# Criteria and awareness in perceptual decision

**Stephen Michael Fleming**

Wellcome Trust Centre for Neuroimaging
Institute of Neurology
University College London

Dissertation submitted for the degree of
Doctor of Philosophy
of
University College London

June 2011

# Declaration

I, Stephen Michael Fleming, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

June 20, 2011

# Abstract

The immediacy of subjective experience belies the complex process of inference and categorisation that our brains undertake every moment of our waking lives, a process that allows the selection of the best course of action in the face of under-determined sensory input. There is much behavioural evidence that humans use the context in which decisions occur to actively shape links between perception and action. However, there are several remaining questions as to how this process occurs in the brain, and how such decision-making is linked to subjective reports, four of which are addressed in this thesis. It is unknown at which stage along the path from sensory to motor areas a loss function is integrated into the perceptual decision process. Using fMRI I show that asymmetries in value affect a fronto-parietal-basal ganglia network, rather than impacting upon the coding of visual categories. Theoretical models predict that the basal ganglia adjust the link between decision and action on the basis of contextual variables, but supporting empirical evidence is scarce. In two further imaging studies I show that the subthalamic nucleus modulates action control when default expectations are violated. That links between perception and action may be labile leads one to ask to what extent the observer has metacognitive access to these stages of the decision process, and which brain structures might mediate this access. I show that a second-order signal detection model can capture some, but not all, features of metacognitive confidence. Finally, I show that individual differences in metacognitive ability are associated with the structure of anterior prefrontal cortex. Comparing the levels of perceptual and metacognitive decision is critical for understanding how the mechanisms of decision-making are linked to awareness and self-report. The thesis concludes with a brief discussion of future challenges in this direction.

# Acknowledgements

It is very hard to reduce the laughter and learning I have shared with the people surrounding me over the past four years into a few paragraphs of acknowledgements. The FIL has become a second home, and its members have taught me the privilege and pleasure of doing good science. My supervisor, Ray Dolan, has given me both freedom and guidance, and imparted his wisdom on the finer things in life and academia. My second supervisor, Chris Frith, has gently nudged me down interesting avenues without me always being aware of it. I have been additionally lucky to study for a PhD in a department at the cutting edge of theory in the neurosciences: Karl Friston has slowly but surely turned my psychologist's conception of mind upside down, and done so with grace and kindness.

The collaborative opportunities I have stumbled upon have been second-to-none: a chance discussion over strong beer in a Swiss monastery led to a wonderful first project conducted with Ollie Hulme and Louise Whiteley. In Ollie I found a patient and knowledgable teacher in all things SPM, and now have a friend with a shared passion for consciousness and shared confusion for free energy. Working closely with my friend Charlotte Thomas during her medical elective project on the default bias, and Rimona Weil and Geraint Rees on metacognition has been in equal parts energising and enjoyable. The rotation project I carried out in Patrick Haggard's lab has led to a fruitful collaboration and exchange of ideas over the years, a relationship that I hope will continue.

My office mates in the second-floor front room have been a pleasure to work with. The FIL occasional XI have provided competition on the football pitch, and first-class banter off of it. Ray's group has provided a constant source of ideas and friendship; in particular, the cottage crew – Sara Bengtsson, Marc Guitart-Masip, Rosalyn Moran, Tali Sharot, Tamara Shiner, Bryan Strange and Nick Wright. Mkael Symmonds has been both a patient sounding board and a good friend during my time here. Nick – you've made my time at the FIL immeasurably more fun. I hate to imagine an alternate universe where our PhD studies didn't collide. I'll miss our gossips cunningly disguised as gym sessions, and I selfishly hope that I'll be able to continue bugging you for many years to come. Tali – the highs and lows of this PhD were shared most of all by you.

To my housemates, I'd like to say you keep me sane but it's probably not the case. Suffice to say you've kept me smiling, and reminded me of the importance of both friendship and life outside academia. To my family, thank you for your support in whatever odd scheme for a career change I dreamt up, even if you secretly knew it wouldn't last. Your love and kindness is a constant source of strength.

Finally, this thesis would not have been possible without my undergraduate tutor at Oxford, Paul Azzopardi. His introductory course in signal detection theory and consciousness given over eight exhilarating weeks in Hilary Term 2006 made me realise that combining psychophysics with philosophy was possible. His influence permeates each chapter.

Thanks are due to Ōiwi Parker Jones for his crash course in LaTeX, and to Tom Fitzgerald for his assistance with my methods chapter. Marina Anderson has helped me navigate each stage of thesis-related bureaucracy, and more besides. For reading part or all of this thesis in various forms, I would like to thank Ray Dolan, Sven Bestmann, Patrick Haggard, Karl Friston, Chris Frith, Zoltan Nagy, Ged Ridgway and Tali Sharot.

# Contents

# List of Figures

# List of Tables

# Contributions

The work reported in this thesis is entirely my own unless otherwise indicated. All chapters have benefited from guidance and advice from my supervisor, Prof. Ray Dolan. The experiment reported in Chapter 4 was a result of a collaboration between myself, Dr. Louise Whiteley and Dr. Maneesh Sahani of the Gatsby Unit for Computational Neuroscience, and Dr. Oliver Hulme of the Institute for Opthalmology. The project built upon a paradigm developed by Louise as part of her PhD (Whiteley & Sahani, 2008). Louise developed the Bayesian model comparison software and methods used to analyse the psychophysics data reported in this chapter. I collected behavioural and fMRI data, and led the project. Louise, Oliver and I contributed to the implementation and interpretation of the fMRI analysis.

The experiment reported in Chapter 5 was partly based on Dr. Charlotte Thomas' medical elective project co-supervised by myself and Prof. Ray Dolan. Charlotte carried out behavioural data collection and analysis during development of the paradigm for the scanner. The work reported in Chapter 6 was a result of a collaboration between myself and Dr. Christian Lambert at the Wellcome Trust Centre for Neuroimaging. Christian developed methods for localising the subthalamic nucleus on high-resolution anatomical scans, allowing extraction and analysis of fMRI signal from this region on a subject-by-subject basis. I designed the behavioural paradigm, and carried out data collection and imaging analysis.

Chapters 7 and 8 were a result of a collaboration between myself and Dr. Zoltan Nagy at the Wellcome Trust Centre for Neuroimaging, and Dr. Rimona Weil and Prof. Geraint Rees at the Institute for Cognitive Neuroscience. Rimona and I collaborated closely on all aspects of data collection and analysis. Zoltan advised on DTI acquisition and developed the artefact correction routines applied in Chapter 8. The signal detection model and additional analysis of reaction time data is my own.

# Publications during the PhD

Chapter 4 has been published in *Journal of Neurophysiology* (Fleming *et al.*, 2010c). Chapter 5 has been published in *Proceedings of the National Academy of Sciences* (Fleming *et al.*, 2010a). A further collaboration based on the paradigm presented in Chapter 5 is in press at *Journal of Neuroscience* (Nicolle *et al.*, 2011).

Data from Chapters 7 and 8 have been published in *Science* (Fleming *et al.*, 2010b), and a behavioural study of metacognitive ability that builds on this work carried out by Ms. Chen Song is in press at *Consciousness and Cognition* (Song *et al.*, 2011). The signal detection model presented in Chapter 7 is derived in part from an analysis of post-decision wagering published in *Consciousness and Cognition* (Fleming & Dolan, 2010). Part of the discussion in Chapter 9 is drawn from a commentary I wrote on Liston & Stone (2008) in *Frontiers in Human Neuroscience* (Fleming, 2009). I have also maintained close collaboration with Prof. Patrick Haggard at the Institute for Cognitive Neuroscience, resulting in two papers based on work carried out prior to my PhD (Fleming *et al.*, 2009; Wenke *et al.*, 2010).

# Chapter 1

# Introduction

## 1.1 Conceptual overview

Decision-making is usually assumed to involve concerted deliberation, such as choosing which job to take, or which café to go to for lunch. Darwin famously engaged in this type of effortful decision-making when choosing whether to marry, by weighing up the pros and cons of having a wife[1]. But from the moment we open our eyes in the morning and embark upon a continuous series of glances, looks, categorisations, judgements, misjudgements and deliberations, we are making decisions, both automatic and effortful. Our brain is able to classify a voice as male or female, a face in the crowd as being a familiar or unfamiliar, without our conscious 'self' necessarily having insight into this process (Pylyshyn, 2003). Here we adopt this broader definition of decision, one that allows us to probe how the brain multiplexes various sources of 'evidence' to guide our everyday behaviour. This thesis addresses the relationship between perception – 'seeing' – and action – 'doing', and one's awareness of this relationship. By analysing the neural basis of simple visual decisions both in terms of first-order decision-making, and second-order commentaries on this decision process, I aim to form bridges between mechanism on the one hand, and subjectivity on the other.

The first part of the thesis examines how sense data (visual input) interacts with context to bias connections between stimuli and actions in the brain. The role of the basal ganglia, a collection of nuclei that play a key role in the control of action, forms the focus of these chapters. The second part of the thesis then asks to what extent we are able to access and report aspects of the relationship between perception and action, and how this ability is linked to individual differences in brain structure. On one view, introspection is only weakly linked to the underlying characteristics of the decision process (Nisbett & Wilson, 1977; Johansson *et al.*, 2005). In contrast, data reported here show that under the right testing conditions, subjects are able to report the fine-grained and graded nature of processes underlying their decision-

[1] He eventually decided the pros outweighed the cons.

**Figure 1.1:** England appeal for the dismissal of South African batsman Graeme Smith in a 2010 Test match.

making. How this access arises from the mechanisms that underpin 'first-order' perceptual decisions is a question that will be pondered at multiple junctures in this thesis.

Despite perceptual decisions being based on information arriving from the outside world, the context in which we receive sensory information fundamentally affects our judgments, and how these judgments connect to action. Biases in the link between perception and action are most noticeable when decisions are made under uncertainty. Intuitively, when the correct decision is clearly specified by incoming sense data, the context in which the decision is made plays little role in the outcome. Alternatively, when the data only weakly specifies one course of action over another, *a priori* knowledge of the costs and benefits of choosing one response over another are given more weight (Kersten *et al.*, 2004). This can be illustrated in a simple example. The umpire at a Test cricket match is often called upon to make a perceptual judgment under uncertainty, such as whether the ball struck the bat or not following an appeal for 'caught behind' (see figure 1.1). If the sensory input is equivocal, the influence of the crowd, the state of the game, and the number of times a similar incident has occurred up to this point may sway his or her decision[2]. In other words, extraneous factors impinge upon the decision process, biasing the action taken. If the evidence that the batsman hit the ball were stronger, leading to greater certainty in the mind of the umpire, these factors can be assumed to carry less weight when arriving at a decision.

