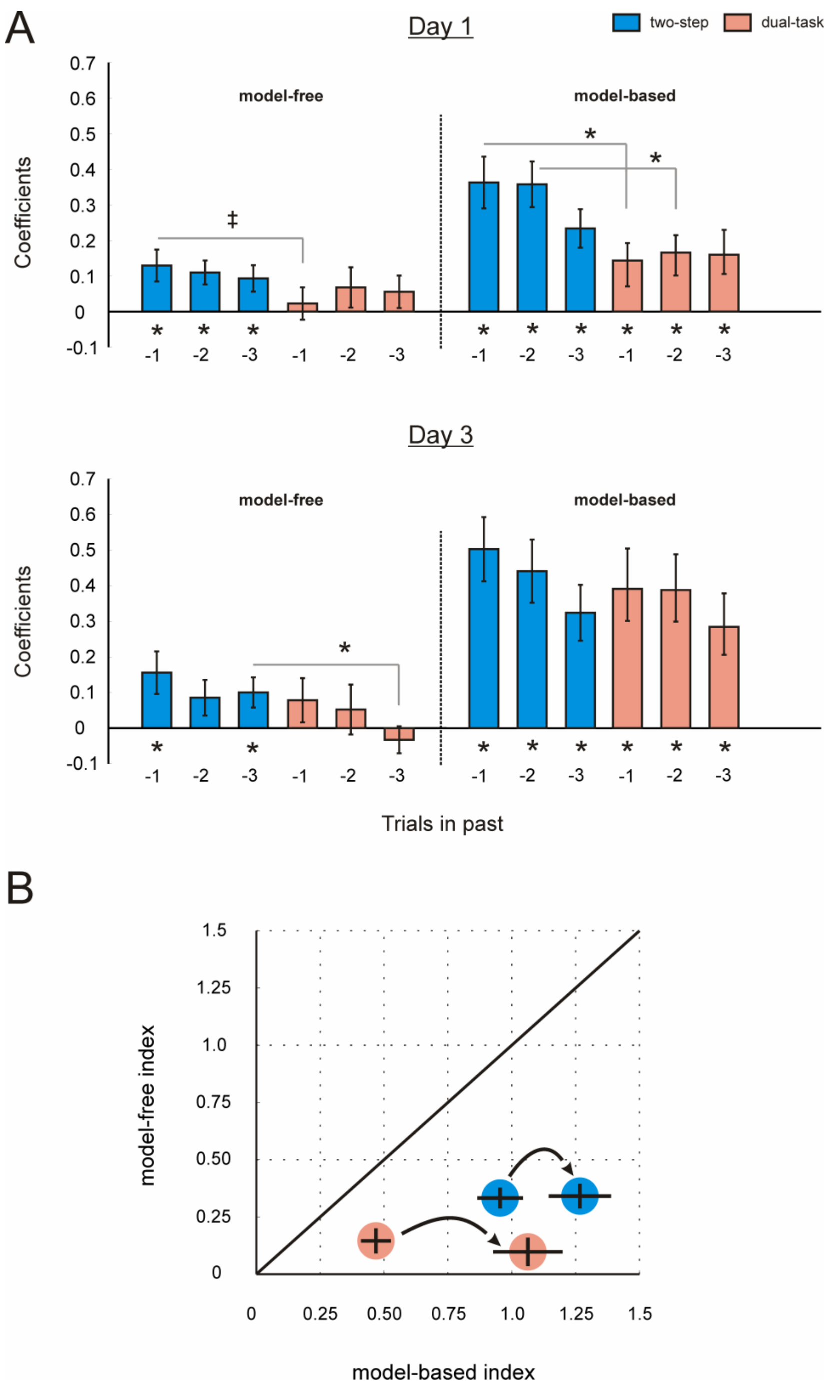


Figure 6.5 I performed a logistic regression to estimate the relationship between choice on trial *t* and events occurring on trial t_{-1} up to t_{-3}. Here, regression coefficients can be interpreted as reflecting a model-free or model-based influence on choice, where larger coefficients indicate a stronger influence. (**A**) In the two-step condition, model-free and model-based coefficients were significantly different from 0 (up to 3 trials in the past), suggesting that subjects used a hybrid of both strategies (upper panel). In the dual-task condition, I observed no significant influence of a model-free system, and a diminished influence of a model-based system in untrained subjects. This impairment in model-based

reasoning was suppressed following task training (day 3), while the absence of a model-free influence remained insensitive to training (lower panel). (**B**) For each condition, and separately for days 1 and 3, I summed (individually) the coefficients corresponding to trial t_{-1}, t_{-2} and t_{-3}, and derived single estimates of the degree to which model-free and model-based control were dominant in choice. I observed a larger relative shift towards model-based control with training in the dual-task condition compared to the two-step condition, consistent with the previously discussed analyses. Vertical lines represent SEM. * denotes p < 0.05, ‡ denotes p = 0.09.

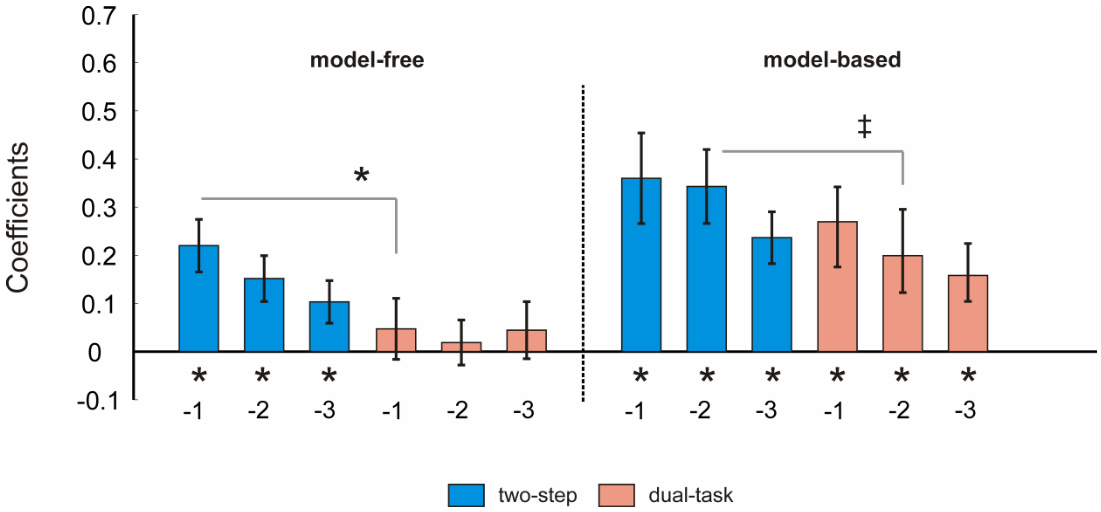


Figure 6.6 I performed a logistic regression on data from 'low load group' and day 3 of training to estimate the relationship between choice on trial *t* and events occurring on trial t_{-1} up to t_{-3}. Here, regression coefficients can be interpreted as reflecting a model-free or model-based influence on choice, where larger coefficients indicate a stronger influence. In the two-step condition (blue bars), model-free and model-based coefficients were significantly different from 0 (up to 3 trials in the past), suggesting that subjects used a hybrid of both strategies. In the dual-task (high load) condition (orange bars), I observed a significant influence of a model-based system, that did not differ from the two-step condition, up to 3 trials in the past. In contrast, I found no significant influence of a model-free system. These results are consistent with data from the 'high load group' (see Figure 6.5, p. 170). Vertical lines represent SEM. * denotes p =< 0.05, ‡ denotes p = 0.08.

Numerical Stroop performance

Mean numerical Stroop accuracy during dual-task trials was 81.9% on day 1, 85.5% on day 2, and 89.5% on day 3 for the ‘high load group’. Thus, performance on the secondary task demonstrated an approximately linear increase across training days (all $p < 0.05$). Mean numerical Stroop accuracy for the ‘low load group’, in which subjects only experienced the dual-task condition on day 3 of training, was 83.2%, and thus comparable to the ‘high load group’.

Despite performance on the Stroop task being high overall (~80 - 90%), I hypothesized that the ability for subjects to respond accurately would depend on, or interact with, events occurring on the two-step task. To explore this I performed a logistic regression, using data from the ‘high load group’, to quantify whether Stroop task performance on the current trial (coded as ‘0’ for incorrect and ‘1’ for correct) could be explained by choice or events related to the concurrent two-step task. The explanatory factors in this model included whether the numbers presented on the current trial were congruent or incongruent with respect to ‘SIZE’ and ‘VALUE’, whether subjects choose to repeat or switch their first-stage choice with respect to the previous trial, the response time at the first-stage, whether the transition on the current trial was common or uncommon, whether the current trial was rewarded or not, and whether the data were from day 1, 2 or 3 of training. The resulting coefficients are presented in Table 6.6 (p. 173).

Regressors	*High load group*		
	Coef	*SE*	*p*
intercept	1.5854	0.3084	**< 0.001**
switch/stay	-0.1122	0.0958	0.255
rew	-0.0094	0.1068	0.931
trans	0.1592	0.0831	0.069
congruency	-0.0613	0.1991	0.761
RT	-0.0004	0.0002	**0.020**
day	0.3377	0.0766	**< 0.001**

Table 6.6 Group-level output of a logistic regression on the probability of performing the Stroop task correctly. Bold-face denotes p < 0.05 uncorrected for multiple comparisons. *rew = reward*; *trans = transition*; *RT = first-stage reaction time*; *Coef = regression coefficient*.

This regression analysis revealed that subjects were more likely to perform the Stroop task correctly with increasing task experience (a main effect of day). Performance was also more accurate when they were faster to make their first-stage choice on the two-step task. This could imply that deploying more cognitive resources to the two-step task (or for a longer period of time) - owing to reduced decision confidence - might negatively impact on the encoding or maintenance of distracting information. Alternatively, subjects may have generally been more aroused and attentive on trials where they were faster to respond on the two-step task. Further, I identified a main effect of transition that trended towards significance (p = 0.07). This suggests that subjects may have been less likely to perform the Stroop task correctly on trials where they experienced an uncommon transition on the two-step task. The Stroop task requires maintaining spatial information about numerical cues in working memory. Thus, one possibility is that the experience of an uncommon transition interfered with subjects' ability to correctly maintain location-specific (left versus right) information required for the Stroop task, perhaps because it increased the likelihood of mentally switching these locations. However, subjects had no explicit reason to map common or uncommon transitions onto specific spatial locations since they did not transition through the Markov structure "spatially" when performing the task. A more likely explanation is that uncommon transitions utilize more cognitive resources than common transitions. For example, it is feasible that subjects prepare a second-stage response following their first-stage choice based on the fractal pair that would result from a common transition. In the event of an uncommon transition, subjects might have to deploy additional cognitive resources to retrieve and compare cached values at the second stage on the fly.

6.4 Discussion

A prominent account of decision-making posits that humans and other animals use (at least) two distinct strategies for making choices (Balleine & O'Doherty, 2010; Daw et al., 2005; Dolan & Dayan, 2013; van der Meer, Kurth-Nelson, & Redish, 2012). In this view, a habitual or model-free decision system resides primarily in the basal ganglia and does not depend on limited executive resources (Stalnaker et al., 2010; Yin & Knowlton, 2006). Meanwhile, a goal-directed or model-based decision system engages prefrontal cortical areas with capacity limits that might arise from serial processing (Alvarez & Emory, 2006; Glascher et al., 2010; Jones et al., 2012; Sigman & Dehaene, 2008; Smittenaar et al., 2013; Wunderlich, Dayan, et al., 2012).

Here, I asked whether reliance on a limited pool of executive resources is a universal property of model-based choice, or whether, as task familiarity increases, the brain can execute model-reasoning in a way that depends less on these limited-resource functions. I found that model-based choice was preserved under load in subjects that had acquired familiarity, through prior training, with the structure of a two-stage Markov decision task (Daw et al., 2011), suggesting a change in the implementation of model-based behaviour. This finding was independently replicated in two cohorts of subjects (who received training either with or without load) and using different methodological approaches.