The combined influence of prior beliefs (how likely is it, on average, that a batsman nicks the ball), and costs and benefits (will giving the batsman 'out' significantly change the course of the game?) is known as the 'loss function'. As indicated in figure 1.2, the mechanisms through which the loss function may affect perceptual

[2] Such effects have been quantified for Premiership football referees, and indeed show a significant bias away from penalising the home side (Boyko *et al.*, 2007).

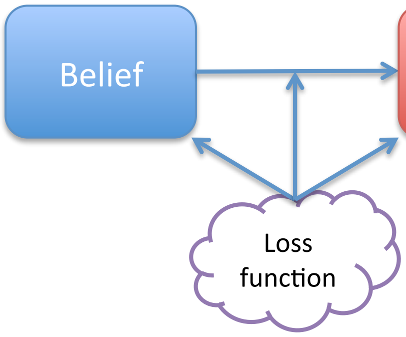

**Figure 1.2:** Cartoon of how the loss function could be applied to various stages of the perceptual decision process. Bias could either affect an early sensory stage, prior to response specification, a decision stage intermediate between stimulus and response, or a post-decision stage at the level of response specification.

decisions under uncertainty are unknown, and several potential systems-level models have been articulated (Maloney, 2002). On the one hand, the loss function could shape early sensory processing, leading to a change at the input stage. On the other hand, beliefs about the stimulus could be constructed in relative isolation from the loss function, before adjustments in the decision are made during action selection. In Chapter 4 we explore this question, concluding that the asymmetries in costs affect cortico-basal ganglia circuitry, rather than early perceptual mechanisms, consistent with the latter view.

As we will see in Chapter 5, the weight attached to the loss function is inversely proportional to the subject's current level of uncertainty about the correct answer (see also section 2.2.4). However, the extent to which subjects are aware of (i.e. able to report) their fluctuating level of uncertainty and/or interactions with the loss function during perceptual decision-making is unknown. In the second part of this thesis (Chapters 7 and 8), I investigate the extent to which subjects are able to communicate their level of uncertainty about their own decision process. To the extent that this communication is accurate, it provides convergent evidence that the brain encodes and use decision uncertainty to shape ongoing behaviour. In Chapter 7 I extend simple models of perceptual decision-making used to analyse behavioural data in the first part of the thesis to encompass metacognitive judgments. In Chapter 8 I go on to apply this analysis to investigate how metacognitive sensitivity is related to brain structure.

## 1.2 Unifying principles

### 1.2.1 Bayesian decision theory

Imagine you are a radar operator at the height of the Cold War. Tensions are high, and there is a real sense of danger from unannounced enemy missiles. Your job is to detect the earliest, faint visual signals of these objects on your screen. Often you have only the merest glance, a fleeting visual impression of the signal, yet have to decide whether the blip you just saw was truly a missile or not. How does your brain compute the right course of action?

Signal detection theory (Green & Swets, 1966), a special case of Bayesian decision theory (see section 2.2.1), provides us with a natural way of thinking about forced-choice decision scenarios such as these, giving a prescriptive framework for converting single observations of noisy evidence into a categorical choice. Like deciding whether to accept or reject a scientific hypothesis (Neyman & Pearson, 1933), a trade-off needs to be made between saying something is true when it is not (a false positive) and failing to identify something as true when it is (a false negative). Deciding whether to convert instantaneous sensory evidence into the decision of a missile, or some other harmless object such as a flock of birds, requires adoption of a decision criterion. By assuming certain factors about observers' internal representation of sensory evidence, signal detection theory can both proscribe how this criterion should be set (Macmillan & Creelman, 2005), and provide an experimental measure of its usage.

The decision criterion is known to be affected by the loss function (Gold & Shadlen, 2007), such that two contextual influences impinge upon the decision-making process. The first is the respective value of particular outcomes: in a wartime context, the repercussions for missing a potential missile firing are catastrophic, whereas in peacetime, the costs associated with this outcome are likely to be downgraded. Second, prior beliefs affect the inference process: is one area of the screen (such as the location of enemy airspace) more likely to generate missiles than any other area?

In Chapters 4-6, I investigate changes in decision thresholds induced by action costs and prior probabilities for simple perceptual decisions. I will review the literature specific to the neural mechanisms underlying perceptual decision-making in Chapter 2. Briefly, competing theories place the basal ganglia, and associated cortical afferents, at the centre of a network adjusting a threshold for action initiation on the basis of decision context (Bogacz, 2007; Lo & Wang, 2006). However, the implementation of these adjustments in the brain remains unclear (Bogacz *et al.*, 2010).

### 1.2.2 Higher-order awareness

We regularly engage in higher-order reflection on our thoughts, memories and perceptions, or 'thinking about thinking'. Here I use the terms 'second-order' and 'metacognitive' interchangeably to refer to (self-referent) cognition about a first-order mental state that has the potential to be communicated to others (Jack & Roepstorff, 2002). Humans (and possibly some animal species; Smith *et al.* 2004) are able to use metacognitive commentaries to communicate confidence in their own decision process (Flavell, 1979; Kunimoto *et al.*, 2001; Persaud *et al.*, 2007). Furthermore, explicit knowledge of uncertainty can optimise performance on a task. For instance, the opportunity to 'opt-out' of a decision is often given in multiple-choice exams; if there is a penalty for wrong answers then knowing you do not know is highly beneficial (Metcalfe, 1996; Higham *et al.*, 2009).

The accuracy of metacognitive commentaries suggests awareness of the antecedents of the decision process. Several researchers have used variants of metacognitive report to measure self-knowledge about a decision process, in particular, the ability to discriminate correct from incorrect decisions using confidence ratings (Dienes & Seth, 2010; Evans & Azzopardi, 2007; Kunimoto *et al.*, 2001), wagers (Persaud *et al.*, 2007) or signalling of errors (Ullsperger *et al.*, 2010). Second-order judgments can be usefully analysed within an extension of SDT known as Type 2 signal detection theory (Type 2 SDT; Clarke *et al.* 1959). In Type 2 SDT, the 'evidence' which is being discriminated is the subjects own mental state, rather than sensory evidence in the world. For post-decision confidence, a 'hit' is then a high-confidence correct judgment, and a 'false alarm' is a high-confidence incorrect judgment (see section 3.2). Little is known about the psychological and neural processes underlying metacognitive judgments of perceptual decision-making.

## 1.3 Outline of the thesis

First, I review the literature on both perceptual (first-order) and metacognitive (second-order) decision-making in Chapter 2. Chapter 3 provides background on general methodology, allowing each subsequent experimental Chapter to take up the baton with specifics on methods used.

The empirical work contained in this thesis is roughly divided into two interconnected halves. The first, in Chapters 4-6, presents studies characterising the psychological and neural effects of changes in decision criteria during human perceptual decision-making using psychophysics and functional magnetic resonance imaging (fMRI). Chapter 4 asks at what stage in the perceptual decision system does a loss function exert its effects (cf. figure 1.2). Next, Chapters 5 and 6 'zoom in' on basal ganglia mechanisms hypothesised to control the threshold for action initiation under asymmetric priors. This work isolates novel prefrontal and basal ganglia mechanisms that may play a role in setting decision criteria, with a particular focus on the STN.

The second half of the thesis builds a paradigm for investigating (second-order) introspective assessments about perceptual performance. Chapter 7 presents a theoretical framework for analysing second-order decisions, and applies these methods to the analysis of confidence rating data collected in the context of uncertain perceptual decisions. In brief, the method I outline holds objective decision uncertainty constant across subjects, and asks to what extent subjects can access local fluctuations in performance using a subjective confidence scale. I ask whether a 'bottom-up' model based on first-order accounts of decision-making is enough to explain metacognitive ability, and conclude that there exists variance in metacognitive reports that cannot be explained by a 'direct translation' approach alone (cf. Higham *et al.* 2009). Chapter 8 then applies second-order decision analysis to investigate the neural basis of this partially separable metacognitive component using voxel-based morphometry (VBM) and diffusion-tensor imaging (DTI). Finally, the discussion (Chapter 9) draws together the two halves of the thesis, examining connections between first- and second-order decision-making. The computational and psychological implications of these mechanistic insights are discussed.

# Chapter 2

# Literature review

## 2.1 Overview

The studies within this thesis are situated at the interface of perception and action. In this chapter, I begin with an outline of normative theoretical concepts that can be considered common to all types of decisions. I then review recent work on the neural implementation of perceptual decision-making, with a particular focus on work that has examined interactions between visual information and contextual biases induced by prior beliefs and asymmetric costs.

Post-decision, or metacognitive, commentaries provide insight into the psychological structure of links between perception and action The concepts and models that are used to analyse the decision itself can also be applied to metacognitive commentaries, casting the latter as ‘decisions about decisions’. In section 2.4 I review behavioural studies that have quantified metacognitive decision-making, and the few recent studies that have begun to examine the neural correlates of second-order decision processes. I conclude by indicating how the work contained in this thesis is positioned to address outstanding questions in the field.

## 2.2 Statistical decision theory

At any moment in time, the brain is being bombarded with noisy neural signals from the outside world. It must use this shadowy impression of reality to guide behaviour and, in so doing, maximising the likelihood of its continued existence. This is a formidable problem, as action guidance depends on perceptual inference (Kersten *et al.*, 2004). If we misperceive a poisonous plant as a tasty snack, then we will act upon it as if it was a tasty snack. There is no way of knowing otherwise. This, in a nutshell, is the problem of decision-making and inference: going from the data (the noisy signals), back to its cause (the species of plant), and then to the selection of an appropriate action (eating or avoiding). Many researchers agree that solving this problem requires a statistical approach (Doya *et al.*, 2007)[1], which is reviewed next.

### 2.2.1 Bayes' rule

The basic idea behind Bayes' theorem was presented in section 1.2.1. The core insight is that the probability of the cause given a piece data can be inferred from the likelihood of the data arising from a particular cause (MacKay, 2003). Consider a random variable $x$ that encodes the evidence supporting a particular choice. For example, if an observer is discriminating whether a dim light is present or absent, $x$ might be the average firing of his or her retinal ganglion cells. If this firing is related to the actual state of the world (present or absent), then it should covary with this state; it is on average higher if the light is actually present. The key qualifier here is 'on average': $x$ is corrupted by noise at all stages of signal transduction, and therefore is described by a distribution, rather than a deterministic transition. If we consider the presence or absence of the light to be indicated by two hypotheses, $h_1$ and $h_2$, the probability that each gave rise to the data $x$ is equal to:

$$\begin{aligned} p(h_1|x) &= \frac{p(x|h_1).p(h_1)}{p(x)} \\ p(h_2|x) &= \frac{p(x|h_2).p(h_2)}{p(x)} \end{aligned} \tag{2.1}$$

These equations specify the *posterior* probability of the state of the world ($h_1$ or $h_2$):

$$\text{posterior} \propto \text{likelihood} \times \text{prior} \tag{2.2}$$

On this view, *perception* is the computation of a posterior belief in the state of the world given incoming sense data. But this computation is not the end of the story – brains are not passive perceivers, but instead function to maximise their chances of survival in a changing world (Smith, 1982). Achieving this requires *action* – given a particular belief that the plant in front of me is harmless, I should probably eat it in order not to go hungry. Converting a posterior belief into action requires integration of a loss function (Kording, 2007; Berger, 1985; Davison & Tustin, 1978), which summarises the costs and benefits of each possible decision outcome. Contextual biasing of perception has broad historical precedent in psychology, beginning with the 'New Look' school of the 1950s (Bruner & Goodman, 1947; Bruner, 1957).

[1] I will not be concerned with proving or disproving what is known as the 'Bayesian brain hypothesis' – the suggestion that the brain faithfully implements Bayesian inference – in this thesis (cf. Whiteley 2009). Instead I use Bayes as an organising framework for the link between perception and action, and note that much of the empirical work presented in the subsequent chapters can be interpreted in strictly psychological, rather than computational, terms.

These studies emphasised the role of 'needs and desires' in altering perception, and have been echoed in recent work showing that, for example, being motivated to receive a particular outcome leads to perceptual biases (Balcetis & Dunning, 2006; Changizi & Hall, 2001). That perceptual decision processes are affected by changes in context has been repeatedly confirmed by experimentation in psychology (Balcetis & Dunning, 2006; Bohil & Maddox, 2001; Johnstone & Alsop, 1996, 2000; Maddox & Bohil, 2003; Proshansky & Murphy, 1942). Indeed, as Balcetis & Dunning (2007) note, 'the time might be ripe to explore [the New Look] hypotheses with theories and methods that are more nuanced and sophisticated than what was available 50 years ago'.

### 2.2.2 Decision criteria

I now consider a simple choice between two alternatives (such as a stimulus being present or absent) in order to illustrate the concept of a decision criterion. Under flat, or uninformative, priors, the posterior weight of evidence favouring decisions $h_1$ and $h_2$ can be computed as the likelihood ratio ($l$):

$$l = \frac{p(x|h_1)}{p(x|h_2)} \tag{2.3}$$

The Neyman-Pearson lemma states that the likelihood ratio is the optimal (most sensitive) test to apply when choosing between two alternatives under uncertainty (Neyman & Pearson, 1933). However, this ratio can take on one of many continuous values, and does not yet specify which of the two decisions the observer should take. Instead, a decision rule (criterion) needs to be applied to the likelihood ratio, which I denote $\beta$. The observer makes the choice of $h_1$ when $l > \beta$, and $h_2$ otherwise. For example, when testing scientific hypotheses, the rule often used is to accept the alternative hypothesis if the weight of evidence is greater than 20:1, or $\alpha < 0.05$. Here, if the goal is to maximise the accuracy of the choice, we should choose $h_1$ when $\beta > 1$, the point at which the evidence supporting each hypothesis is equal.

Except in certain constrained psychophysics experiments, the goal is not usually to maximise accuracy. Instead, the loss function is used to adjust the decision criterion to maximise the potential reward to the decision-maker given knowledge of the costs and benefits of each course of action, subsuming, for instance, the motivational factors that might impinge on our umpire in figure 1.1. The optimal value of the likelihood ratio to maximise expected reward weights the values and probabilities of each outcome (hypothesis) in the following fashion (Green & Swets, 1966; Dayan & Daw, 2008):

$$\beta_0 = \frac{V(h_2, r_2) - V(h_2, r_1)}{V(h_1, r_1) - V(h_1, r_2)} . \frac{P(h_2)}{P(h_1)} \tag{2.4}$$

where $V(h_i, r_j)$ indicates the value of response $r_j$ made when the true state of the world was $h_i$.

Note that decision theory is agnostic as to how values and priors are dealt with in the algorithms underlying the decision process (Maloney 2002; figure 1.2). Indeed, a recent theoretical suggestion holds that cost functions and priors perform a common role in Bayesian inference (Friston *et al.*, 2006). In other words, costs may be implemented 'as if' they were prior expectations (and vice versa). On one view, the expected value of potential outcomes is taken into account when computing a posterior belief (equation 2.1). Alternatively, the expected value can affect a decision stage beyond coding of the sensory likelihood (equation 2.4). In other words, values and priors could affect perception, or affect a post-perceptual stage (see Liston & Stone 2008; Summerfield & Koechlin 2010 and Chapter 4 for experiments bearing on this question), while remaining faithful to the ideal observer model outlined above. In the former case, we might expect the individual to resolve a belief state[2] about the cause of his or her sense data (the movement of the ball after hitting the bat in figure 1.1), before taking into account asymmetries in the value of each potential decision and acting accordingly (Henderson & Hollingworth, 1999; Lu & Dosher, 2008). On the latter view, a belief in a particular state of the world is itself adjusted by the value of that state; action is then concerned with communicating the decision with the strongest supporting belief. The crucial fact here is that while behaviour would remain the same in both cases, the internal state variables of our observer would differ.

One way of arbitrating between the two hypotheses is through use of functional brain imaging. The logic here is that biases to belief states should manifest as changes in the activity profile of sensory regions known to be sensitive to particular perceptual categories, such as faces and houses (Kanwisher *et al.*, 1997; Grill-Spector *et al.*, 2001).

### 2.2.3 Sensitivity and bias

Signal detection theory (SDT) is a special case of a more general Bayesian inference scheme, and connects concepts of the likelihood ratio and criterion to equations that can easily be applied to the analysis of behavioural data (see Chapter 3). I will make use of SDT in the analysis of behavioural and brain imaging data in Chapters 4, 5, 7 and 8. Classic psychophysical paradigms for investigating sensitivity and bias ask observers to complete several forced-choice judgments, such as whether a visual stimulus is present or absent. The difficulty of this judgment is adjusted to induce uncertainty in the observer. Initial work on validating SDT measures confirmed that for both visual and auditory stimuli, sensitivity ($d'$) is independent of an observer's criterion ($c$), as predicted by theory (Green & Swets, 1966; Macmillan & Creelman, 2005). More generally, the goodness-of-fit of the receiver operating characteristic (ROC) model in a wide range of psychophysics experiments provides strong support

[2] We remain agnostic as to the *reportability* of this belief state.

for the existence of underlying posterior distributions of evidence values to which a changing threshold or criterion is applied (figure 2.1; Green & Swets 1966).

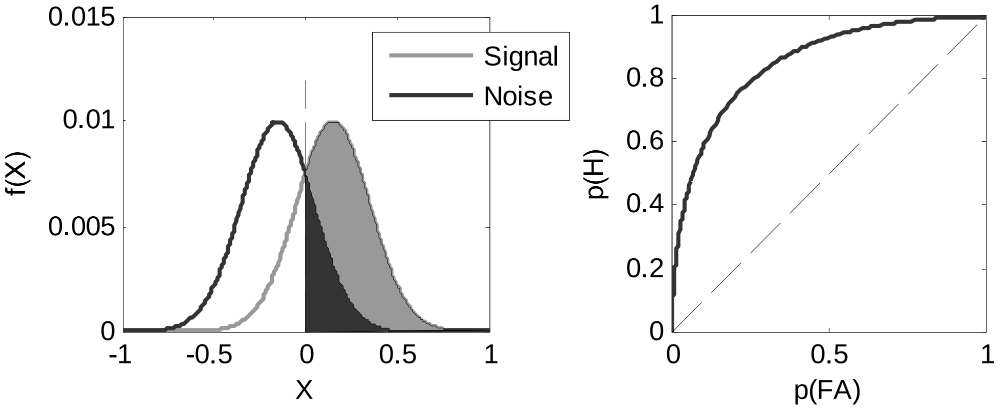

**Figure 2.1:** Left panel: Type 1 SDT distributions for hypothetical 'signal+noise' and 'noise' classes over a random variable $X$. $d'$ increases in direct proportion to the distance between the two distributions and in inverse proportion to the variance. The dotted line represents the criterion value that an ideal observer would choose to maximise accuracy ($\beta = 0$). Light and dark shaded areas represent the proportion of hits and false alarms for this particular criterion value. Right panel: receiver operating characteristic (ROC) curve that tracks out the relationship between hit and false alarm rate for all possible values of decision criteria.

We can loosely categorise these biasing influences, or components of the loss function, into two classes: prior probabilities and prospective costs. Changes to both prior probabilities of the occurrence of one or other target, and/or the costs or benefits associated with one or other decision, have been shown to affect response criteria in categorisation tasks (Alsop & Davison, 1991; Alsop & Porritt, 2006; Bohil & Maddox, 2001; Johnstone & Alsop, 2000). Similarly, evidence that observers adopt an optimal criterion for perceptual judgments based both on their categorisation uncertainty and payoff matrix has been found in low-level visual decision tasks (Landy *et al.*, 2007; Simen *et al.*, 2006; Whiteley & Sahani, 2008). Furthermore, Whiteley & Sahani (2008) found that the best model of the data required only a single psychometric function slope (a measure of sensitivity) for multiple criteria, supporting the independence of sensitivity and bias. A similar experiment in monkeys has also reported optimal integration of asymmetric value with perceptual sensitivity (Feng *et al.*, 2009).

An open question is how bias and sensitivity interact. For instance, object classification is often sufficiently determined to make context irrelevant; when categorising an object as a bus, I do not need to worry about the relative probabilities of the object being a bus to avoid misclassification, because the perceptual input is unequivocal. The basic premise here is that the confluence of perceptual conflict (what Bruner called 'equivocality') and the need to maximise expected gain in the face of this conflict, leads to the prediction of systematic biases in some decision contexts, but not others. It is an unclear whether bias is always necessary, and applied 'blindly' regardless of the evidence the system is operating with (as in SDT), or whether the mechanisms governing bias are themselves dynamically shaped by changes in sensitivity.

One result bearing on this question is that empirically measured criteria often tend to be conservative – they are closer to neutral ($\beta = 1$) than would be expected based on equation 2.4 (Bohil & Maddox, 2001; Green & Swets, 1966; Healy & Kubovy, 1981; Maloney & Thomas, 1991). One explanation of this effect is that observers also place weight on being accurate, rather than only on maximising reward, leading to a competition between reward and accuracy maximisation, or 'COBRA' (Maddox & Bohil, 2004; Maddox, 2002). In COBRA it is proposed that bias should be applied in inverse proportion to the long-running $d'$; this scheme makes intuitive sense, in that it is only when states are uncertain (low $d'$) that penalties associated with misperception become relevant. Such an interaction has been confirmed in animal studies: when pigeons are trained to discriminate between two intensities of red light, asymmetries in the rewards available for each response induce biases in behaviour only when stimulus discriminability is low but not high (Alsop & Porritt, 2006; Davison & Tustin, 1978). The same researchers have reported similar results in humans (Johnstone & Alsop, 2000; Lie & Alsop, 2010). Such interactions between sensitivity and bias pose problems for the standard versions of SDT.