There are several possible explanations for this result. First, following training, subjects may become more efficient at using the task structure to plan ahead. From a neural perspective, this might entail shifting model calculations away from executive brain areas that are limited by serial processing (Dux, Ivanoff, Asplund, & Marois, 2006; Glascher et al., 2010; Sigman & Dehaene, 2008; Smittenaar et al., 2013; Valentin et al., 2007; Wunderlich, Dayan, et al., 2012). Interestingly, task training has previously been shown to cause "off-loading" to other neural circuits in tasks requiring executive resources, including a shift from the prefrontal cortex to parietal and striatal regions (Kelly & Garavan, 2005; Yildiz & Beste, 2014). In particular, recent evidence suggests the striatum, in contrast to the prefrontal cortex, may be more optimized for parallel processing in dual-task conditions (Yildiz & Beste, 2014). If model calculations remain in the same brain areas, it is possible that the coding within these areas becomes more efficient with experience, for example, if only a fraction of neurons are required to fire in order to achieve the same representational fidelity (Beauchamp, Dagher, Aston, & Doyon, 2003; Bush et al., 1998; Poldrack, 2000).

Second, resilience to load could emerge with training if other implicitly necessary processes (other than reasoning with the Markov transition matrix of the task itself) become more efficient. For example, some cognitive resources may be required for identifying the various fractals and remembering which is which, for maintaining events that occurred on the previous trial in working memory, and for retrieving cached values at the second-stage during planning. There may also be resources involved in maintaining belief distributions over meta-parameters, such as whether the task structure might change or new fractals might appear, what appropriate learning rates are, when model-based reasoning should be deployed (Daw et al., 2005; Keramati, Dezfouli, & Piray, 2011) and how attentional resources should be allocated within a trial. Since all these processes are likely dependent on executive brain regions to some degree (Badre, 2008; Behrens et al., 2007; Knight et al., 1995; E. E. Smith & Jonides, 1999; Waskom et al., 2014), gaining efficiency in any of these domains may free resources for model-based computations, or for maintaining task-relevant content in working memory.

Third, subjects could learn to perform model-based calculations at the end of each trial (i.e., "offline"), rather than the beginning of the next trial. The results of this calculation could update a cached or habitual value accessed for the next choice, relieving the need to store the current reward in memory until the beginning of the next trial. In turn this could allow increased allocation of executive resources to the concurrent task. Further, choice under load after training may not be truly model-based. Increasingly sophisticated choice heuristics (for example, applying Q-value updates to the opposite first-stage transition following an uncommon transition), permit behaviour that is increasingly difficult to distinguish from fully model-based in the simple two-step task (K. J. Miller, Erlich, Kopec, Botvinick, & Brody, 2013). Although not realizing the full Markov model of the task, these strategies implicitly embody partial models of the task structure.

While my data cannot currently disambiguate between these divergent mechanisms (many of which are at least to some degree overlapping and by no means mutually exclusive), it seems likely that task training could instigate a combination of the above. Future experiments could aim to investigate their respective predictions, for example via the use of neuroimaging. In the remainder of the discussion I will elaborate on a number of more subtle features of the data that, while not affecting the main conclusions of the paper, are nevertheless of interest.

Using computational modelling, I found that ***w*** (a parameter indexing the balance between model-based and model-free control) was reduced by load, and this deficit was eliminated by prior task training. However, a subsequent 3-back regression analysis suggests the possibility that the observed reduction in ***w*** under load could reflect a marginal weakening of model-free reasoning, in addition to a more pronounced disruption of model-based reasoning. This contrasts with previous studies showing that model-based, but not model-free learning, is prone to interference in a range of contexts (Otto, Gershman, et al., 2013; Otto, Raio, et al., 2013; Smittenaar et al., 2013; Voon et al., 2014; Wunderlich, Smittenaar, et al., 2012). One possibility is that this subtle difference may be a consequence of dissimilarities in task design. For example, while Otto and colleagues utilized interleaved trials of low and high load (Otto, Gershman, et al., 2013), I employed alternating blocks of either condition. If subjects make choices by integrating over a history of trials, then enforcing a high load over a longer history of trials could have more diffuse consequences on choice. Similarly, I found that task training boosted model-based but not model-free reasoning under load. While I do not wish to draw strong conclusions, one possibility is that subjects were encouraged to overcompensate for load during dual-task trials, and that this suppressed the influence of a model-free system following training.

Through computational modelling, I found that load affected not just ***w*** but also learning rates and choice noise parameters. These changes were not eliminated by prior task training. Slower learning rates and more stochastic choice appeared during high load trials, independent of training (in the ‘low load group’). Slower learning rates raise the possibility that subjects under load inferred lower environmental volatility (perhaps placing stronger weight on priors) (Behrens et al., 2007). Alternatively, it may reflect a tradeoff between executive processes, such as updating the contents of working memory, and more incremental learning processes that exhibit longer time-constants. Noisier choice might be associated with a reduction in decision-making confidence (De Martino et al., 2013; Kepecs, Uchida, Zariwala, & Mainen, 2008). It is also possible that the underlying choice strategy used by subjects was not fully captured by the winning model, leading some other form of variability to be absorbed by the model parameters. In addition, although prior task training had a large effect on behaviour under load, it had little effect on behaviour without load (comparing day 3 to day 1 in the 'low load group'). This suggests that training induced latent changes that were made apparent by the addition of load, for example, due to an effective ceiling on model-based choice.

I found slightly higher ***w*** values on day 3 of training in the 'high load group' (who received training both with and without load), than the 'low load group' (who only received training on the two-step task), in both trial types, indicating a higher degree of model-based relative to model-free control. This could in part be explained by the fact that subjects in the 'high load group' received more overall training (256 additional trials; 4 blocks of 64 dual-task trials) than the 'low load group'. However, it is also possible that training under load induced neural adaptations that permitted improved dual-task performance independent from a change in the implementation of model-based choice (Hazeltine, Teague, & Ivry, 2002). Finally, I cannot exclude the possibility that a component of the behavioural change I observed across days in the 'low load group' was unrelated to familiarity with the primary decision-task. For example, subjects may have simply become more comfortable in the laboratory setting after consecutive visits. Further, by merely practicing a cognitively demanding task, subjects' working memory capacity may have improved, reducing the burden associated with concurrent task performance. However, evidence that working memory training generalizes to novel tasks or contexts remains at best controversial (Melby-Lervag & Hulme, 2013).

In summary, I present data that challenge a prominent notion in decision-making that goal-directed or model-based reasoning is necessarily reliant on a finite pool of executive resources. Instead, I show this reliance is linked to the degree of prior experience with the model of the world, where more experience may enable different (and potentially less costly) neural mechanisms for the implementation of model-based choice. These data may have implications for therapies to restore normal decision-making in psychiatric disorders, where deficits in model-based reasoning are thought to play a key role.

CHAPTER 7

GENERAL CONCLUSIONS

A major focus of the work presented in this thesis has been the contribution of multiple neural systems to decision-making in an adaptive context. My first two experiments focused on the processes supporting value encoding and action selection, while the latter experiment investigated learning. In the following sections I will give an overview of the significance of my findings within the broader context of neuroeconomics and proceed to discuss the limitations of the work and relevant future directions.

7.1 Overview, limitations, and further work

Contemporary theories formalize decision-making in terms of an evaluation of the predicted rewards, punishments and costs associated with different choice options, which are then compared so that the best option can be chosen. Over the last decade, a number of influential experiments have sought to reveal the neural correlates of such valuations by combining behavioural paradigms, in which individuals choose between options that differ in value, with neuroimaging techniques such as fMRI, or single-unit recordings.

Although a wide array of brain regions have been implicated, two ubiquitous regions include the ventromedial prefrontal cortex (vmPFC) and the striatum. For example, in a set of studies by Padoa-Schioppa and colleagues (Padoa-Schioppa & Assad, 2006, 2008), activity in the orbitofrontal cortex of non-human primates correlated with the subjective value of different rewards or “goods”. Here, monkeys were made to choose between a set of two or three juice options where the types and amounts of juices varied across trials and sessions. This allowed the experimenters to infer the subjective value of different options based upon each monkey’s actual choices. Within the orbitofrontal cortex, neurons were identified that encoded the subjective value of one of the offered rewards, or alternatively the chosen reward, or the type of juice chosen. This led the authors to conclude that this region in the brain is a good candidate for the type of value assignment that is thought to underlie economic choice. Similarly, Kable and Glimcher (Kable & Glimcher, 2007) showed that comparable brain areas encode subjective value in humans. In their experiment, subjects were required to choose between an immediately available sum of money and a larger but delayed sum. Here, the subjective value of a delayed outcome typically declines as the imposed delay to its delivery increases, a phenomenon termed delay discounting. By estimating a discounting factor for each individual subject, Kable and Glimcher were able to show that the BOLD response in regions including the vmPFC and striatum correlated with the subjective value of the delayed choice option. Thus, the authors similarly concluded that these regions form a common valuation system that assigns values to choice options in the environment and ultimately guides action selection.

Many value-guided decision-making paradigms involve discrete, binary decisions between two or more choice options, where the outcome of each decision (or trial) is independent. Thus, rewards and punishments accumulate discretely over time, or alternatively, a single random trial is realized at the end of the experiment. Although this approach is highly efficient, it fails to capture several components of real-life decision-making, where choice is often sequential and highly context-dependent. For example, drinking coffee may have a high utility in the morning (due to increased alertness), but a low utility in the evening (as one’s sleep is likely to be disrupted). Similarly, actions that lead to an immediate reward or punishment may also generate a complex set of delayed consequences. For instance, smoking a cigarette may bestow immediate rewards but may also imperil long-term health and confer negative social consequences. Bearing in mind these complexities, we know less about how vmPFC and other valuations regions encode different components of value in different contexts, or how this contributes towards adaptive decision-making.

In this thesis I took the relatively novel approach of designing sequential decision-making paradigms in which the value of a given choice option can fluctuate according to changes in context. The data I present in Chapters 4 and 5 support a notion that vmPFC (and other valuation regions) represent the immediate or stimulus-driven value of choice options in such contexts. These low-level representations are then moderated by higher-order prefrontal (and possibly parietal) networks that track specific contextual or task-related computations, in a manner that allows individuals to overcome prepotent responses, maintain hierarchical goals, and adaptively switch their choices. These results are consistent with previous computational theories that decision values reflect the accumulation and integration of multiple sources of information in the brain (Ratcliff & McKoon, 2008), and with empirical evidence that dorsal prefrontal cortex incorporates more abstract decision components within vmPFC (Hare et al., 2009; McClure et al., 2004). However, it should not be overlooked that in some cases, there may be substantial inter-individual variability in the degree to which stimulus-driven valuations are modulated, and consequently, in the extent to which individuals are able to dynamically adjust their behaviour (as shown in Chapter 5). Future experiments could address whether this variability stems from an impaired representation of context, or from a deficit in in the functional integration of multiple decision components.

Importantly, while several previous studies have shown that dorsal prefrontal cortex is involved in flexible decision-making when changes in context are externally cued (Aron et al., 2004; Badre & Wagner, 2004; Kerns et al., 2004), few paradigms have explored the mechanisms underpinning switches in choice when the agent has to infer that the context has changed, a decision process important in real life. Here, I show that both instances recruit similar regions of prefrontal cortex and likely utilize equivalent choice architectures. Although I used a computational approach, it remains unclear to what extent the functional contribution of dorsal prefrontal cortex generalizes to other tasks, or how it relates to other executive functions supported by the neocortex, such as working memory (Curtis & D'Esposito, 2003). These issues are discussed more fully in Chapter 4.

Further, in Chapter 5, I showed that value representations within some brain regions (such as the striatum) remain insensitive to changes in context, and that this may lead to suboptimalities in decision-making. In this framework, choice might instigate a competition between value systems supporting short-term gains and long-term goals respectively. This dual-system framework is analogous, albeit computationally distinct, to other dual-system theories in the brain, such as the increasingly evident division between model-free and model-based reinforcement learning (Daw et al., 2011; Dolan & Dayan, 2013; Doll et al., 2012). Further work is needed to understand, i) how these “short-term” reward representations in striatum manifest in choice, ii) whether they are in fact ‘functional’ as opposed to merely ‘content’ representations (deCharms & Zador, 2000), iii) whether they are exaggerated in impulsive individuals, iv) whether they are enhanced by stress or fatigue, and v) whether they can be suppressed through training. Further, we currently know little about how the brain resolves competition between systems supporting short-term and long-term goals respectively, or put differently, why one system prevails on some trials but not others.

Finally, in Chapter 6 I asked whether the mechanism by which the brain uses feedback to learn how to make optimal decisions in a given environment changes with experience. Laboratory-based experiments that probe learning and planning often utilize tasks in which the underlying structure -that is, the relationship between actions and their future trajectories - are entirely novel to the subject. Thus, the neural implementation of learning in this context may represent a special case that is less commonly encountered in everyday life. To address this, I trained subjects on a value-guided decision-making task for 3 consecutive days. The data were suggestive of a shift in the implementation of value-guided planning with training, from a more cumbersome, resource-dependant mechanism, to a more efficient and robust process that remained resistant to attentional load.

While very few previous studies have explored the effects of task-experience on value-guided choice, I believe the data speak as an important reminder that the brain is a dynamic machine, and that by reporting averaged data we are merely taking a snapshot of the underlying cognitive processes and could thus be overlooking important subtleties. Naturally, a vital follow-up will be to use neuroimaging techniques to identify the underlying changes in neural representations that occur with training. In addition, given these data, it may be important to determine whether equivalent changes in the neural architecture supporting value-guided choice occur outside of learning tasks. Even in experiments not designed to probe the effects of training, one could contrast fMRI activation maps obtained from the first and last session of a task (albeit at more liberal thresholds), and test a null hypothesis that no differences would be observed in a more explorative manner. Finally, a highly pertinent follow-up would be to determine whether similar training effects would generalize to other planning (model-based) tasks, as this may have implications for enhancing model-based reasoning in psychiatric disorders.

7.2 Challenges in fMRI

Although fMRI has become a widely used tool for understanding the association between brain activity and cognition, it is not without limitations. Whilst providing a detailed account goes beyond the scope of this thesis (for an excellent review see (Constable, 2012)), I outline some of the major and more general challenges.

Firstly, fMRI measures changes in cerebral blood flow (the BOLD signal) and is thus merely a correlative measure of neuronal activity. Here, the change in the MR signal from neuronal activity is called the hemodynamic response (HR) and lags the neuronal event triggering it by 1-2 seconds, typically rising to a peak 5-6 seconds after event onset, and dropping back to baseline some 16-20 seconds later. During fMRI analysis, task-relevant cues are convolved with an HR function so as to capture the time-course of the BOLD response in the brain. Action potentials on the other hand, fire over the course of milliseconds. Since the BOLD response cannot capture detailed aspects of evoked responses or specific spike timings, it has poor temporal resolution. The relationship between neural activity and BOLD response is also a complex one. It is typically assumed that an increase or decrease in BOLD signal stems from an equivalent increase or decrease in the spiking of many task or stimulus-specific neurons. While this may be true in many cases, it is also important to consider that cortical microcircuits consist of multiple interacting neuronal populations, each with a specific set of excitatory and inhibitory connections (Douglas & Martin, 2004). Thus, activation of a microcircuit sets in motion a sequence of excitation and inhibition in every neuron of the module, and the proportional changes in this excitation-inhibition will likely impact the haemodynamic response. In some cases, this can lead to an increase in BOLD response without a net excitatory increase in task-related cortical output (Logothetis, 2008).

Further, as previously mentioned, fMRI is susceptible to multiple sources of unwanted noise. Consequently, fMRI studies require multiple repetitions of the same events to improve the signal-to-noise ratio. Common sources of noise include the scanner, random brain activity, and large blood vessels where blood flow is often highly variable due to factors of no interest. Other physiological sources include signal changes as a function of both the cardiac cycle and respiration pattern. Lastly, head movement by the subject is an invariable problem in fMRI experiments. Although one can attempt to rectify this using spatial realignment algorithms (see Chapter 3, *Pre-processing*, p. 65), improved motion-correction methods and motion-limiting devices are needed as the field moves towards higher spatial resolution imaging.

In addition to the practical limitations of fMRI, it is important to also consider interpretational difficulties. A common phenomenon from electrophysiological studies that measure the activity of single neurons in non-human primates is that of opponent encoding schemes. Thought to be a fundamental feature of decision-making networks, opponent encoding describes the phenomenon that individual neurons within a given population frequently encode the same neural computation with opposing signs. For example, in the context of value-guided decision-making, a neuron in the orbitofrontal cortex (OFC) may increase its firing rate as the probability of reward goes up, whilst a neighbouring neuron may decrease its firing rate under the same conditions (Kennerley et al., 2009). Thus, averaging across the activity of these neurons within a given brain region may average away meaningful signals. A related issue is that spatially adjacent neurons may track distinct computations. Returning to the example of value-guided decision-making, neighbouring neurons in OFC may track the probability or the magnitude of an expected reward respectively. fMRI relies on voxel-based analyses, and typically hundreds of thousands of neurons are included in a single voxel. Thus, different groups of neurons may be activated by different tasks within a single voxel, making it difficult to distinguish different functional roles.

Related to several of these concepts, significant attention has recently been given towards the notion that many neuroimaging experiments may be under-powered. Low statistical power not only reduces the chances of detecting true effects but also increase the chances of finding statistically significant effects that are in fact false positives. Recently, Button and colleagues used meta-analytic studies to estimate the 'true' power of a given effect, and then prospectively calculated the power of each individual study in the meta-analysis based on the associated sample size (Button et al., 2013). The authors reported that the mean statistical power across 461 neuroimaging studies (from 41 separate meta-analyses published from 2006-2009) was as low as 8%. While other researchers claim this statistic may be inflated, it nevertheless highlights the need for increasingly efficient designs, sufficient numbers of subjects, a shift from univariate to multivariate methods, and (where possible) paradigms that allow for replications.

7.3 Challenges in computational modelling

Classical behavioural analyses rely on averaging data over a number of trials to achieve a reasonable statistical power with which to quantify a given metric of human performance, such as a reaction time, a preference for action A over action B, or the number of correctly executed choices. However, it is evident that the complexities of the human mind cannot be understood strictly though observing human behaviour. Computational models on the other hand, seek to address how human performance comes about. That is, they represent a description of the underlying representations, mechanisms and processes that result in cognition. By considering how (latent) variables in the environment influence behavioural responding on a trial-by-trial basis, computational models offer a much more fine-grained explanation of the processes underlying flexible decision-making.

The use of model-based analyses is becoming increasingly popular in fMRI studies of value-based learning and decision-making. Here, the goal is to capture key aspects of behaviour with a computational model, and then to investigate whether different components of the model are realized in the brain. Unlike subtraction analyses which merely report increases or decreases in the BOLD signal in one condition with respect to another, model-based fMRI analyses allow attribution of specific computational processes to the underlying neural activations. For example, this approach has been used to show that the ventral striatum tracks prediction errors associated with temporal difference reinforcement learning (Rutledge, Dean, Caplin, & Glimcher, 2010), that the anterior cingulate cortex tracks the volatility of the environment during foraging (Behrens et al., 2007), and that the dorsolateral prefrontal cortex supports model-based reasoning by encoding the underlying task structure (Glascher et al., 2010). Further, rather than exclusively trying to reason backwards from behaviour to the underlying processes, modelling affords the opportunity to simulate large amounts of data, to adjust small components of the model, and to observe how these changes influence the output of the model. This process can be extremely useful for generating new hypotheses, and for providing novel insights into the interpretation of real experimental data (Sun, 2008).

Computational modelling faces a number of important challenges and limitations. For example, models often encompass a number of free parameters - variables that have to be set a certain value in order for the model to make practical predictions. Since it is assumed that these values can differ between individuals, parameters are often fit to each individuals' choices (e.g. via maximum likelihood estimation), with the resulting parameters then being used to predict the neural data. Of course, this relies on a critical assumption that the behavioural and neural data are fitted accurately by the same parameters, which may not necessarily hold true. Further, many of the optimization (parameter-estimation) procedures (such as the Simplex method) are susceptible to converging on local minima that do not represent the true global minimum. This typically occurs when the error surface is not smooth but rather contains dimples, valleys or plateaus, compromising any meaningful interpretation of parameter values and obscuring the true power of the model (Lewandowsky & Farrell, 2011). This latter problem can be somewhat alleviated by repeating the fitting procedure using multiple random starting values, or using population-level data to constrain unreliable parameter estimates at the individual level (both of which are employed in this thesis).

It is customary when performing computational modelling to not just consider the fit of a single model, but rather to compare the performance of a number of competing models. This raises an obvious question of how to specify which models to include in the set? Given a free choice from a wide range of models that are a priori possible, selected models should be informed by factors such as how successful they are at predicting performance in related tasks, although such data is not always available. Further, model comparison is only meaningful in the extent to which the models tested are plausible, yet falsifiable; in other words, that there are hypothetical outcomes, that if observed, would falsify a candidate model (Lewandowsky & Farrell, 2011). A further complication arises in the instance when more than one model does an equality effective job at explaining the data. For example, the striatal prediction error discussed above is in some cases similar to Shannon surprise (Shannon, 1948). This presents the danger that the result of a model comparison could represent a bias in a particular data set, as opposed to a true preference for one or the other algorithm. Thus the output of the comparison process warrants, in this case, a more general, as opposed to specific, conclusion about the type of algorithm implemented in the brain.

A related issue is that at the behavioural level, model-based analyses often remain neutral about how, or even whether, some components of the model are realized in neural algorithms, or whether the implementation of a given model is biophysically plausible (Mars, Shea, Kolling, & Rushworth, 2012). Further, in the event that a clear winning model emerges, there is no guarantee that there does not exist a better model that was simply not conceived of by the experimenter. In fact, it has been argued that the number of factors influencing any given cognitive process are far too great to ever allow for a full specification of the "true" underlying model (Burnham & Anderson, 1998). Perhaps most clearly stated by MacCallum, "Regardless of their form or function, or the area in which they are used, it is safe to say that these models all have one thing in common: *They are all wrong*" (MacCallum, 2003). Thus it is important that researchers remain aware of these limitations when drawing conclusions from their data.

A final problem worth discussion is the issue of overfitting. In general, overfitting occurs when a model is excessively complex (owing to too many free parameters), such that it captures elements of the data that are likely driven by random error as opposed to the cause that is of interest. The potential for overfitting is common in the field of decision-making where data are inherently noisy. In our experiments, subjects typically have to perform multiple repetitions of the same trial type in which their response latency, attention span, alertness, and action selection will vary. The problem is illustrated in Figure 7.1 (taken from (Pitt & Myung, 2002)) (p. 190). Plotted on the y-axis is a measure of goodness-of-fit (such as root mean squared error (RMSE), or percentage error accounted for). A common occurrence is that goodness-of-fit will increase in tandem with model complexity (x-axis), but at the risk of fitting noise. The three graphs along the x-axis represent data points (dots) and the corresponding fits (solid lines), as model complexity increases. In the leftmost graph, the model is not complex enough to capture the data. In the second graph, the model and data are well matched, with model generalizability peaking at this point. Here, generalizability refers to the ability of the model to fit all data samples (such as those obtained from a different experimental cohort) generated by the same cognitive process, as opposed to just the current sample. In contrast, the rightmost graph is more complex than the data, and despite providing the best objective fit, is capturing elements of random error.

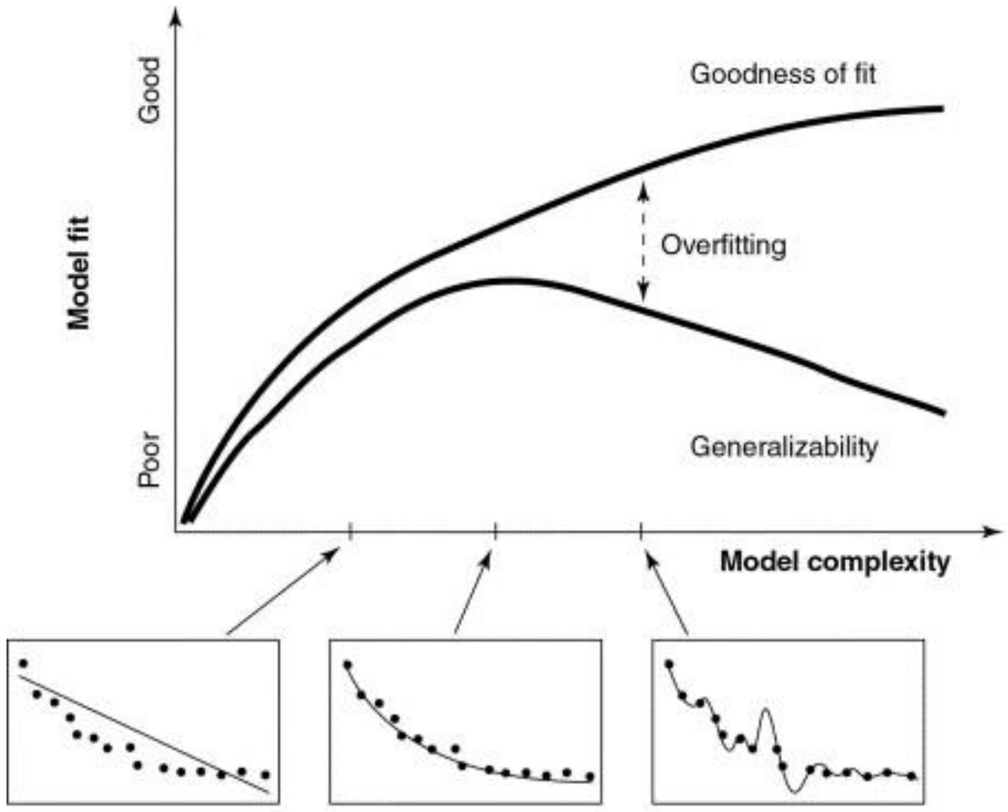


Figure 7.1 Tradeoff between goodness-of-fit and generalizability in cognitive modelling; taken from (Pitt & Myung, 2002). Increasing model complexity improves the objective model fit (y-axis), as a higher percentage of variance is accounted for. By contrast, whilst model generalizability initially follows the same trajectory, peaking when the complexity of the model and the data are well matched (middle graph on the x-axis), it declines with increasing complexity due to fitting of variance (noise) unrelated to the underlying cognitive process.