### 2.2.4 Reconciling interactions between sensitivity and bias

In a full Bayesian model, the relative influence of the prior and likelihood is inversely proportional to their precision (MacKay, 2003). This feature provides a simple explanation for why bias and sensitivity might be expected to interact (Ma, 2010). More generally, the weighting of competing sources of information by their precision is at the heart of probabilistic approaches to decision-making (Daw *et al.*, 2005). A common example is visual capture: as vision is usually more reliable than hearing for spatial localisation, the location of an object is usually biased towards that suggested by visual information, partially explaining the well-documented ability of ventriloquists to seemingly 'throw' their voices (Pick *et al.*, 1969; Welch & Warren, 1980; Bertelson, 1999). Psychophysical experiments have demonstrated that this weighting is dependent on the uncertainty of the cue (e.g. Knill & Saunders 2003; Jacobs 1999; Ernst & Banks 2002), and that observers often assign weights in a Bayes-optimal fashion. If the sensory likelihood and loss function both have associated uncertainties, we might expect each source to be weighted in proportion to its precision, exactly as observed in SDT experiments (Alsop & Porritt, 2006; Lie & Alsop, 2010; Johnstone & Alsop, 2000).

However, both COBRA and related approaches run into both mathematical and algorithmic problems, as they do not specify how observers monitor this uncertainty, or precision, to adjust $\beta$ accordingly. Indeed, the problem of criterion setting is problematic in general, as it assumes self-knowledge of the posterior belief distributions involved (Lau, 2008; Ma, 2010). This knowledge can be considered equivalent to observers having access to the antecedents of their decision process. We will turn to these questions in section 2.4.

### 2.2.5 Evidence accumulation models

Intuitively, accumulating further information pertaining to the discrimination can reduce uncertainty. The static models discussed above do not permit this type of accumulation; instead, SDT assumes that the decision-maker has instantaneous access to the underlying signal ($x$). As such, it is silent on temporal issues such as a subject's reaction time (RT). In contrast, a broad class of models known as sequential sampling or evidence accumulation models takes the evolution of evidence into account, forming a more complete model of the decision-process (Link, 1975; Ratcliff, 1978). These models may also have an advantage in connecting more naturally to neural implementation of a decision, as the brain itself deals with dynamic rather than static data (Gold & Shadlen, 2001, 2007).

The evidence accumulation framework treats each sample of data as an independent piece of evidence (Wald & Wolfowitz, 1948). It is thus possible to update a decision variable (e.g. the likelihood ratio), at each point in time. Assuming that $x$ is on average informative about the true state of the world, the decision variable will tend to diverge from zero, despite being corrupted by noise at each timestep. More formally, by transforming the probability ratio into a logarithm (thus permitting additivity), its evolution over time is given by:

$$\begin{aligned}\log(l) &= \log\frac{p(x_1, x_2, \ldots x_n|h_1)}{p(x_1, x_2, \ldots x_n|h_2)}\\ \log(l) &= \sum_{i=1}^{n}\log\frac{p(x_i|h_1)}{p(x_i|h_2)} \qquad (2.5)\end{aligned}$$

This 'decision variable' (DV) increases (decreases) in proportion to the strength of evidence favouring $h_1$ ($h_2$). As for the static theory discussed in section 2.2.2, we need to apply a decision rule to the DV. A simple rule in this case is to accept one or other hypothesis when the DV reaches a 'barrier' or threshold that is mirror symmetric around a neutral starting point (only the positive barrier B is shown in figure 2.2). The further the excursion of this barrier from the starting point, the fewer errors are made, but the longer it takes to make the decision. The rate of evidence accumulation in this type of model can be considered equivalent to $d'$, and the position of the bound equivalent to the criterion (Palmer *et al.*, 2005). Increasing

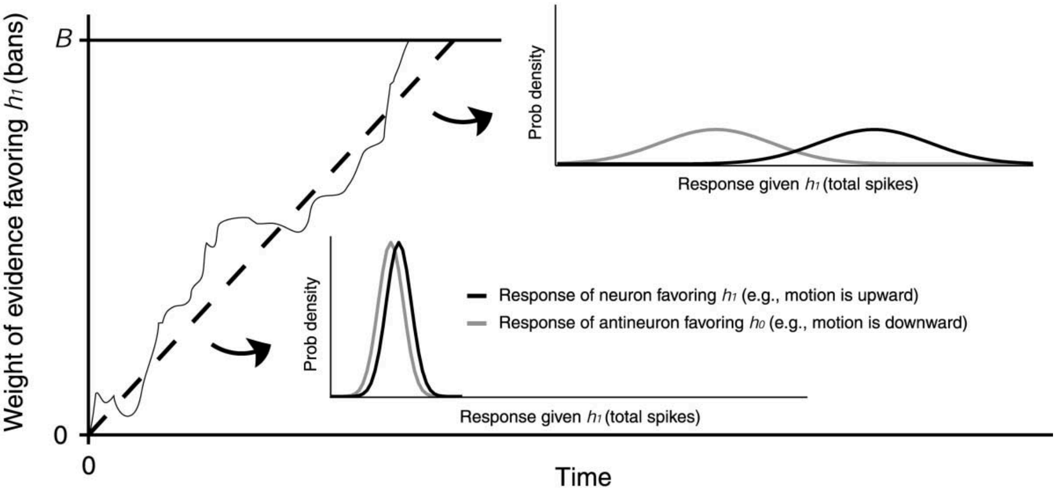

**Figure 2.2:** Graphical representation of the evolution of a decision variable over time in an evidence accumulation model. The two probability density plots indicate the increased separation of instantaneous SDT distributions that occurs with time, demonstrating the benefit to the organism of accumulating evidence. Reproduced with permission from Gold & Shadlen (2002).

the speed of the decision can be achieved by reducing the baseline-to-threshold distance in evidence accumulation models; conversely, emphasis on accuracy is thought to raise the baseline-to-threshold distance (Bogacz *et al.*, 2010). Thus the evidence accumulation model provides a principled account of the ubiquitous speed-accuracy tradeoff (SAT) in decision-making (Luce, 1991), a topic I return to in section 2.3.3.

As we will see in section 2.3.2, activity in several cortical regions involved in decision-making shows properties one might expect of an accumulating decision variable (see Gold & Shadlen 2007 for a review). However, the evidence accumulation framework applies to a restricted range of settings, especially given that the a likelihood ratio computation assumes a small, discrete set of options (two in equation 2.5; see Churchland *et al.* 2008 for an extension to a four-choice scenario). Reframing the accumulation model as a special case of a full probabilistic representation of the options available to the decision-maker (Dayan & Daw 2008; see also Beck *et al.* 2008) might help to strengthen the interpretation of these neural activities in more general terms.

## 2.3 Neural mechanisms for perceptual decision-making

A general experimental approach for investigating perceptual decision-making is to ask observers to make sensory discriminations under conditions of greater or lesser uncertainty. Various paradigms have been developed, including vibrotactile frequency discrimination (Romo *et al.*, 2002a), visual motion discrimination (Newsome *et al.*, 1989) and face-house discrimination (Heekeren *et al.*, 2004). All rely on degrading the sensory evidence available to the observer to investigate the labile link between perception and action. The next sections review the current state of knowledge of the neural mechanisms underlying perceptual decision-making. As the empirical work in this thesis manipulates visual uncertainty, I focus on studies in the visual domain, while noting that many of the conclusions about the neural coding of uncertainty and value often transcend domains (see Heekeren *et al.* 2008 for a review). A general scheme for decision-making is shown in figure 2.3 (adapted from Heekeren *et al.* 2008 and Rangel *et al.* 2008); here, a distinction is made between the representation of evidence, evidence accumulation, and action selection. As emphasised in section 2.2.2, valuation could affect one of several points of this process. Similarly, uncertainty is associated with both the representation of evidence, and the selection of particular actions, and such signals may be used to adaptively resolve competition among response options (Botvinick *et al.*, 2001; Frank, 2006).

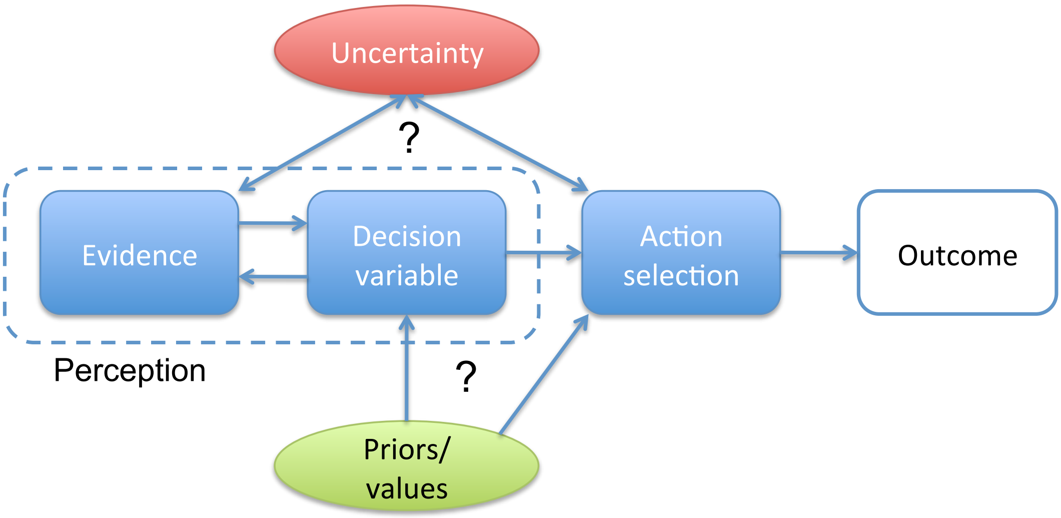

**Figure 2.3:** Hypothetical relationship between various dissociable stages of the decision process. For perceptual decision-making, posterior beliefs over the state of the world are used to compute decision variables that reflect the likelihood the observer will take each possible decision. How value and priors are integrated into this process is unknown, and uncertainty in the sensory-to-motor mapping may arise at both the level of stimulus representation and action selection. Adapted from Heekeren *et al.* (2008) and Rangel *et al.* (2008).

### 2.3.1 Encoding of evidence

One of the most widely used experimental paradigms in perceptual decision-making research is the random dot kinematogram (RDK) discrimination. In an RDK experiment, the subject is required to detect the global direction of motion (usually from two alternatives) in the noisy movement of the dots. Using such stimuli, single-unit recordings in monkeys have identified area MT (V5) as containing neurons that encode the strength of evidence for a given direction of motion (Britten *et al.*, 1996; Newsome *et al.*, 1989), and electrical microstimulation of this area is sufficient to bias the animals choice (Salzman & Newsome, 1994). Analogously, the sensory coding of frequency in vibrotactile frequency (VTF) discrimination is reflected in the firing of primary somatosensory cortical neurons (de Lafuente & Romo, 2006). Consistent with S1 representing the sensory evidence during this task, replacing one interval's tactile stimulus with microstimulation of neurons responsive to the same frequency was sufficient to replicate the endpoint of the behavioural decision (Romo *et al.*, 2002a). Finally, for higher-level categories, stimulation of face-selective neurons in extrastriate visual cortex biased the monkey's decisions towards the face category in a face/non-face discrimation task (Afraz *et al.*, 2006), consistent with these neurons representing evidence in support of the face category.

Functional MRI experiments have revealed that particular sensory dimensions are represented in localised regions of cortex, for example, house and face object categories (Epstein & Kanwisher, 1998; Kanwisher *et al.*, 1997). Building on the VTF paradigm in monkeys, Pleger and colleagues demonstrated that primary somatosensory cortex is active in proportion to the strength of evidence supporting the decision alternative (Pleger *et al.*, 2006). More generally, these results on the representation of sensory evidence accord with the 'standard model' in cognitive neuroscience, where, despite massive recurrence, visual cortex is organised in a hierarchy of regions that are tuned to increasingly complex features (Zeki & Bartels, 1998), from motion, colour and orientation up to high-level categories such as faces and places (Grill-Spector *et al.*, 2001).

### 2.3.2 Formation of a decision variable

In RDK decisions, neural activity that ramps up in proportion to the strength of motion evidence has been found in the lateral intraparietal sulcus (LIP; Shadlen & Newsome 2001). This ramping is consistent with these neurons encoding the likelihood ratio in evidence accumulation models (section 2.2.5). Furthermore, the monkey's eye movement decision can be predicted by this activity reaching a threshold level (Roitman & Shadlen, 2002). Similar neural responses have been documented in dorsolateral prefrontal cortex (dlPFC) (Kim & Shadlen, 1999) and frontal eye fields (FEF) (Thompson & Schall, 2000). Secondary somatosensory (S2) cortical neurons show activity that is proportional to a comparison of the two frequencies in a VTF task (Romo *et al.*, 2002b). Similar results have been found in medial and ventral premotor cortex (Hernandez *et al.*, 2002; Lemus *et al.*, 2007; Romo *et al.*, 2004). Indeed, in a review integrating findings from several recording sites, the transition from representation of the sensory evidence to computation of a decision variable (DV) in the VTF task is characterised as gradual, proceeding from somatosensory to premotor cortex via prefrontal regions (de Lafuente & Romo, 2006). The localisation of an evolving DV within premotor cortical regions suggests decisions are partly embodied within the particular sensorimotor pathway used to make the response (Cisek, 2007; Gold & Shadlen, 2003; Tosoni *et al.*, 2008).

Using multiple criteria for the triangulation of a decision region, Heekeren and colleagues demonstrated that the BOLD signal in left dlPFC tracks the difference between activity in fusiform face area (FFA) and parahippocampal place area (PPA) (Heekeren *et al.*, 2004) during identification of noisy images. However, it is unclear whether the 'decision' may already be computed in the category-invariant responses of neurons in FFA and PPA for high-level object categories such as faces and houses (McKeeff & Tong, 2007; Afraz *et al.*, 2006; Ploran *et al.*, 2007). Other studies have provided evidence (similar to the primate literature) that evidence accumulation is not restricted to dlPFC, but is spread out across a fronto-parietal network (Philiastides *et al.*, 2006; Philiastides & Sajda, 2007; Ploran *et al.*, 2007; Thielscher & Pessoa, 2007; Tosoni *et al.*, 2008). It remains an open question as to which are the critical nodes in this network that facilitate accurate perceptual decision-making.

An important and unanswered question is how decision variables interact with motor plans (Freedman & Assad, 2011). As noted above, prefrontal and parietal areas thought to encode decision variables overlap with those involved in the planning, selection and implementation of motor responses (e.g. Hernandez *et al.* 2002). However, many studies conflate the decision process with the motor response, for instance by requiring an leftward saccade to signal leftward motion. In non-human primates, a recent study suggests that activity in LIP encodes the monkey's categorical decision independent of activity related to motor programming (Bennur & Gold, 2011). In humans, Heekeren *et al.* (2006) varied response modality in the same motion-discrimination task and found that a network of left posterior dlPFC, cingulate cortex, left intraparietal sulcus (IPS) and left fusiform/parahippocampal gyrus correlated with the strength of sensory evidence independent of whether responses were given with button presses or eye movements (see also Ho *et al.* 2009). More generally, a central feature of PFC function is the selection of responses on the basis of context, rather than just the gating of a particular motor program (Miller & Cohen, 2001). Thus an important issue is how flexible and multifarious fronto-parietal decision-variables are, how they are linked to specific actions, and whether there are significant species differences in their level of abstraction (Freedman & Assad, 2011; Heekeren *et al.*, 2008). In Chapter 9 I will return to this issue and outline how the 'frame of reference' in which decision variables are encoded may have important implications for relating perceptual decision-making to metacognitive function (section 2.4).

### 2.3.3 Neural basis of the decision threshold

The basal ganglia are a set of subcortical nuclei that play a pivotal role in action selection (figure 2.4; Gurney *et al.* 2001; Redgrave *et al.* 2010). Two opposing pathways – the direct and indirect pathway – facilitate and suppress the selection of actions, respectively (Alexander & Crutcher, 1990). Recent models of the basal ganglia have additionally placed emphasis on 'hyperdirect' inputs into the STN (Nambu *et al.*, 2000, 2002) that may modulate the flow of activity around the cortico-basal ganglia circuits. Cortico-basal ganglia loops have been proposed to be central in the setting of the decision threhsold (Simen *et al.*, 2006; Lo & Wang, 2006), consistent with neurons in the dorsal striatum responding to changes in response criteria (Lauwereyns *et al.*, 2002; Pasquereau *et al.*, 2007).

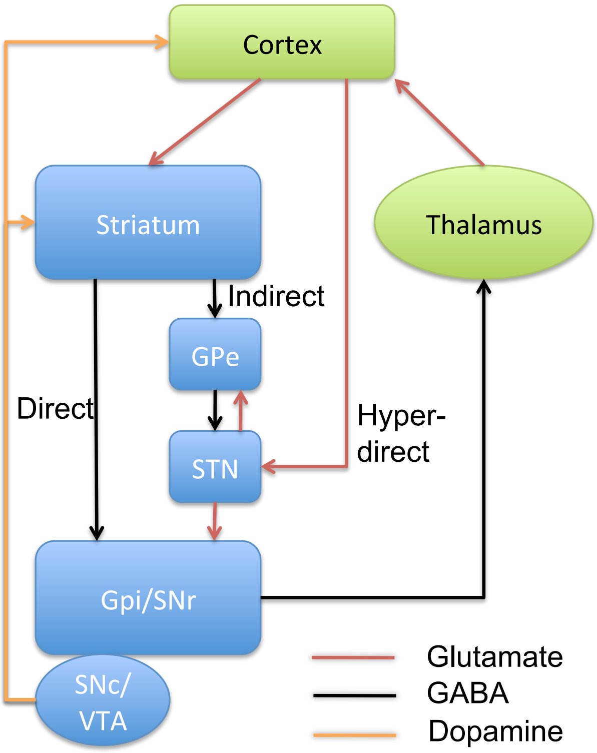

**Figure 2.4:** A subset of known connections between basal ganglia nuclei, thalamus and cortex. The direct, indirect and hyperdirect pathways are marked. Adapted from Redgrave *et al.* (2010).

At least two non-mutually exclusive hypotheses have been proposed with regard to the basal ganglia's role in setting decision thresholds (Bogacz *et al.*, 2010). First, the striatal hypothesis predicts that increased excitatory input from cortex is the neural instantiation of an increase in an evidence accumulation baseline. Forstmann and colleagues found that the BOLD signal in the pre-supplementary motor area (pre-SMA) and striatum was increased when speed was emphasised in the RDK task; this pre-SMA increase correlated negatively with an individual differences measure of response caution, or the weight subjects ascribed to being accurate (Forstmann *et al.*, 2008). Furthermore, a recent structural imaging study found that the white matter connections between pre-SMA and caudate were stronger in subjects who displayed greater alterations in response thresholds (Forstmann *et al.*, 2010), leading the authors to suggest that the pre-SMA provides a controlling input to the striatum to adjust SAT.

Other studies have proposed the STN is an important node in the setting of response thresholds, with increased STN activity producing slower and more accurate choices (Bogacz *et al.*, 2010). Specifically, Frank and colleagues have proposed that areas of frontal cortex detect the need for cognitive control, and activate the STN to slow down decision-making (Frank, 2006; Frank *et al.*, 2007). This view is supported by the fact that deep-brain stimulation (DBS) of the STN for treatment of Parkisons disease leads to deficits in impulse control when conflict is high (Alberts *et al.*, 2008; Ballanger *et al.*, 2009; Frank *et al.*, 2007; Hershey *et al.*, 2004), and lesions of the STN in rodents produce impairments in high-conflict decision-making (Baunez *et al.*, 2001; Eagle *et al.*, 2008).

Potential sources of basal ganglia modulation include the inferior frontal cortex (IFC) and pre-SMA. The IFC, particularly the right IFC (although see Swick *et al.* 2008), has been specifically implicated in the inhibition of motor responses (Aron *et al.*, 2003; Garavan *et al.*, 1999; Menon *et al.*, 2001; Rubia *et al.*, 2001; Hodgson *et al.*, 2007; Leung & Cai, 2007; Swann *et al.*, 2009). Studies of the stop-signal reaction time (SSRT) task using fMRI have isolated both the right IFC and STN as critical nodes the stopping of ongoing responses (Aron & Poldrack, 2006; Li *et al.*, 2008). Deep brain-stimulation of the STN in patients with Parkinson's disease directly modulates SSRTs (Ray *et al.*, 2009; van den Wildenberg *et al.*, 2006). An important study confirmed that, as would be predicted by an adjustment of decision threshold, the rIFC and STN are active both for outright stopping and slowing of responses (Aron *et al.*, 2007), and that these functionally activated regions overlapped with interconnected regions identified using white-matter tractography.

However, recent studies have begun to question the specificity to which rIFC activation can be ascribed solely to motor inhibition. rIFC is implicated in a wide variety of cognitive tasks, including attentional reorienting (Corbetta & Shulman, 2002; Hampshire & Owen, 2006), oddball detection (Bledowski *et al.*, 2004; Hampshire *et al.*, 2007) and updating actions in the light of new information (Mars *et al.*, 2007). Indeed, these latter two functions may be compatible with a role for IFC and/or STN in action reprogramming, rather than inhibition per se (Mostofsky & Simmonds, 2008). With respect to the IFC, two recent lines of evidence support this suggestion. First, when comparing two types of trials using fMRI, both requiring detection of a novel cue but only one requiring inhibition, only the pre-SMA but not the IFC was found to be specific to inhibition (Dodds *et al.*, 2010; Sharp *et al.*, 2010). Dodds *et al.* (2010) built upon these results by showing that the rIFC was active both during attentional and motor shifts, and that this region was even more active during trials requiring increased response control (an additional response) compared to trials only requiring inhibition (no-go trials). Second, pre-SMA, but not rIFC, was shown to have heightened functional connectivity with the basal ganglia in the SSRT (Duann *et al.*, 2009), in keeping with its role in resolving response competition (Nachev *et al.*, 2007; Sumner *et al.*, 2007).