Pitt and Myung further demonstrated the importance of overfitting using the principles of model recovery (Pitt & Myung, 2002). Here, the authors used a one-parameter model, *M_x*, to generate a sample of data, and added some sampling noise. Next, they fit the original model, *M_x*, to the data, in addition to a more complex model, *M_y*, which contained three parameters. The authors then asked whether the simulated data was fit best by *M_x*, the true model, or the competing alternative, *M_y*. To their surprise, Pitt and Myung found the data were better accommodated by the more complex model, as measured by RMSE, compared to the true underlying model. In this scenario, *Mx* and *M_y* were equally probable a priori, and the simulation thus serves to demonstrate the potential dangers of overfitting. One can at least partially safeguard from these dangers but deriving measures of model fit that account for complexity (the number of free parameters), such as the Bayesian Information Criterion. However, Pitt and Muyng argue that a model's data-fitting abilities are also affected by other properties of the model, such as its functional form, which are often too easily overlooked (Pitt & Myung, 2002).

7.4 Concluding remarks

In this thesis, I combined economic paradigms with computational modelling (and neuroimaging) to draw conclusions about the cognitive and neural mechanisms that support adaptive economic choice. Behavioural, neural and computational data contribute differentially towards our understanding of individual components of decision-making. Yet, when combined, these divergent methodologies form a largely unified framework which speaks to the usefulness of this approach. I have extensively reviewed previous evidence that humans and other animals make choices by assigning 'values' to potential choice options which then compete for action selection. I then presented data in support of a framework where multiple interacting neural systems contribute to this valuation. In chapter 6, I focus on how the behavioural manifestation of these systems evolves with task training.

In chapters 4 and 5, I show that one system, involving the striatum (and possibly ventromedial prefrontal cortex), is short-sighted and responds to basic and immediate outcomes. This system appears insensitive to context-dependent information and may contribute towards choices that are inconsistent with higher-order goals. On the other hand, a second system, likely involving the ventromedial and dorsal prefrontal cortex (vmPFC/dPFC), is far-sighted and is associated with abstract or delayed outcomes and goals. One possibility is that dPFC modulates value representations within vmPFC, enacting controlled choice. In this framework, an enhancement of the representation of long-term (or context-specific) value in vmPFC, and a suppression of a representation of short-term (or prepotent) value in striatum, may contribute to more controlled choice, while impulsive choice may arise from the reverse. Further, a finding that anterior cingulate cortex (ACC) modulates value representations in vmPFC in response to a change in environmental context (see Chapter 4) contributes to an ongoing debate regarding the precise role of this region in decision-making. In particular, while previous work proposes ACC signals a non-specific conflict signal (or Bayesian surprise) (Ide et al., 2013; Shenhav et al., 2013), my data suggests this interpretation could instead be re-framed as signalling a need to switch behaviour away from a default or prepotent action. Future work might expand on this idea by re-examining previous data in light of such a framework.

Further experiments are also needed to elucidate exactly how these dual systems interact during choice, and why the representation of long-term value is disrupted in impulsive individuals. For example, one could use magnetoencephalography (MEG) to better characterize the temporal dynamics of value representation in vmPFC and striatum during decisions that require self-control. Imaging of deep or superficial structures using MEG is becoming increasingly feasible with more advanced signal processing techniques (Kanal, Sun, Ozkurt, Jia, & Sclabassi, 2009). Further, it is unclear whether long-term value computations in vmPFC can arise without a modulatory influence from dPFC. One could test this by disrupting dPFC function with transcranial magnetic stimulation (TMS), and assessing whether choice becomes increasingly impulsive.

In chapter 6, I draw on a parallel dual-systems account of value-based decision-making in the context of reinforcement learning. In this framework, one system is thought to be model-free (favourable actions are memorized and not flexibly updated with new information) and resource-independent, while the other system is thought to be model-based (flexibly transforming new information through a model of the world) and resource-dependent (Daw et al., 2005). Although these systems are computationally distinct from those discussed in Chapters 4 and 5, it is likely that their implementation involves overlapping neural substrates and mechanisms. Here, I provide at least provisional evidence that the implementation of model-based choice changes as task familiarity increases, raising the possibility that model-based reasoning becomes less reliant on executive resources over time. An alternative interpretation is that the model-free system is able to instigate increasingly complex heuristics that progressively resemble a model-based computation, perhaps by incorporating increasingly sophisticated models of the world. Given a mounting interest in model-free versus model-based decision-making, including a growing number of experiments utilizing tasks that dissociate both systems (Daw et al., 2011; S. W. Lee et al., 2014; K. J. Miller et al., 2013; Otto, Gershman, et al., 2013; Otto, Raio, et al., 2013; Skatova et al., 2013; Smittenaar et al., 2013; Smittenaar et al., 2014; Wunderlich, Smittenaar, et al., 2012), and evidence that model-based reasoning is disrupted in psychiatric disorders (Voon et al., 2014), the data I present provides strong incentive for follow-up experiments (perhaps using neuroimaging) with the aim of teasing apart these divergent mechanisms.

References

Adams, C. D., & Dickinson, A. (1981). Instrumental Responding Following Reinforcer Devaluation. *Quarterly Journal of Experimental Psychology Section B-Comparative and Physiological Psychology, 33*(May), 109-121.

Adams, D. L., & Horton, J. C. (2003). A precise retinotopic map of primate striate cortex generated from the representation of angioscotomas. *J Neurosci, 23*(9), 3771-3789.

Alvarez, J. A., & Emory, E. (2006). Executive function and the frontal lobes: a meta-analytic review. *Neuropsychol Rev, 16*(1), 17-42. doi: 10.1007/s11065-006-9002-x

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends Cogn Sci, 8*(4), 170-177. doi: 10.1016/j.tics.2004.02.010

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends Cogn Sci, 12*(5), 193-200. doi: 10.1016/j.tics.2008.02.004

Badre, D., & D'Esposito, M. (2009). Is the rostro-caudal axis of the frontal lobe hierarchical? *Nat Rev Neurosci, 10*(9), 659-669. doi: 10.1038/nrn2667

Badre, D., & Wagner, A. D. (2004). Selection, integration, and conflict monitoring; assessing the nature and generality of prefrontal cognitive control mechanisms. *Neuron, 41*(3), 473-487.

Baliki, M. N., Mansour, A., Baria, A. T., Huang, L., Berger, S. E., Fields, H. L., & Apkarian, A. V. (2013). Parceling human accumbens into putative core and shell dissociates encoding of values for reward and pain. *J Neurosci, 33*(41), 16383-16393. doi: 10.1523/JNEUROSCI.1731-13.2013

Balleine, B. W. (2005). Neural bases of food-seeking: affect, arousal and reward in corticostriatolimbic circuits. *Physiol Behav, 86*(5), 717-730. doi: 10.1016/j.physbeh.2005.08.061

Balleine, B. W., Delgado, M. R., & Hikosaka, O. (2007). The role of the dorsal striatum in reward and decision-making. *J Neurosci, 27*(31), 8161-8165. doi: 10.1523/JNEUROSCI.1554-07.2007

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology, 37*(4-5), 407-419.

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology, 35*(1), 48-69. doi: 10.1038/npp.2009.131

Barbas, H., Ghashghaei, H., Dombrowski, S. M., & Rempel-Clower, N. L. (1999). Medial prefrontal cortices are unified by common connections with superior temporal cortices and distinguished by input from memory-related areas in the rhesus monkey. *Journal of Comparative Neurology, 410*(3), 343-367.

Barbas, H., & Pandya, D. N. (1989). Architecture and Intrinsic Connections of the Prefrontal Cortex in the Rhesus-Monkey. *Journal of Comparative Neurology, 286*(3), 353-375. doi: DOI 10.1002/cne.902860306

Barber, A. D., & Carter, C. S. (2005). Cognitive control involved in overcoming prepotent response tendencies and switching between tasks. *Cereb Cortex, 15*(7), 899-912. doi: 10.1093/cercor/bhh189

Barbey, A. K., Koenigs, M., & Grafman, J. (2012). Dorsolateral prefrontal contributions to human working memory. *Cortex*. doi: 10.1016/j.cortex.2012.05.022

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage, 76*, 412-427. doi: 10.1016/j.neuroimage.2013.02.063

Basar, K., Sesia, T., Groenewegen, H., Steinbusch, H. W., Visser-Vandewalle, V., & Temel, Y. (2010). Nucleus accumbens and impulsivity. *Prog Neurobiol, 92*(4), 533-557. doi: 10.1016/j.pneurobio.2010.08.007

Basten, U., Biele, G., Heekeren, H. R., & Fiebach, C. J. (2010). How the brain integrates costs and benefits during decision making. *Proc Natl Acad Sci U S A, 107*(50), 21767-21772. doi: 10.1073/pnas.0908104107

Bates, J. F., & Goldman-Rakic, P. S. (1993). Prefrontal connections of medial motor areas in the rhesus monkey. *Journal of Comparative Neurology, 336*(2), 211-228. doi: 10.1002/cne.903360205

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: is the active self a limited resource? *J Pers Soc Psychol, 74*(5), 1252-1265.

Bear, M. F., Connors, B. W., & Paradiso, M. A. (2007). *Neuroscience: Exploring the Brain* (Third Edition ed.): Lippincott Williams & Wilkins.

Beauchamp, M. H., Dagher, A., Aston, J. A., & Doyon, J. (2003). Dynamic functional changes associated with cognitive skill learning of an adapted version of the Tower of London task. *Neuroimage, 20*(3), 1649-1660.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition, 50*(1-3), 7-15.

Bechara, A., Damasio, H., Tranel, D., & Anderson, S. W. (1998). Dissociation Of working memory from decision making within the human prefrontal cortex. *J Neurosci, 18*(1), 428-437.

Bechara, A., Tranel, D., & Damasio, H. (2000). Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain, 123 (Pt 11)*, 2189-2202.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nat Neurosci, 10*(9), 1214-1221. doi: 10.1038/nn1954

Beierholm, U., Guitart-Masip, M., Economides, M., Chowdhury, R., Duzel, E., Dolan, R., & Dayan, P. (2013). Dopamine modulates reward-related vigor. *Neuropsychopharmacology, 38*(8), 1495-1503. doi: 10.1038/npp.2013.48

Bellman, R. (1957). *Dynamic Programming*: Princeton University Press, Princeton, N.J.

Bjorck, A. (1994). Numerics of Gram-Schmidt Orthogonalization. *Linear Algebra and Its Applications, 198*, 297-316.

Blanchard, T. C., & Hayden, B. Y. (2014). Neurons in dorsal anterior cingulate cortex signal postdecisional variables in a foraging task. *J Neurosci, 34*(2), 646-655. doi: 10.1523/JNEUROSCI.3151-13.2014

Boorman, E. D., Behrens, T. E., Woolrich, M. W., & Rushworth, M. F. (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), 733-743. doi: 10.1016/j.neuron.2009.05.014

Boorman, E. D., Rushworth, M. F., & Behrens, T. E. (2013). Ventromedial Prefrontal and Anterior Cingulate Cortex Adopt Choice and Default Reference Frames during Sequential Multi-Alternative Choice. *J Neurosci, 33*(6), 2242-2253. doi: 10.1523/JNEUROSCI.3022-12.2013

Boschin, E. A., Piekema, C., & Buckley, M. J. (2015). Essential functions of primate frontopolar cortex in cognition. *Proc Natl Acad Sci U S A, 112*(9), E1020-1027. doi: 10.1073/pnas.1419649112

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), 179-181. doi: 10.1038/46035

Botvinick, M. M. (2007). Conflict monitoring and decision making: reconciling two perspectives on anterior cingulate function. *Cogn Affect Behav Neurosci, 7*(4), 356-366.