With respect to the STN, a single-neuron recording study in monkeys found that STN neurons were active during trials requiring action reprogramming (Isoda & Hikosaka, 2008), with similar activity reported in the pre-SMA (Isoda & Hikosaka, 2007). Employing a similar task, Neubert and colleagues used paired-pulse transcranial magnetic stimulation (TMS) to reveal that pre-SMA and rIFC have facilitatory and inhibitory effects respectively on motor output during trials requiring a switch in a planned movement (Neubert *et al.*, 2010). Furthermore, subjects who showed the greatest inhibitory effects also showed greater white-matter connecitivity between rIFC/pre-SMA and the STN region. However, it is unknown whether the STN signal seen in human fMRI experiments reflects motor inhibition per se (Aron & Poldrack, 2006; Li *et al.*, 2008), or whether, as may be the case for the rIFC/pre-SMA, it plays a more general role in action reprogramming.

### 2.3.4 Effects of asymmetric priors

Functional imaging studies have revealed activity that is systematically modulated by prior beliefs in a stimulus class. A classic and simple example of this effect is the oddball response, or mismatch negativity (Näätänen *et al.*, 1987). By defining surprise as an information theoretic quantity $[I = -log(P(x_i)]$, Strange *et al.* (2005) showed that the activity of a widespread corticothalamic network correlated with the conditional surprise of an event in a stimulus sequence. A similar analysis revealed that the P300 component of the electroencephalography (EEG) signal correlates with trial-by-trial surprise (Mars *et al.*, 2008). Surprising events are accompanied by slowed reaction times (Bestmann *et al.*, 2008; Mars *et al.*, 2008), and Bestmann *et al.* (2008) found that this slowing was mediated by a decrease in the excitability of the cortico-spinal motor tract using TMS, consistent with decreased drive to the motor system and a raised decision threshold (see section 2.2.5). More broadly, this work indicates that the brain is sensitive to the probabilistic context of stimuli, consistent with predictive coding models of cognition (Friston, 2009; Rao & Ballard, 1999), and indicates expectation violation may be a key driver of action reprogramming (see previous section 2.3.3).

Another line of research has investigated how predictions adjust the perceptual decision process. Classical attentional paradigms use cues to bias perception towards or away from a particular stimulus (e.g. Posner *et al.* 1980), altering activity in both spatially-selective and feature-selective visual cortex via top-down modulatory connections originating in prefrontal cortex (see Desimone & Duncan 1995; Corbetta & Shulman 2002 for reviews). Similarly, Summerfield *et al.* (2006a) found that when subjects were expecting to see ('looking for') faces in a stream of visual stimuli, the BOLD signal in FFA was selectively increased. This effect was accompanied by increased backward connectivity from frontal cortex as a function of changes in expectation (see also Summerfield & Koechlin 2008). Changes in the probability of face/house stimuli have been found to activate a network of frontoparietal regions (Puri *et al.*, 2009), as well as causing baseline shifts in the activity of FFA and PPA (Egner *et al.*, 2010; Esterman & Yantis, 2010; Puri *et al.*, 2009).

### 2.3.5 Effects of asymmetric costs

In perceptual discrimination tasks, participants often receive feedback that signals whether their response was correct or not. Increasing the reward available for correct judgments can motivate increases in performance and induce attention-like modulations of early sensory activity in both somatosensory (Pleger *et al.*, 2008) and visual (Weil *et al.*, 2010) domains, an effect that may be dependent on dopamine (Pleger *et al.*, 2009). However, relatively less is known about how asymmetric costs associated with one or other alternative affect the neural hierarchy subserving perceptual decisions. As outlined in section 2.2.2, computational models of behaviour are agnostic as to how value is incorporated. Changes in value linked to particular regions of space have been shown to modulate spatially selective regions of early visual (Serences, 2008) and somatosensory (Pleger *et al.*, 2008) cortex. Using elegant psychophysical analysis, Liston & Stone (2008) demonstrated that associating one region of space with greater reward probability increased the subjective perception of brightness of targets at that location (measured via a post-decision report), supporting the proposal that asymmetric rewards have effects on early visual processing. Such modulation may occur via recruitment of fast attention-like mechanisms (Serences, 2008), but the extent to which attentional biases can be dissociated from knowledge of the spatial distribution of rewards is an open question (Maunsell, 2004).

When stimulus value is defined by identity, rather than spatial location, the picture is less clear. In a single-unit recording study, Rorie *et al.* (2010) demonstrated that asymmetric payoffs bias the initial firing rate of individual neurons in the intraparietal sulcus coding for a saccadic response to one of two particular targets. Similar effects (albeit induced via changes to category boundaries, rather than asymmetric reward) have been observed in FEF neurons (Ferrera *et al.*, 2009). These modulations are consistent with changes in the starting point and/or barrier of evidence accumulation circuits, rather than the accumulation rate (Brodersen *et al.*, 2008; Gold & Shadlen, 2002), a hypothesis supported by model fits to subjects' reaction times in asymmetric reward tasks (Feng *et al.*, 2009; Simen *et al.*, 2009; Summerfield & Koechlin, 2010). Indeed, computational simulations of the decision process have suggested that asymmetries in potential rewards optimise the height of the decision threshold (barrier), making responses associated with higher rewards more likely irrespective of the evidence obtained (Bogacz & Gurney, 2007; Simen *et al.*, 2006). Together, these studies predict that the effects of asymmetric costs in perceptual decision-making might be observed in brain regions involved in setting the decision threshold (section 2.3.3), rather than representing sensory evidence.

### 2.3.6 Effects of increasing task difficulty

Decisions can be easy or difficult. One influential suggestion is that the decision-making system should be sensitive to the current level of difficulty to mobilise additional 'cognitive control' resources in an adaptive fashion (Botvinick *et al.*, 2001). Studies that have directly investigated this component of perceptual decision-making have found a network of regions centred on dorsomedial prefrontal cortex (dmPFC) and anterior insula are engaged when reaction times increase despite the stimulus remaining constant (Binder *et al.*, 2004; Philiastides *et al.*, 2006; Philiastides & Sajda, 2007) or stimuli are closer to a category boundary (Grinband *et al.*, 2006). Such activity naturally connects to the research discussed in section 2.3.3, where conflict-related activity was proposed to activate the STN and raise decision thresholds (Frank, 2006). Indeed, dmPFC is often coactivated with the lateral frontal cortex in neuroimaging studies (Koski & Paus, 2000), and is thought to recruit lateral PFC to implement increases in cognitive control (Kouneiher *et al.*, 2009; MacDonald *et al.*, 2000).

As discussed above, uncertainty can enter into the decision process at at least two distinct loci. First, uncertainty over the stimulus may occur, despite response mappings being clear and unambiguous (as for perceptual decisions involving degraded stimuli). Second, the stimulus might be unambiguous, but conflict is induced due to response mappings being similar (Botvinick *et al.*, 2001). Conflict between multiple potential responses activates the ACC even when the sensory evidence is unambiguous (Botvinick *et al.*, 2001), and may be related to the monitoring of potential and actual errors in choosing (Brown & Braver, 2005; Gehring *et al.*, 1993; Magno *et al.*, 2006; Yeung *et al.*, 2004). Wendelken and colleagues used a novel version of the RDK task to attempt to dissociate these contributions to decision difficulty (Wendelken *et al.*, 2009). It was found that when conflict between irrelevant and relevant motion information was high ('stimulus conflict'), the middle temporal area (human MT) and right IFC was active; conversely, the parietal cortex and dmPFC were selectively active when the two kinematograms indicated conflicting responses compared to congruent responses ('response conflict'). A central role for posterior parietal areas in resolving response-specific conflict has been demonstrated through careful testing of patients with lesions to this region (Coulthard *et al.*, 2008).

One limitation of studies of cognitive control is that they often rely on manipulations of either the stimulus and/or correlations with reaction time to identify regions that are related to adaptive adjustments to task difficulty. Thus an alternative interpretation is that these regions participate in the process of decision-making itself (Thielscher & Pessoa, 2007), rather than exerting cognitive control over an independent sensorimotor pathway (Miller & Cohen, 2001). To the extent to which uncertainty optimises a single 'closed loop' system, the adaptive effects of conflict will be confined to a single sensorimotor pathway, and may be unreportable. Conversely, to the extent that the latter hypothesis is true, we might expect that subjects can report domain-general uncertainty, or conflict, through other modalities such as verbal reports. One way to gain traction on this distinction is to explicitly ask subjects to comment on their own decision process. By doing so we can ask to what extent variables such as perceptual and response uncertainty are encoded in a higher-order frame of reference (and thus are available to be reported). This approach will be discussed next.

## 2.4 Metacognition during decision-making

Metacognition refers to the monitoring of one's own cognition and behaviour. Humans (and possibly some animal species, see Smith 2009) are able to use metacognitive commentaries to communicate confidence in their own decision process (Flavell, 1979; Hampton, 2001; Kunimoto *et al.*, 2001; Persaud *et al.*, 2007). The neural mechanisms contributing to metacognition remain largely unknown (Shimamura, 2000). Metacognitive reports are particularly intriguing, as they allow for a fuller characterisation of the subjective milieu of a decision, by assessing what the observer can tell the experimenter about their decision at a given point in time. The theory and methods underlying this approach are outlined in the next section.

### 2.4.1 Metacognitive measures

Metacognitive reports about perceptual decisions can take many forms, but in general require the subject to give an additional report or commentary over and above their initial forced-choice response. For example, Peirce & Jastrow (1885) asked observers to rate their degree of confidence in their judgment on the following scale:

> '0' denoted absence of any preference for one answer over its opposite, so that it seemed nonsensical to answer at all. '1' denoted a distinct leaning to one alternative. '2' denoted some little confidence of being right. '3' denoted as strong a confidence as one would have about such sensations.
>
> Peirce & Jastrow (1885)

Since this seminal work, asking for confidence-in-accuracy has become a standard tool for analysing judgments of performance. Several studies have documented a systematic relationship between confidence and task features such as difficulty and processing time (Baranski & Petrusic, 2001; Graziano & Sigman, 2009; Vickers, 1979), where under most conditions, confidence decreases with both increasing difficulty and response time (see Pleskac & Busemeyer 2010 for a review). However, there are several methods of elicitation that can be used to more or less accurately tap the observers assumed underlying confidence. I consider a selection of these methods next.

#### 2.4.1.1 Free report

The simplest measure of confidence-in-accuracy is a subjective rating. This can be obtained using a binary categorisation ('Sure', 'Unsure'), or a continuous rating scale such as that used by Peirce & Jastrow (1885). One drawback of subjective reports is that the scale is open to interpretation: a confidence rating of 'Sure' might mean different things to different people. Measuring an observer's metacognitive sensitivity (ability to discriminate correct from incorrect decisions) is largely unaffected by this subjectivity, provided the observer is encouraged to give a range of relative ratings (see Chapter 7). One other potential measure of interest is whether an observer's confidence judgments are accurately calibrated. Answering this question requires an absolute scale, such as eliciting the numerical probability that a decision was answered correctly. This probability can then be compared to the true probability, computed as an aggregate measure over several trials. The typical finding is (relative) over-confidence for difficult decisions, and (relative) under-confidence for easy decisions (e.g. Gigerenzer *et al.* 1991).

Subjective reports have also been used to investigate awareness of errors. Again this is usually achieved via a two-stage decision procedure – subjects make a forced-choice decision, and subsequently 'comment' to the experimenter via a separate response if they were aware of having made an error. This would appear to be a similar cognitive phenomenon to the monitoring of subjective confidence. Interestingly, errors can be made in the presence and absence of such a subjective report, which permits analysis of error monitoring within the Type 2 signal detection theory (SDT) framework I go on to discuss below (Steinhauser & Yeung, 2010). A rich body of literature on error awareness has identified the 'error-related positivity' (Pe) as being an ERP component selectively associated with error awareness (Nieuwenhuis *et al.*, 2001; Steinhauser & Yeung, 2010), possibly originating in anterior insula cortex (see Ullsperger *et al.* 2010, for a review).

#### 2.4.1.2 Post-decision wagering

One problem that has dogged subjective reports is their potential unreliability (Eriksen, 1960). Why should the subject be motivated to reveal his true confidence, when there is no incentive to do so? In addition, the necessarily subjective instructions given to the subject in a free report task preclude the use of these measures in children and non-human animal species. Even if instructions can be imparted accurately, they may be misinterpreted, leading to aberrant usage of the scale. To circumvent these issues, Persaud and colleagues introduced post-decision wagering (PDW) as an intuitive measure of metacognitive confidence (Persaud *et al.* 2007; see also Seth 2008). In its simplest form, a participant is asked to gamble on whether their response was correct or not. If the decision was correct, the wager amount is kept; if it was incorrect, this amount is lost. The size of the chosen gamble is assumed to be a reflection of the subject's confidence in his or her decision.

#### 2.4.1.3 Incentive-compatible scoring

In the same spirit as PDW, incentive-compatible scoring aims to elicit 'true' underlying confidence. However, here the goal is less that of an intuitive measure, rather the construction of payoff structure that provides a maximum return when the participant provides an honest rating. This approach is similar to that taken by the Becker-DeGroot-Marschak procedure in behavioural economics to elicit the true value attached to an object (Becker *et al.*, 1964). For example, using the Quadratic Scoring Rule (QSR), the participant's payoff is proportional to his or her confidence if they get the answer correct, but provides an increasing penalty if the answer provided was incorrect. It is, in effect, a graded version of PDW. A more complex procedure is provided by the Lottery Rule (Holt & Smith, 2009): here, participants enter into a two-step lottery following every decision. The elicited probability of being correct is used to set the probability that the first lottery is won. If this lottery is lost, then a second lottery is initiated with a probability set by the random number drawn on the first lottery. It can be shown that this procedure motivates the subject to reveal her true subjective probability that the decision was correct in order to maximise their reward (Holt & Smith, 2009). However, one potential drawback of this procedure is that it requires considerable depth of understanding on the part of the subject. Furthermore, Hollard and colleagues found only minor advantages in use of the Lottery Rule over the considerably more intuitive free report method (Hollard *et al.*, 2010).

#### 2.4.1.4 Implicit metacognitive measurement

Other studies document how subjects are able to monitor their decision uncertainty, and use this monitoring to effectively guide task performance. For example, Barthelmé & Mamassian (2009) asked observers to choose between two stimuli to judge on any given trial, before carrying out a difficult perceptual judgment on the chosen stimulus. Observers were seen to improve their performance by choosing the less uncertain among a pair of visual stimuli, demonstrating that higher-order sensitivity to uncertainty can be used to guide future decision making. Other work has demonstrated that subjects can use knowledge of uncertainty to optimally bias decision-making in both perceptual (Landy *et al.*, 2007; Whiteley & Sahani, 2008) and motor (Trommershauser *et al.*, 2003) tasks. These abilities have also been found in non-human animal species: both rats (Foote & Crystal, 2007; Kepecs *et al.*, 2008) and monkeys (Hampton, 2001; Kiani & Shadlen, 2009; Smith *et al.*, 2004) show indirect signs of metacognitive monitoring through effective use of an 'opt-out' response on uncertain trials.

### 2.4.2 Extending statistical decision theory to metacognitive judgments

The brief review in the previous sections indicates that subjects can monitor and report their uncertainty during decision-making, and use these estimates to guide future behaviour. What then is the optimal strategy for observers to adopt when generating a metacognitive report? When rewards are contingent on the correctness of the previous decision (as in post-decision wagering and incentive-compatible scoring), observers should compute the conditional probability of being correct given their previous choice, $P(\text{correct}|\text{choice})$, and use this quantity to adjust the second-order decision. There are various proposals for how this might be achieved. Most involve tracking the strength of the underlying evidence. In signal detection theory, it is assumed that confidence ratings reflect the existence of multiple decision criteria, each corresponding to an increment in confidence in the category (Green & Swets 1966; see figure 2.5). Similarly, in a dynamic situation, Vickers (1979) proposed that decision confidence could be computed by comparing the absolute distance between the winning and losing integrators in an evidence accumulation model (see also Kepecs *et al.* 2008). A probabilitistic analogue of this approach has demonstrated that the distribution of activity over a population of simulated evidence accumulation cells is sufficient to provide a robust estimate of the likelihood of being correct (Beck *et al.*, 2008; Moreno-Bote, 2010).

While not explicitly stating so, these accounts of decision confidence refer to Type 1, or belief confidence, as it is assumed that confidence is equivalent to a graded belief in the percept. In contrast, post-decision confidence ratings are metacognitive judgments about the subject's response. These second-order judgments can be usefully analysed within the framework of Type 2 signal detection theory (Type 2 SDT). Type 2 SDT was first devised by Clarke *et al.* (1959), but recently there has been a revival of the method, spurred on by an in-depth derivation of the relevant probability distributions by Galvin *et al.* (2003). For post-decision confidence, a 'hit' is then a high-confidence correct judgment, and a 'false alarm' is a high confidence incorrect judgment (table 3.2). As for Type 1 SDT, if continuous confidence ratings are used, an ROC function relating how increasing confidence discriminates between correct and incorrect judgments can be derived (figure 7.2). Often post-decision confidence ratings are used to construct a *Type 1* ROC, by treating them as equivalent to ratings of the original stimulus dimension (Macmillan & Creelman,

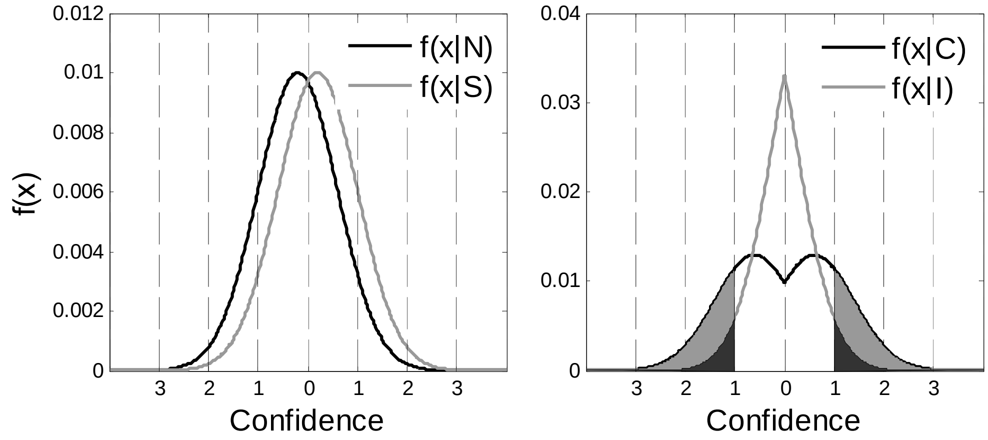

**Figure 2.5:** Left panel: Type 1 distributions as in figure 2.1, but here overlaid with hypothetical post-decision confidence criteria. Confidence is assumed to increase both for post-decision confidence in reporting signal (to the right of the $x$-axis), and post-decision confidence in reporting noise (to the left of the $x$-axis). Right panel: A formal treatment of post-decision confidence is obtained by computing the conditional (Type 2) probability of being correct ($f(x|C)$) or incorrect ($f(x|I)$) given a particular value of $X$ (Clarke *et al.*, 1959; Galvin *et al.*, 2003); here, the symmetry of the distributions naturally maps onto the symmetry of the confidence criteria. The proportions of Type 2 hits (light shading) and false alarms (dark shading) are indicated for a single value of the Type 2 criterion.

2005). However, as Galvin *et al.* point out, 'the difference between a Type 1 and Type 2 task lies in which [mental] events are being evaluated, not in whether the evaluation is binary or a rating' (p. 845). Thus Type 1 decisions can be made on a continuous scale, such as brightness, or a categorical scale, such as present/absent. Similarly, Type 2 decisions can be made on a continuous scale, such as confidence, or on a categorical scale, such as correct/incorrect. It is the referent of these judgments (stimulus or response) that separates them into the classes of Type 1 or Type 2.

Type 2 probability functions (the conditional probability of being correct or incorrect for a given stimulus strength) can be derived theoretically, from a linear transformation of the assumed underlying Type I SDT distributions (Galvin *et al.*, 2003; Kiani & Shadlen, 2009). Knowledge of the underlying distributions is required, through assuming some representation of uncertainty on the part of the subject (Barthelmé & Mamassian, 2009; Whiteley & Sahani, 2008). This derivation – the 'direct translation hypothesis' (Higham *et al.*, 2009) – assumes that confidence is based directly on an assessment of the probability of being accurate, as derived from the perceptual system, and is unaffected by other factors endogenous to the subject. The model is thus 'feed-forward' in that it assumes that perceptual performance is veridically translated into estimates of metacognitive confidence. Further, the model makes the prediction that when perceptual (Type 1) performance increases, metacognitive (Type 2) performance, or the ability to discriminate correct from incorrect decisions, also increases (figure 7.2).