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

Bouret, S., & Richmond, B. J. (2010). Ventromedial and orbital prefrontal neurons differentially encode internally and externally driven motivational values in monkeys. *J Neurosci, 30*(25), 8591-8601. doi: 10.1523/JNEUROSCI.0049-10.2010

Braver, T. S., Barch, D. M., Gray, J. R., Molfese, D. L., & Snyder, A. (2001). Anterior cingulate cortex and response conflict: effects of frequency, inhibition and errors. *Cereb Cortex, 11*(9), 825-836.

Brodmann, K. (1909). *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues.* Leipzig: Johann Ambrosius Barth.

Brovelli, A., Nazarian, B., Meunier, M., & Boussaoud, D. (2011). Differential roles of caudate nucleus and putamen during instrumental learning. *Neuroimage, 57*(4), 1580-1590. doi: 10.1016/j.neuroimage.2011.05.059

Burnham, P. K., & Anderson, R. D. (1998). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*: Springer.

Buschman, T. J., Denovellis, E. L., Diogo, C., Bullock, D., & Miller, E. K. (2012). Synchronous oscillatory neural ensembles for rules in the prefrontal cortex. *Neuron, 76*(4), 838-846. doi: 10.1016/j.neuron.2012.09.029

Bush, G., Whalen, P. J., Rosen, B. R., Jenike, M. A., McInerney, S. C., & Rauch, S. L. (1998). The counting Stroop: an interference task specialized for functional neuroimaging--validation study with functional MRI. *Hum Brain Mapp, 6*(4), 270-282.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci, 14*(5), 365-376. doi: 10.1038/nrn3475

Cai, W., & Leung, H. C. (2009). Cortical activity during manual response inhibition guided by color and orientation cues. *Brain Res, 1261*, 20-28. doi: 10.1016/j.brainres.2008.12.073

Cai, X., & Padoa-Schioppa, C. (2012). Neuronal encoding of subjective value in dorsal and ventral anterior cingulate cortex. *J Neurosci, 32*(11), 3791-3808. doi: 10.1523/JNEUROSCI.3864-11.2012

Cardinal, R. N., Pennicott, D. R., Sugathapala, C. L., Robbins, T. W., & Everitt, B. J. (2001). Impulsive choice induced in rats by lesions of the nucleus accumbens core. *Science, 292*(5526), 2499-2501. doi: 10.1126/science.1060818

Carmichael, S. T., & Price, J. L. (1994). Architectonic subdivision of the orbital and medial prefrontal cortex in the macaque monkey. *Journal of Comparative Neurology, 346*(3), 366-402. doi: 10.1002/cne.903460305

Carmichael, S. T., & Price, J. L. (1996). Connectional networks within the orbital and medial prefrontal cortex of macaque monkeys. *Journal of Comparative Neurology, 371*(2), 179-207. doi: Doi 10.1002/(Sici)1096-9861(19960722)371:2<179::Aid-Cne1>3.0.Co;2-#

Cavada, C., Company, T., Tejedor, J., Cruz-Rizzolo, R. J., & Reinoso-Suarez, F. (2000). The anatomical connections of the macaque monkey orbitofrontal cortex. A review. *Cereb Cortex, 10*(3), 220-242.

Christoff, K., & Gabrieli, J. D. E. (2000). The frontopolar cortex and human cognition: Evidence for a rostrocaudal hierarchical organization within the human prefrontal cortex. *Psychobiology, 28*(2), 168-186.

Colwill, R. M., & Rescorla, R. A. (1986). Associative Structures in Instrumental Learning. *Psychology of Learning and Motivation-Advances in Research and Theory, 20*, 55-104.

Constable, R. T. (2012). Challenges in fMRI and Its Limitations *Functional Neuroradiology: Principles and Clinical Applications* (pp. 331-344): Springer

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends Cogn Sci, 7*(9), 415-423.

D'Esposito, M., & Postle, B. R. (1999). The dependence of span and delayed-response performance on prefrontal cortex. *Neuropsychologia, 37*(11), 1303-1315.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*(6), 1204-1215. doi: 10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci, 8*(12), 1704-1711. doi: 10.1038/nn1560

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), 876-879. doi: 10.1038/nature04766

Dayan, P. (2008). The Role of Value Systems in Decision Making. In C. a. S. Engel, Wolf (Ed.), *Better Than Conscious? Decision Making, the Human Mind, and Implications For Institutions* (pp. 51-70): MIT Press.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Curr Opin Neurobiol, 18*(2), 185-196. doi: 10.1016/j.conb.2008.08.003

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nat Neurosci, 16*(1), 105-110. doi: 10.1038/nn.3279

deCharms, R. C., & Zador, A. (2000). Neural representation and the cortical code. *Annu Rev Neurosci, 23*, 613-647. doi: 10.1146/annurev.neuro.23.1.613

Delgado, M. R., Miller, M. M., Inati, S., & Phelps, E. A. (2005). An fMRI study of reward-related probability learning. *Neuroimage, 24*(3), 862-873. doi: 10.1016/j.neuroimage.2004.10.002

Dickinson, A. (1980). *Contemporary Animal Learning Theory*. Cambridge, United Kingdom Cambridge University Press.

Dickinson, A., Nicholas, D. J., & Adams, C. D. (1983). The Effect of the Instrumental Training Contingency on Susceptibility to Reinforcer Devaluation. *Quarterly Journal of Experimental Psychology Section B-Comparative and Physiological Psychology, 35*(Feb), 35-51.

Diekhof, E. K., & Gruber, O. (2010). When desire collides with reason: functional interactions between anteroventral prefrontal cortex and nucleus accumbens underlie the human ability to resist impulsive desires. *J Neurosci, 30*(4), 1488-1493. doi: 10.1523/JNEUROSCI.4690-09.2010

Dietrich, M. O., & Horvath, T. L. (2013). Hypothalamic control of energy balance: insights into the role of synaptic plasticity. *Trends Neurosci, 36*(2), 65-73. doi: DOI 10.1016/j.tins.2012.12.005

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron, 80*(2), 312-325. doi: 10.1016/j.neuron.2013.09.007

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Curr Opin Neurobiol, 22*(6), 1075-1081. doi: 10.1016/j.conb.2012.08.003

Donald, N. A., & Tim, S. (1986). Attention to action: Willed and automatic control of behavior. . In R. J. Davidson, G. E. Schwartz & D. Shapiro (Eds.), (pp. pp 1-18). New York, NY: Plenum.

Douglas, R. J., & Martin, K. A. (2004). Neuronal circuits of the neocortex. *Annu Rev Neurosci, 27*, 419-451. doi: 10.1146/annurev.neuro.27.070203.144152

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends Neurosci, 23*(10), 475-483.

Dux, P. E., Ivanoff, J., Asplund, C. L., & Marois, R. (2006). Isolation of a central bottleneck of information processing with time-resolved FMRI. *Neuron, 52*(6), 1109-1120. doi: 10.1016/j.neuron.2006.11.009

Economides, M., Guitart-Masip, M., Kurth-Nelson, Z., & Dolan, R. J. (2014). Anterior cingulate cortex instigates adaptive switches in choice by integrating immediate and delayed components of value in ventromedial prefrontal cortex. *J Neurosci, 34*(9), 3340-3349. doi: 10.1523/JNEUROSCI.4313-13.2014

Economo, C., & Koskinas, G. N. (1925). *Die Cytoarchitektonik der Hirnrinde Des Erwachsenen Menschen.* Springer, Wien.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of Noise Letters Upon Identification of a Target Letter in a Nonsearch Task. *Perception & Psychophysics, 16*(1), 143-149. doi: Doi 10.3758/Bf03203267

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nat Neurosci, 8*(11), 1481-1489. doi: 10.1038/nn1579

Farooqi, I. S., Bullmore, E., Keogh, J., Gillard, J., O'Rahilly, S., & Fletcher, P. C. (2007). Leptin regulates striatal regions and human eating Behavior. *Science, 317*(5843), 1355-1355. doi: DOI 10.1126/science.1144599

Featherstone, R. E., & McDonald, R. J. (2005). Lesions of the dorsolateral striatum impair the acquisition of a simplified stimulus-response dependent conditional discrimination task. *Neuroscience, 136*(2), 387-395. doi: 10.1016/j.neuroscience.2005.08.021

Fellows, L. K. (2013). Lesion studies in affective neuroscience. In J. Armony & P. Villeumier (Eds.), *The Cambridge handbook of human affective neuroscience* (pp. 154-167): Cambridge University Press.

Fellows, L. K., & Farah, M. J. (2005). Different underlying impairments in decision-making following ventromedial and dorsolateral frontal lobe damage in humans. *Cereb Cortex, 15*(1), 58-63. doi: 10.1093/cercor/bhh108

FitzGerald, T. H., Friston, K. J., & Dolan, R. J. (2012). Action-specific value signals in reward-related regions of the human brain. *J Neurosci, 32*(46), 16417-16423a. doi: 10.1523/JNEUROSCI.3254-12.2012

Friston, K., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiack, R. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*.

Frith, C. D., Friston, K., Liddle, P. F., & Frackowiak, R. S. (1991). Willed action and the prefrontal cortex in man: a study with PET. *Proc Biol Sci, 244*(1311), 241-246. doi: 10.1098/rspb.1991.0077

Fuhrer, D., Zysset, S., & Stumvoll, M. (2008). Brain activity in hunger and satiety: an exploratory visually stimulated FMRI study. *Obesity (Silver Spring), 16*(5), 945-950. doi: 10.1038/oby.2008.33

Gabriel, M., Burhans, L., Talk, A., & Scalf, P. (2002). The Cingulate Cortex. In V. S. Ramachandran (Ed.), *Encyclopedia of The Human Brain* (pp. 775-791): Elsevier Science.

Gailliot, M. T., & Baumeister, R. F. (2007). The physiology of willpower: linking blood glucose to self-control. *Pers Soc Psychol Rev, 11*(4), 303-327. doi: 10.1177/1088868307303030

Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science, 295*(5563), 2279-2282. doi: 10.1126/science.1066893

Glascher, J., Adolphs, R., Damasio, H., Bechara, A., Rudrauf, D., Calamia, M., . . . Tranel, D. (2012). Lesion mapping of cognitive control and value-based decision making in the prefrontal cortex. *Proc Natl Acad Sci U S A, 109*(36), 14681-14686. doi: 10.1073/pnas.1206608109

Glascher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron, 66*(4), 585-595. doi: 10.1016/j.neuron.2010.04.016

Goldman-Rakic, P. S. (1988). Topography of cognition: parallel distributed networks in primate association cortex. *Annu Rev Neurosci, 11*, 137-156. doi: 10.1146/annurev.ne.11.030188.001033

Guitart-Masip, M., Duzel, E., Dolan, R., & Dayan, P. (2014). Action versus valence in decision making. *Trends Cogn Sci, 18*(4), 194-202. doi: 10.1016/j.tics.2014.01.003

Guitart-Masip, M., Fuentemilla, L., Bach, D. R., Huys, Q. J., Dayan, P., Dolan, R. J., & Duzel, E. (2011). Action dominates valence in anticipatory representations in the human striatum and dopaminergic midbrain. *J Neurosci, 31*(21), 7867-7875. doi: 10.1523/JNEUROSCI.6376-10.2011

Guitart-Masip, M., Huys, Q. J., Fuentemilla, L., Dayan, P., Duzel, E., & Dolan, R. J. (2012). Go and no-go learning in reward and punishment: interactions between affect and effect. *Neuroimage, 62*(1), 154-166. doi: 10.1016/j.neuroimage.2012.04.024

Haber, S. N. (2011). Neuroanatomy of Reward: A View from the Ventral Striatum. In J. A. Gottfried (Ed.), *Neurobiology of Sensation and Reward*. Boca Raton (FL).

Haber, S. N., & Knutson, B. (2010). The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology, 35*(1), 4-26. doi: 10.1038/npp.2009.129

Haber, S. N., Kunishio, K., Mizobuchi, M., & Lynd-Balta, E. (1995). The orbital and medial prefrontal circuit through the primate basal ganglia. *J Neurosci, 15*(7 Pt 1), 4851-4867.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006a). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience, 26*(32), 8360-8367. doi: Doi 10.1523/Jneurosci.1010-06.2006

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006b). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J Neurosci, 26*(32), 8360-8367. doi: 10.1523/JNEUROSCI.1010-06.2006

Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science, 324*(5927), 646-648. doi: 10.1126/science.1168450

Hare, T. A., Hakimi, S., & Rangel, A. (2014). Activity in dlPFC and its effective connectivity to vmPFC are associated with temporal discounting. *Front Neurosci, 8*, 50. doi: 10.3389/fnins.2014.00050

Hare, T. A., Malmaud, J., & Rangel, A. (2011). Focusing attention on the health aspects of foods changes value signals in vmPFC and improves dietary choice. *J Neurosci, 31*(30), 11077-11087. doi: 10.1523/JNEUROSCI.6383-10.2011

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *J Neurosci, 28*(22), 5623-5630. doi: 10.1523/JNEUROSCI.1309-08.2008

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: the early beginnings. *Neuroimage, 62*(2), 852-855. doi: 10.1016/j.neuroimage.2012.03.016

Hazeltine, E., Teague, D., & Ivry, R. B. (2002). Simultaneous dual-task performance reveals parallel response selection after practice. *J Exp Psychol Hum Percept Perform, 28*(3), 527-545.

Heekeren, H. R., Marrett, S., & Ungerleider, L. G. (2008). The neural systems that mediate human perceptual decision making. *Nature Reviews Neuroscience, 9*(6), 467-479. doi: Doi 10.1038/Nrn2374

Hershberger, W. A. (1986). An Approach through the Looking-Glass. *Animal Learning & Behavior, 14*(4), 443-451. doi: Doi 10.3758/Bf03200092

Hofmann, W., Friese, M., & Strack, F. (2009). Impulse and Self-Control From a Dual-Systems Perspective. *Perspectives on Psychological Science, 4*(2), 162-176. doi: DOI 10.1111/j.1745-6924.2009.01116.x

Hofmann, W., & Van Dillen, L. (2012). Desire: The New Hot Spot in Self-Control Research. *Current Directions in Psychological Science, 21*(5), 317-322. doi: Doi 10.1177/0963721412453587

Hunt, L. T., Kolling, N., Soltani, A., Woolrich, M. W., Rushworth, M. F., & Behrens, T. E. (2012). Mechanisms underlying cortical activity during value-guided choice. *Nat Neurosci, 15*(3), 470-476, S471-473. doi: 10.1038/nn.3017

Hunt, L. T., Woolrich, M. W., Rushworth, M. F., & Behrens, T. E. (2013). Trial-type dependent frames of reference for value comparison. *PLoS Comput Biol, 9*(9), e1003225. doi: 10.1371/journal.pcbi.1003225

Hutton, C., Josephs, O., Stadler, J., Featherstone, E., Reid, A., Speck, O., . . . Weiskopf, N. (2011). The impact of physiological noise correction on fMRI at 7 T. *Neuroimage, 57*(1), 101-112. doi: 10.1016/j.neuroimage.2011.04.018

Huys, Q. J., Cools, R., Golzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS Comput Biol, 7*(4), e1002028. doi: 10.1371/journal.pcbi.1002028

Ide, J. S., Shenoy, P., Yu, A. J., & Li, C. S. (2013). Bayesian prediction and evaluation in the anterior cingulate cortex. *J Neurosci, 33*(5), 2039-2047. doi: 10.1523/JNEUROSCI.2201-12.2013

Ito, S., Stuphorn, V., Brown, J. W., & Schall, J. D. (2003). Performance monitoring by the anterior cingulate cortex during saccade countermanding. *Science, 302*(5642), 120-122. doi: 10.1126/science.1087847

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short- and long-term benefits of cognitive training. *Proc Natl Acad Sci U S A, 108*(25), 10081-10086. doi: 10.1073/pnas.1103228108

Jenison, R. L., Rangel, A., Oya, H., Kawasaki, H., & Howard, M. A. (2011). Value encoding in single neurons in the human amygdala during decision making. *J Neurosci, 31*(1), 331-338. doi: 10.1523/JNEUROSCI.4461-10.2011

Jezzard, P., Smith, S. M., & Matthews, P. W. (2003). *Functional MRI: an introduction to methds.* . Oxford: Oxford University Press.

Jocham, G., Furlong, P. M., Kroger, I. L., Kahn, M. C., Hunt, L. T., & Behrens, T. E. (2014). Dissociable contributions of ventromedial prefrontal and posterior parietal cortex to value-guided choice. *Neuroimage, 100C*, 498-506. doi: 10.1016/j.neuroimage.2014.06.005

Johnson, D. (2001). Frontal lobes and executive function. *Pediatr Rehabil, 4*(3), 101-103.

Jones, J. L., Esber, G. R., McDannald, M. A., Gruber, A. J., Hernandez, A., Mirenzi, A., & Schoenbaum, G. (2012). Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science, 338*(6109), 953-956. doi: 10.1126/science.1227489

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nat Neurosci, 10*(12), 1625-1633. doi: 10.1038/nn2007

Kanal, E. Y., Sun, M., Ozkurt, T. E., Jia, W., & Sclabassi, R. (2009). Magnetoencephalographic imaging of deep corticostriatal network activity during a rewards paradigm. *Conf Proc IEEE Eng Med Biol Soc, 2009*, 2915-2918. doi: 10.1109/IEMBS.2009.5334490

Kass, R., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association, 90*.

Kelly, A. M., & Garavan, H. (2005). Human functional neuroimaging of brain changes associated with practice. *Cereb Cortex, 15*(8), 1089-1102. doi: 10.1093/cercor/bhi005

Kennerley, S. W., Behrens, T. E., & Wallis, J. D. (2011). Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. *Nat Neurosci, 14*(12), 1581-1589. doi: 10.1038/nn.2961

Kennerley, S. W., Dahmubed, A. F., Lara, A. H., & Wallis, J. D. (2009). Neurons in the frontal lobe encode the value of multiple decision variables. *J Cogn Neurosci, 21*(6), 1162-1178. doi: 10.1162/jocn.2009.21100

Kennerley, S. W., & Walton, M. E. (2011). Decision making and reward in frontal cortex: complementary evidence from neurophysiological and neuropsychological studies. *Behav Neurosci, 125*(3), 297-317. doi: 10.1037/a0023575

Kennerley, S. W., Walton, M. E., Behrens, T. E., Buckley, M. J., & Rushworth, M. F. (2006). Optimal decision making and the anterior cingulate cortex. *Nat Neurosci, 9*(7), 940-947. doi: 10.1038/nn1724

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature, 455*(7210), 227-231. doi: 10.1038/nature07200

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Comput Biol, 7*(5), e1002055. doi: 10.1371/journal.pcbi.1002055

Keramati, M., & Gutkin, B. S. (2011). A Reinforcement Learning Theory for Homeostatic Regulation. *Neural Information Processing Systems*, 82 - 90.

Kerns, J. G., Cohen, J. D., MacDonald, A. W., 3rd, Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science, 303*(5660), 1023-1026. doi: 10.1126/science.1089910

Kilner, J. M., Kiebel, S. J., & Friston, K. J. (2005). Applications of random field theory to electrophysiology. *Neurosci Lett, 374*(3), 174-178. doi: 10.1016/j.neulet.2004.10.052

Kimchi, E. Y., & Laubach, M. (2009). Dynamic encoding of action selection by the medial striatum. *J Neurosci, 29*(10), 3148-3159. doi: 10.1523/JNEUROSCI.5206-08.2009

Knight, R. T., Grabowecky, M. F., & Scabini, D. (1995). Role of human prefrontal cortex in attention control. *Adv Neurol, 66*, 21-34; discussion 34-26.

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J Neurosci, 21*(16), RC159.

Koechlin, E., Basso, G., Pietrini, P., Panzer, S., & Grafman, J. (1999). The role of the anterior prefrontal cortex in human cognition. *Nature, 399*(6732), 148-151. doi: 10.1038/20178

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science, 302*(5648), 1181-1185. doi: 10.1126/science.1088545

Kolling, N., Behrens, T. E., Mars, R. B., & Rushworth, M. F. (2012). Neural mechanisms of foraging. *Science, 336*(6077), 95-98. doi: 10.1126/science.1216930

Kringelbach, M. L. (2005). The human orbitofrontal cortex: linking reward to hedonic experience. *Nat Rev Neurosci, 6*(9), 691-702. doi: 10.1038/nrn1747

Kringelbach, M. L., O'Doherty, J., Rolls, E. T., & Andrews, C. (2003). Activation of the human orbitofrontal cortex to a liquid food stimulus is correlated with its subjective pleasantness. *Cereb Cortex, 13*(10), 1064-1071.

Kringelbach, M. L., & Rolls, E. T. (2004). The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. *Prog Neurobiol, 72*(5), 341-372. doi: 10.1016/j.pneurobio.2004.03.006

LaBar, K. S., Gitelman, D. R., Parrish, T. B., Kim, Y. H., Nobre, A. C., & Mesulam, M. M. (2001). Hunger selectively modulates corticolimbic activation to food stimuli in humans. *Behav Neurosci, 115*(2), 493-500.

Lam, T. K. T., Schwartz, G. J., & Rossetti, L. (2005). Hypothalamic sensing of fatty acids. *Nature Neuroscience, 8*(5), 579-584. doi: Doi 10.1038/Nn1456

Lavenex, P. B., Amaral, D. G., & Lavenex, P. (2006). Hippocampal lesion prevents spatial relational learning in adult macaque monkeys. *J Neurosci, 26*(17), 4546-4558. doi: 10.1523/JNEUROSCI.5412-05.2006

Lee, D., Seo, H., & Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annu Rev Neurosci, 35*, 287-308. doi: 10.1146/annurev-neuro-062111-150512

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. *Neuron, 81*(3), 687-699. doi: DOI 10.1016/j.neuron.2013.11.028

Lewandowsky, S., & Farrell, S. (2011). *Computational Modeling In Cognition: Principles and Practice*: SAGE Publications.

Lim, S. L., O'Doherty, J. P., & Rangel, A. (2011). The decision value computations in the vmPFC and striatum use a relative value code that is guided by visual attention. *J Neurosci, 31*(37), 13214-13223. doi: 10.1523/JNEUROSCI.1246-11.2011

Liston, C., Matalon, S., Hare, T. A., Davidson, M. C., & Casey, B. J. (2006). Anterior cingulate and posterior parietal cortices are sensitive to dissociable forms of conflict in a task-switching paradigm. *Neuron, 50*(4), 643-653. doi: 10.1016/j.neuron.2006.04.015

Litt, A., Plassmann, H., Shiv, B., & Rangel, A. (2011). Dissociating valuation and saliency signals during decision-making. *Cereb Cortex, 21*(1), 95-102. doi: 10.1093/cercor/bhq065

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes, 65*(3), 272-292. doi: DOI 10.1006/obhd.1996.0028

Logan, G. D., Cowan, W. B., & Davis, K. A. (1984). On the ability to inhibit simple and choice reaction time responses: a model and a method. *J Exp Psychol Hum Percept Perform, 10*(2), 276-291.

Logothetis, N. K. (2003). The underpinnings of the BOLD functional magnetic resonance imaging signal. *J Neurosci, 23*(10), 3963-3971.

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature, 453*(7197), 869-878. doi: Doi 10.1038/Nature06976

Lu, M. T., Preston, J. B., & Strick, P. L. (1994). Interconnections between the Prefrontal Cortex and the Premotor Areas in the Frontal-Lobe. *Journal of Comparative Neurology, 341*(3), 375-392. doi: DOI 10.1002/cne.903410308

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research, 38*(1), 113-139. doi: Doi 10.1207/S15327906mbr3801_5

Manes, F., Sahakian, B., Clark, L., Rogers, R., Antoun, N., Aitken, M., & Robbins, T. (2002). Decision-making processes following damage to the prefrontal cortex. *Brain, 125*(Pt 3), 624-639.

Mansouri, F. A., Tanaka, K., & Buckley, M. J. (2009). Conflict-induced behavioural adjustment: a clue to the executive functions of the prefrontal cortex. *Nat Rev Neurosci, 10*(2), 141-152. doi: 10.1038/nrn2538

Mars, R. B., Shea, N. J., Kolling, N., & Rushworth, M. F. S. (2012). Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. *Quarterly Journal of Experimental Psychology, 65*(2), 252-267. doi: Doi 10.1080/17470211003668272

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science, 306*(5695), 503-507. doi: 10.1126/science.1100907

McDannald, M. A., Lucantonio, F., Burke, K. A., Niv, Y., & Schoenbaum, G. (2011). Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *J Neurosci, 31*(7), 2700-2705. doi: 10.1523/JNEUROSCI.5499-10.2011

McLaren, D. G., Ries, M. L., Xu, G., & Johnson, S. C. (2012). A generalized form of context-dependent psychophysiological interactions (gPPI): a comparison to standard approaches. *Neuroimage, 61*(4), 1277-1286. doi: 10.1016/j.neuroimage.2012.03.068

Melby-Lervag, M., & Hulme, C. (2013). Is Working Memory Training Effective? A Meta-Analytic Review. *Developmental Psychology, 49*(2), 270-291. doi: Doi 10.1037/A0028228

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu Rev Neurosci, 24*, 167-202. doi: 10.1146/annurev.neuro.24.1.167

Miller, K. J., Erlich, J. C., Kopec, C. D., Botvinick, M. M., & Brody, C. D. (2013). *A multi-step decision task to distinguish model-based from model-free reinforcement learning in rats*. Poster. Presented at Society for Neuroscience, San Diego (855.13).

Miller, N. E., & Kessen, M. L. (1952). Reward effects of food via stomach fistula compared with those of food via mouth. *J Comp Physiol Psychol, 45*(6), 555-564.

Minokoshi, Y., Alquier, T., Furukawa, N., Kim, Y. B., Lee, A., Xue, B. Z., . . . Kahn, B. B. (2004). AMP-kinase regulates food intake by responding to hormonal and nutrient signals in the hypothalamus. *Nature, 428*(6982), 569-574. doi: Doi 10.1038/Nature02440

Moeller, F. G., Barratt, E. S., Dougherty, D. M., Schmitz, J. M., & Swann, A. C. (2001). Psychiatric aspects of impulsivity. *Am J Psychiatry, 158*(11), 1783-1793.

Monosov, I. E., & Hikosaka, O. (2012). Regionally distinct processing of rewards and punishments by the primate ventromedial prefrontal cortex. *J Neurosci, 32*(30), 10318-10330. doi: 10.1523/JNEUROSCI.1801-12.2012

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

Moore, T. L., Schettler, S. P., Killiany, R. J., Rosene, D. L., & Moss, M. B. (2009). Effects on executive function following damage to the prefrontal cortex in the rhesus monkey (Macaca mulatta). *Behav Neurosci, 123*(2), 231-241. doi: 10.1037/a0014723

Morris, R. W., Dezfouli, A., Griffiths, K. R., & Balleine, B. W. (2014). Action-value comparisons in the dorsolateral prefrontal cortex control choice between goal-directed actions. *Nat Commun, 5*, 4390. doi: 10.1038/ncomms5390

Morrison, S. E., & Salzman, C. D. (2009). The convergence of information about rewarding and aversive stimuli in single neurons. *J Neurosci, 29*(37), 11471-11483. doi: 10.1523/JNEUROSCI.1815-09.2009

Narayanan, N. S., Guarnieri, D. J., & DiLeone, R. J. (2010). Metabolic hormones, dopamine circuits, and feeding. *Front Neuroendocrinol, 31*(1), 104-112. doi: 10.1016/j.yfrne.2009.10.004

Narayanan, N. S., Prabhakaran, V., Bunge, S. A., Christoff, K., Fine, E. M., & Gabrieli, J. D. (2005). The role of the prefrontal cortex in the maintenance of verbal working memory: an event-related FMRI analysis. *Neuropsychology, 19*(2), 223-232. doi: 10.1037/0894-4105.19.2.223

Newman, S. D., Carpenter, P. A., Varma, S., & Just, M. A. (2003). Frontal and parietal participation in problem solving in the Tower of London: fMRI and computational modeling of planning and high-level perception. *Neuropsychologia, 41*(12), 1668-1682.

Niv, Y., Daw, N., & Dayan, P. (2005). How fast to work: Response vigor, motivation and tonic dopamine. *Neural Information Processing Systems*, 1019 - 1026.

Niv, Y., Joel, D., & Dayan, P. (2006). A normative perspective on motivation. *Trends Cogn Sci, 10*(8), 375-381. doi: 10.1016/j.tics.2006.06.010

Noonan, M. P., Walton, M. E., Behrens, T. E., Sallet, J., Buckley, M. J., & Rushworth, M. F. (2010). Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. *Proc Natl Acad Sci U S A, 107*(47), 20547-20552. doi: 10.1073/pnas.1012246107

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, 304*(5669), 452-454. doi: 10.1126/science.1094285

O'Doherty, J. P., Buchanan, T. W., Seymour, B., & Dolan, R. J. (2006). Predictive neural coding of reward preference involves dissociable responses in human ventral midbrain and ventral striatum. *Neuron, 49*(1), 157-166. doi: 10.1016/j.neuron.2005.11.014

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

O'Doherty, J. P., Deichmann, R., Critchley, H. D., & Dolan, R. J. (2002). Neural responses during anticipation of a primary taste reward. *Neuron, 33*(5), 815-826.

O'Keefe, J., & Nadel, L. (1978). *The Hippocampus as a Cognitive Map*: Oxford University Press.

O'Reilly, J. X., Schuffelgen, U., Cuell, S. F., Behrens, T. E., Mars, R. B., & Rushworth, M. F. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proc Natl Acad Sci U S A, 110*(38), E3660-3669. doi: 10.1073/pnas.1305373110

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc Natl Acad Sci U S A, 87*(24), 9868-9872.

Olds, J., & Milner, P. (1954). Positive Reinforcement Produced by Electrical Stimulation of Septal Area and Other Regions of Rat Brain. *J Comp Physiol Psychol, 47*(6), 419-427. doi: Doi 10.1037/H0058775

Ongur, D., An, X., & Price, J. L. (1998). Prefrontal cortical projections to the hypothalamus in macaque monkeys. *Journal of Comparative Neurology, 401*(4), 480-505.

Ongur, D., & Price, J. L. (2000). The organization of networks within the orbital and medial prefrontal cortex of rats, monkeys and humans. *Cereb Cortex, 10*(3), 206-219. doi: DOI 10.1093/cercor/10.3.206

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. *Psychol Sci, 24*(5), 751-761. doi: 10.1177/0956797612463080

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proc Natl Acad Sci U S A, 110*(52), 20941-20946. doi: 10.1073/pnas.1312011110

Owen, A. M. (1997). Cognitive planning in humans: neuropsychological, neuroanatomical and neuropharmacological perspectives. *Prog Neurobiol, 53*(4), 431-450.

Padoa-Schioppa, C. (2011). Neurobiology of economic choice: a good-based model. *Annu Rev Neurosci, 34*, 333-359. doi: 10.1146/annurev-neuro-061010-113648

Padoa-Schioppa, C., & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature, 441*(7090), 223-226. doi: 10.1038/nature04676

Padoa-Schioppa, C., & Assad, J. A. (2008). The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. *Nat Neurosci, 11*(1), 95-102. doi: 10.1038/nn2020

Pandya, D. N., & Seltzer, B. (1982). Intrinsic connections and architectonics of posterior parietal cortex in the rhesus monkey. *Journal of Comparative Neurology, 204*(2), 196-210. doi: 10.1002/cne.902040208

Parent, A., & Hazrati, L. N. (1995). Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop. *Brain Res Brain Res Rev, 20*(1), 91-127.

Petrides, M. (1987). *Conditional learning and the primate frontal cortex* (E. Perecman Ed.). New York: ERBN Press.

Petrides, M. (1994). *Frontal lobes and working memory: evidence from investigations of the effects of cortical excisions in nonhuman primates* (F. Boller & J. Grafman Eds. Vol. 9). Amsterdam.

Petrides, M. (1996). Specialized systems for the processing of mnemonic information within the primate frontal cortex. *Philos Trans R Soc Lond B Biol Sci, 351*(1346), 1455-1461; discussion 1461-1452. doi: 10.1098/rstb.1996.0130

Petrides, M. (2000). Dissociable roles of mid-dorsolateral prefrontal and anterior inferotemporal cortex in visual working memory. *J Neurosci, 20*(19), 7496-7503.

Petrides, M. (2005). Lateral prefrontal cortex: architectonic and functional organization. *Philos Trans R Soc Lond B Biol Sci, 360*(1456), 781-795. doi: 10.1098/rstb.2005.1631

Petrides, M., Alivisatos, B., & Frey, S. (2002). Differential activation of the human orbital, mid-ventrolateral, and mid-dorsolateral prefrontal cortex during the processing of visual stimuli. *Proc Natl Acad Sci U S A, 99*(8), 5649-5654. doi: 10.1073/pnas.072092299

Petrides, M., & Pandya, D. N. (1999). Dorsolateral prefrontal cortex: comparative cytoarchitectonic analysis in the human and the macaque brain and corticocortical connection patterns. *Eur J Neurosci, 11*(3), 1011-1036.

Petrides, M., & Pandya, D. N. (2002). Comparative cytoarchitectonic analysis of the human and the macaque ventrolateral prefrontal cortex and corticocortical connection patterns in the monkey. *Eur J Neurosci, 16*(2), 291-310.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6*(10), 421-425. doi: Pii S1364-6613(02)01964-2

Doi 10.1016/S1364-6613(02)01964-2

Plassmann, H., O'Doherty, J., & Rangel, A. (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *J Neurosci, 27*(37), 9984-9988. doi: 10.1523/JNEUROSCI.2131-07.2007

Plassmann, H., O'Doherty, J. P., & Rangel, A. (2010). Appetitive and aversive goal values are encoded in the medial orbitofrontal cortex at the time of decision making. *J Neurosci, 30*(32), 10799-10808. doi: 10.1523/JNEUROSCI.0788-10.2010

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature, 400*(6741), 233-238. doi: 10.1038/22268

Pochon, J. B., Riis, J., Sanfey, A. G., Nystrom, L. E., & Cohen, J. D. (2008). Functional imaging of decision conflict. *J Neurosci, 28*(13), 3468-3473. doi: 10.1523/JNEUROSCI.4195-07.2008

Poldrack, R. A. (2000). Imaging brain plasticity: conceptual and methodological issues--a theoretical review. *Neuroimage, 12*(1), 1-13. doi: 10.1006/nimg.2000.0596

Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience, 139*(1), 23-38. doi: 10.1016/j.neuroscience.2005.06.005

Price, J. L., & Drevets, W. C. (2010). Neurocircuitry of mood disorders. *Neuropsychopharmacology, 35*(1), 192-216. doi: 10.1038/npp.2009.104

Ramnani, N., & Owen, A. M. (2004). Anterior prefrontal cortex: Insights into function from anatomy and neuroimaging. *Nature Reviews Neuroscience, 5*(3), 184-194. doi: Doi 10.1038/Nrn1343

Rangel, A. (2013). Regulation of dietary choice by the decision-making circuitry. *Nat Neurosci, 16*(12), 1717-1724. doi: 10.1038/nn.3561

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nat Rev Neurosci, 9*(7), 545-556. doi: 10.1038/nrn2357

Rangel, A., & Hare, T. (2010). Neural computations associated with goal-directed choice. *Curr Opin Neurobiol, 20*(2), 262-270. doi: 10.1016/j.conb.2010.03.001

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput, 20*(4), 873-922. doi: 10.1162/neco.2008.12-06-420

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

Reverberi, C., Gorgen, K., & Haynes, J. D. (2012). Distributed representations of rule identity and rule order in human frontal cortex and striatum. *J Neurosci, 32*(48), 17420-17430. doi: 10.1523/JNEUROSCI.2344-12.2012

Rich, E. L., & Wallis, J. D. (2014). Medial-lateral organization of the orbitofrontal cortex. *J Cogn Neurosci, 26*(7), 1347-1362. doi: 10.1162/jocn_a_00573

Roesch, M. R., Calu, D. J., & Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat Neurosci, 10*(12), 1615-1624. doi: 10.1038/nn2013

Roesch, M. R., Singh, T., Brown, P. L., Mullins, S. E., & Schoenbaum, G. (2009). Ventral striatal neurons encode the value of the chosen action in rats deciding between differently delayed or sized rewards. *J Neurosci, 29*(42), 13365-13376. doi: 10.1523/JNEUROSCI.2572-09.2009

Rothwell, J. C. (2011). The motor functions of the basal ganglia. *J Integr Neurosci, 10*(3), 303-315.

Rushworth, M. F., Noonan, M. P., Boorman, E. D., Walton, M. E., & Behrens, T. E. (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron, 70*(6), 1054-1069. doi: 10.1016/j.neuron.2011.05.014

Rushworth, M. F., Walton, M. E., Kennerley, S. W., & Bannerman, D. M. (2004). Action sets and decisions in the medial frontal cortex. *Trends Cogn Sci, 8*(9), 410-417. doi: 10.1016/j.tics.2004.07.009

Rutledge, R. B., Dean, M., Caplin, A., & Glimcher, P. W. (2010). Testing the reward prediction error hypothesis with an axiomatic model. *J Neurosci, 30*(40), 13525-13536. doi: 10.1523/JNEUROSCI.1747-10.2010

Saleem, K. S., Kondo, H., & Price, J. L. (2008). Complementary circuits connecting the orbital and medial prefrontal networks with the temporal, insular, and opercular cortex in the macaque monkey. *Journal of Comparative Neurology, 506*(4), 659-693. doi: Doi 10.1002/Ene.21577

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science, 310*(5752), 1337-1340. doi: 10.1126/science.1115270

Saper, C. B., Iversen, S., & Frackowiack, R. (2000). Integration of sensory and motor function: The association areas of the cerebral cortex and the cognitive capabilities of the brain. *Principles of Neural Science, 4th Edition*: New York: McGraw-Hill.

Schoenbaum, G., Takahashi, Y., Liu, T. L., & McDannald, M. A. (2011). Does the orbitofrontal cortex signal value? *Ann N Y Acad Sci, 1239*, 87-99. doi: 10.1111/j.1749-6632.2011.06210.x

Schultz, W. (2000). Multiple reward signals in the brain. *Nat Rev Neurosci, 1*(3), 199-207. doi: 10.1038/35044563

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

Seger, C. A. (2008). How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. *Neuroscience and Biobehavioral Reviews, 32*(2), 265-278. doi: DOI 10.1016/j.neubiorev.2007.07.010

Seo, H., & Lee, D. (2007). Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. *J Neurosci, 27*(31), 8366-8377. doi: 10.1523/JNEUROSCI.2369-07.2007

Serences, J. T. (2008). Value-based modulations in human visual cortex. *Neuron, 60*(6), 1169-1181. doi: 10.1016/j.neuron.2008.10.051

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *J Neurophysiol, 86*(4), 1916-1936.

Shallice, T. (1982). Specific Impairments of Planning. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 298*(1089), 199-209. doi: DOI 10.1098/rstb.1982.0082

Shallice, T., & Burgess, P. (1996). The domain of supervisory processes and temporal organization of behaviour. *Philos Trans R Soc Lond B Biol Sci, 351*(1346), 1405-1411; discussion 1411-1402. doi: 10.1098/rstb.1996.0124

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*(4), 623-656.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron, 79*(2), 217-240. doi: 10.1016/j.neuron.2013.07.007

Sigman, M., & Dehaene, S. (2008). Brain mechanisms of serial and parallel processing during dual-task performance. *J Neurosci, 28*(30), 7585-7598. doi: 10.1523/JNEUROSCI.0948-08.2008

Skatova, A., Chan, P. A., & Daw, N. D. (2013). Extraversion differentiates between model-based and model-free strategies in a reinforcement learning task. *Front Hum Neurosci, 7*, 525. doi: 10.3389/fnhum.2013.00525

Smith, E. E., & Jonides, J. (1999). Storage and executive processes in the frontal lobes. *Science, 283*(5408), 1657-1661.

Smith, S. M. (2004). Overview of fMRI analysis. *Br J Radiol, 77 Spec No 2*, S167-175.

Smittenaar, P., FitzGerald, T. H., Romei, V., Wright, N. D., & Dolan, R. J. (2013). Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron, 80*(4), 914-919. doi: 10.1016/j.neuron.2013.08.009

Smittenaar, P., Prichard, G., FitzGerald, T. H., Diedrichsen, J., & Dolan, R. J. (2014). Transcranial direct current stimulation of right dorsolateral prefrontal cortex does not affect model-based or model-free reinforcement learning in humans. *PLoS One, 9*(1), e86850. doi: 10.1371/journal.pone.0086850

Sokol-Hessner, P., Hutcherson, C., Hare, T., & Rangel, A. (2012). Decision value computation in DLPFC and VMPFC adjusts to the available decision time. *Eur J Neurosci, 35*(7), 1065-1074. doi: 10.1111/j.1460-9568.2012.08076.x

Stalnaker, T. A., Calhoon, G. G., Ogawa, M., Roesch, M. R., & Schoenbaum, G. (2010). Neural correlates of stimulus-response and response-outcome associations in dorsolateral versus dorsomedial striatum. *Front Integr Neurosci, 4*, 12. doi: 10.3389/fnint.2010.00012

Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. *Neuron, 78*(2), 364-375. doi: 10.1016/j.neuron.2013.01.039

Strait, C. E., Blanchard, T. C., & Hayden, B. Y. (2014). Reward Value Comparison via Mutual Inhibition in Ventromedial Prefrontal Cortex. *Neuron, 82*(6), 1357-1366. doi: 10.1016/j.neuron.2014.04.032

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643-662. doi: Doi 10.1037/0096-3445.121.1.15

Strother, S. C. (2006). Evaluating fMRI preprocessing pipelines - Review of preprocessing steps for BOLD fMRI. *Ieee Engineering in Medicine and Biology Magazine, 25*(2), 27-41. doi: Doi 10.1109/Memb.2006.1607667

Sun, R. (2008). Introduction to Computational Cognitive Modeling. *Cambridge Handbook of Computational Psychology*, 3-19. doi: Book_Doi 10.1017/Cbo9780511816772

Sutton, R. S. (1988). Learning to Predict by the Methods of Temporal Differences. *Machine Learning, 3*, 9-44.

Sutton, R. S. B., A. G. (1998). *Reinforcement Learning: An Introduction*: MIT Press, Cambridge, Massachusetts.

Takahashi, Y. K., Chang, C. Y., Lucantonio, F., Haney, R. Z., Berg, B. A., Yau, H. J., . . . Schoenbaum, G. (2013). Neural estimates of imagined outcomes in the orbitofrontal cortex drive behavior and learning. *Neuron, 80*(2), 507-518. doi: 10.1016/j.neuron.2013.08.008

Takechi, H., Onoe, H., Shizuno, H., Yoshikawa, E., Sadato, N., Tsukada, H., & Watanabe, Y. (1997). Mapping of cortical areas involved in color vision in non-human primates. *Neurosci Lett, 230*(1), 17-20.

Tanaka, S. C., Balleine, B. W., & O'Doherty, J. P. (2008). Calculating consequences: brain systems that encode the causal effects of actions. *J Neurosci, 28*(26), 6750-6755. doi: 10.1523/JNEUROSCI.1808-08.2008

Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., & Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nat Neurosci, 7*(8), 887-893. doi: 10.1038/nn1279

Tanji, J., & Hoshi, E. (2008). Role of the lateral prefrontal cortex in executive behavioral control. *Physiol Rev, 88*(1), 37-57. doi: 10.1152/physrev.00014.2007

Tekin, S., & Cummings, J. L. (2002). Frontal-subcortical neuronal circuits and clinical neuropsychiatry: an update. *J Psychosom Res, 53*(2), 647-654.

Thorndike, E. L. (1911). *Animal intelligence; experimental studies*: New York, The Macmillan company.

Toates, F. M. (1986). *Motivational Systems: Problems in the behavioral sciences*: Cambridge University Press, New York.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychol Rev, 55*(4), 189-208.

Tricomi, E. M., Delgado, M. R., & Fiez, J. A. (2004). Modulation of caudate activity by action contingency. *Neuron, 41*(2), 281-292.

Valentin, V. V., Dickinson, A., & O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *J Neurosci, 27*(15), 4019-4026. doi: 10.1523/JNEUROSCI.0564-07.2007

van den Heuvel, O. A., Groenewegen, H. J., Barkhof, F., Lazeron, R. H., van Dyck, R., & Veltman, D. J. (2003). Frontostriatal system in planning complexity: a parametric functional magnetic resonance version of Tower of London task. *Neuroimage, 18*(2), 367-374.

van der Meer, M., Kurth-Nelson, Z., & Redish, A. D. (2012). Information processing in decision-making systems. *Neuroscientist, 18*(4), 342-359. doi: 10.1177/1073858411435128

Vogt, B. A., Rosene, D. L., & Pandya, D. N. (1979). Thalamic and Cortical Afferents Differentiate Anterior from Posterior Cingulate Cortex in the Monkey. *Science, 204*(4389), 205-207. doi: DOI 10.1126/science.107587

Voon, V., Derbyshire, K., Ruck, C., Irvine, M. A., Worbe, Y., Enander, J., . . . Bullmore, E. T. (2014). Disorders of compulsivity: a common bias towards learning habits. *Mol Psychiatry*. doi: 10.1038/mp.2014.44

Waldron, E. M., & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: evidence for multiple category learning systems. *Psychon Bull Rev, 8*(1), 168-176.

Walker, A. E. (1940). A cytoarchitectural study of the prefrontal area of the macaque monkey. *Journal of Comparative Neurology, 73*(1), 59-86. doi: DOI 10.1002/cne.900730106

Wallis, J. D., & Miller, E. K. (2003). Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *Eur J Neurosci, 18*(7), 2069-2081.

Walton, M. E., Croxson, P. L., Behrens, T. E., Kennerley, S. W., & Rushworth, M. F. (2007). Adaptive decision making and value in the anterior cingulate cortex. *Neuroimage, 36 Suppl 2*, T142-154. doi: 10.1016/j.neuroimage.2007.03.029

Waskom, M. L., Kumaran, D., Gordon, A. M., Rissman, J., & Wagner, A. D. (2014). Frontoparietal representations of task context support the flexible control of goal-directed cognition. *J Neurosci, 34*(32), 10743-10755. doi: 10.1523/JNEUROSCI.5282-13.2014

Williams, Z. M., Bush, G., Rauch, S. L., Cosgrove, G. R., & Eskandar, E. N. (2004). Human anterior cingulate neurons and the integration of monetary reward with motor responses. *Nat Neurosci, 7*(12), 1370-1375. doi: 10.1038/nn1354

Wise, S. P. (2008). Forward frontal fields: phylogeny and fundamental function. *Trends Neurosci, 31*(12), 599-608. doi: 10.1016/j.tins.2008.08.008

Wunderlich, K., Dayan, P., & Dolan, R. J. (2012). Mapping value based planning and extensively trained choice in the human brain. *Nat Neurosci, 15*(5), 786-791. doi: 10.1038/nn.3068

Wunderlich, K., Rangel, A., & O'Doherty, J. P. (2009). Neural computations underlying action-based decision making in the human brain. *Proc Natl Acad Sci U S A, 106*(40), 17199-17204. doi: 10.1073/pnas.0901077106

Wunderlich, K., Smittenaar, P., & Dolan, R. J. (2012). Dopamine enhances model-based over model-free choice behavior. *Neuron, 75*(3), 418-424. doi: 10.1016/j.neuron.2012.03.042

Yildiz, A., & Beste, C. (2014). Parallel and serial processing in dual-tasking differentially involves mechanisms in the striatum and the lateral prefrontal cortex. *Brain Struct Funct*. doi: 10.1007/s00429-014-0847-0

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nat Rev Neurosci, 7*(6), 464-476. doi: 10.1038/nrn1919

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *Eur J Neurosci, 19*(1), 181-189.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2005). Blockade of NMDA receptors in the dorsomedial striatum prevents action-outcome learning in instrumental conditioning. *Eur J Neurosci, 22*(2), 505-512. doi: 10.1111/j.1460-9568.2005.04219.x

Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *Eur J Neurosci, 22*(2), 513-523. doi: 10.1111/j.1460-9568.2005.04218.x

Yoshida, W., & Ishii, S. (2006). Resolution of uncertainty in prefrontal cortex. *Neuron, 50*(5), 781-789. doi: 10.1016/j.neuron.2006.05.006

Zhang, J., Kriegeskorte, N., Carlin, J. D., & Rowe, J. B. (2013). Choosing the rules: distinct and overlapping frontoparietal representations of task rules for perceptual decisions. *J Neurosci, 33*(29), 11852-11862. doi: 10.1523/JNEUROSCI.5193-12.2013