Conversely, models that permit partial independence between perception and post-decision confidence have recently been proposed (Pleskac & Busemeyer, 2010; Del Cul *et al.*, 2009; Insabato *et al.*, 2010). For example, Pleskac & Busemeyer (2010) set out to devise an evidence accumulation model that could account for a wide range of empirical regularities governing the relationship between choice and confidence ratings using the same underlying computational variable. The solution here was to allow a decision variable to continue to accumulate (and be perturbed by noise) beyond the point at which the decision is made[3], until it is accessed to form a confidence judgment. The model provided a good fit to behavioural data, including differences in confidence for correct and incorrect responses. However, neural data supporting these predictions are yet to be obtained. Testing this model would require analysis of neural activity after a decision is made, but before a confidence rating is elicited.

Indeed, studies using methods other than analysis of post-decision confidence also suggest that the direct translation account is incomplete (Baranski & Petrusic, 1998; Busey *et al.*, 2000; Busey & Arici, 2009; Wilimzig *et al.*, 2008). For example, Baranski & Petrusic (1998) concluded based on analysis of reaction times that confidence is partly constructed after the perceptual judgment is made. Metacognitive processes can also be manipulated by factors that are orthogonal to the task of interest: Bengtsonn and colleagues found that priming subjects as 'clever' or 'stupid' altered the monitoring of errors, but not basic task performance, which was held constant throughout the experiment by use of a staircase procedure (Bengtsson *et al.*, 2010); similarly, manipulating the ease of processing can affect metacognitive reports while leaving task performance relatively unaffected (Alter & Oppenheimer, 2009; Busey & Arici, 2009; Wenke *et al.*, 2010).

The antecedents of metacognitive knowledge of decision-making would thus appear to be complex and multi-faceted. In my view, a reductionist approach to defining these processes is preferable: by building on current knowledge of perceptual decision-making (section 2.3), the components of metacognitive processes should become more transparent. Such logic is harnessed by the direct-translation account derived from Type 2 SDT, but this model is likely to be too simplistic. The inter-relationships between first-order decision making, confidence and metacognitive ability are yet to be fully characterised.

### 2.4.3 Neural mechanisms of metacognitive report

In this section I expand my scope to encompass studies of memory and other forms of decision-making, as relatively little is known about the neural bases of metacogni-

[3] Recent analysis of changes of mind in perceptual decision-making make a strong case for continual accumulation of evidence beyond the initiation of a decision (Resulaj *et al.*, 2009), lending plausibility to this proposal.

tive processes in any one given domain. This approach is also motivated by the suggestion that metacognitive function relies on hierarchical processing of 'first-order' cognitive operations. As such, metacognitive variables may be encoded in a fashion that transcends task-specific mechanisms, suggesting a domain-general rather than a domain-specific process (Alter & Oppenheimer, 2009; Song *et al.*, 2011).

#### 2.4.3.1 Neural basis of metamemory

Knowledge of one's own memory capacity has been dubbed 'metamemory' (Flavell, 1979). Often metamemory is studied using ancillary reports (section 2.4.1) or by recording whether individuals engage in particular strategies to manage their own learning process (Metcalfe, 1996). Shimamura & Squire (1986) reported that Korsakoff's syndrome patients presented with impaired feeling-of-knowing (FOK), a measure of whether they will be able to recognise the correct answer to a question, even if they cannot the answer at the time (Nelson & Narens, 1990). Other amnesics were not impaired in FOK judgments. One suggestion is that frontal lobe damage specific to Korsakoff's underlies the impairment in metamemory, consistent with non-amnesic frontal patients also displaying impaired FOK (Janowsky *et al.*, 1989).

Separate evidence for frontal lobe involvement in metamemory comes from fMRI studies. Activation of the PFC during memory encoding predicts future memory performance (Wagner *et al.*, 1998). One suggestion is that the PFC is recruited when second-order, or metacognitive, operations are carried out on first-order memory representations (Christoff & Gabrieli, 2000; Fletcher & Henson, 2001). Consistent with this suggestion, Chua *et al.* (2009) found that activity in anterior PFC (BA 10/9) and IFC (BA 45) showed greater activity for high compared to low prospective FOK judgments. Interestingly, these same regions were not modulated by the level of retrospective confidence in a memory judgment, suggesting they are sensitive to a particular type of first-order task information. Other studies have examined confidence at recognition, finding increased activation of memory-related medial temporal lobe (MTL) regions as confidence increases for true memories, but activation of frontoparietal regions for high confidence false memories (Kim & Cabeza, 2007; Moritz *et al.*, 2006). In a related paradigm, Kao *et al.* (2005) scanned participants while they made predictions about whether they would later remember a variety of different pictures. Objective success in recognition correlated with activity in the MTL, whereas predicted success (irrespective of objective success) activated medial PFC activity. No region showed activity related to the accuracy of the metacognitive judgment (Type 2 hits > false alarms); however, across subjects, vmPFC activity correlated with metamemory ability (the coupling between confidence and memory performance). In summary, these studies suggest that confidence in true memories may derive from a process of recognition (mediated by the MTL), but that modulation of PFC activity can additionally alter meta-level processes to affect confidence, for example when a false memory is held with high confidence.

#### 2.4.3.2 Metacognitive mechanisms in decision-making

Kiani & Shadlen (2009) adapted the RDK paradigm to allow the investigation of decision confidence. They found that monkeys could learn to use an 'opt-out' response on trials where they were unsure about the right answer. The likelihood of using this response was causally related to intermediate firing of LIP cells, previously shown to reflect the accumulation of evidence supporting one or other response (section 2.2.5). However, the extent to which this activity signals confidence in the stimulus or confidence in the decision is unclear, as no post-decision judgment was collected. Another recent study recorded the neural firing of rat orbitofrontal cortex (OFC) neurons during the period after a difficult perceptual decision (Kepecs *et al.*, 2008). Firing was modulated by the difficulty of the decision, and activity differed between correct compared to incorrect trials before the rat had received feedback. While these results might also be interpreted as a signal of expected value – which is higher on correct trials, and has previously been found in OFC (Schoenbaum *et al.*, 2007) – a signal relating to the correctness of a decision in the absence of feedback would also be expected from a circuit computing subjective confidence in the decision. Similar signals showing differences between correct and error trials in the absence of feedback have also been identified in monkey BA10 and dorsolateral PFC (Tsujimoto *et al.*, 2010; Middlebrooks & Sommer, 2010).

An alternative perspective on the encoding of metacognitive variables during decision-making has emerged from studies employing hierarchical computational models. For example, Behrens and colleagues had subjects learn which of two response options delivered greater average reward over time. Unbeknownst to the subjects, the rate at which the best options switched over time, or their volatility, was altered. A hierarchical model which tracks both the reward and its associated volatility provided the best fit to subjects' behaviour, suggesting that they encoded higher-order knowledge about their uncertainty about reward estimates and used this to bias their learning. Furthermore, this higher-order volatility parameter correlated with activity in the anterior cingulate cortex (ACC) (Behrens *et al.*, 2007). In a study employing a similar model-driven analysis, uncertainty about one's position in a computerised maze was correlated with BA10 activity (Yoshida & Ishii, 2006), a region where activity is also linked to exploration during learning (Daw *et al.*, 2006), suggesting that the brain might use a current estimate of belief uncertainty to guide future exploration (Boorman *et al.*, 2009).

Finally, studies aimed at dissociating components of subjective awareness have provided evidence that metacognitive ability – the ability to link performance to confidence – may by partially dissociable from perceptual performance, and may depend on prefrontal function. Rounis *et al.* (2010) showed that within individuals, metacognitive sensitivity in a perceptual task could be selectively decreased through bilateral TMS of dorsolateral PFC, despite perceptual performance remaining intact. Similarly, patients with lesions to anterior prefrontal cortex show an increased threshold for producing a subjective 'seen' response compared to controls in a visual task, despite objective performance being matched between groups (Del Cul *et al.*, 2009). The peak correlation of the lesion with the decrease in subjective report threshold was seen in left BA10.

#### 2.4.3.3 Neural basis of insight

While conceptually broad in scope, studies investigating the neural correlates of self-reflection may provide important information as to where we might expect metacognitive processing to occur. A common network consisting of dmPFC and anterior PFC is active during reflections on mental states (Christoff & Gabrieli, 2000; Fletcher & Henson, 2001; Northoff & Bermpohl, 2004; Passingham *et al.*, 2010). This network is highly reminiscent of that engaged when assessing the mental states of others (Amodio & Frith, 2006), suggesting commonalities between processes engaged by metacognition and mindreading (Carruthers, 2009). A lack of self-reflective capacity is thought to underlie the anosagnosias, neurological conditions in which subjects are unaware of their own deficits. One extreme anosagnosia is Anton's syndrome, where objectively blind subjects believe that they can see. While the varieties of anosagnosic subtypes are beyond the scope of this review, it is interesting to note that anosagnosia is often accompanied by frontal and/or parietal lesions (Vuilleumier, 2004), and is closely related to the concept of insight in neuropsychiatric disorders. For example, in schizophrenia, lack of insight of one's condition is only weakly associated with the severity of delusions, with correlative studies instead implicating prefrontal pathology (Amador & David, 2004).

## 2.5 Outstanding questions

1. Several behavioural studies employing asymmetric payoffs have probed the integration of reward with perceptual accuracy (Davison & Tustin, 1978; Green & Swets, 1966; Johnstone & Alsop, 2000; Maddox, 2002; Maddox & Bohil, 2004). However, little is known about how value interacts with perceptual uncertainty in the brain (Gold & Shadlen 2002; see section 2.3.5); in particular, whether value affects an early or late stage of sensorimotor integration. Indeed, in a review of the field, Heekeren and colleagues note that value 'might affect sensory representations, as well as motor planning or action selection; however, how this occurs in the human brain is still an open question' (Heekeren *et al.*, 2008). In Chapter 4, I report an experiment that employed a modification of the design used by Whiteley & Sahani (2008) to ask how value biases the

neural stages of perceptual decisions.

2. As discussed in section 2.2.2, there are several competing hypotheses for how asymmetries in the loss function alter the decision threshold. One influential view is that cortico-basal ganglia mechanisms adjust decision criteria (e.g. Lo & Wang 2006), but evidence for the operation of such mechanisms in the human brain during decision-making is currently sparse (Bogacz *et al.*, 2010). In Chapters 5 and 6 I determine how biases in the criteria for acting during perceptual decisions affects basal ganglia activity measured using fMRI, with a particular focus on the STN.

3. Control of decision criteria is generally only required when incoming evidence is uncertain (section 2.2.4). The sensitivity of observers to dynamic changes in uncertainty suggests that the brain is able to monitor uncertainty and use this knowledge to effectively adjust decision strategies (section 2.4). It is unknown, however, the extent to which this uncertainty is available for use by other response systems, such as through verbal reports. By using metacognitive reports, I ask in Chapter 7 whether observers' post-decision ratings of decision confidence can be concisely described by a Type 2 SDT model, and whether such a model can identify differences in metacognitive ability across subjects.

4. The neural mechanisms of metacognitive monitoring are poorly understood (section 2.4.3). In Chapter 8, I harness the individual differences in metacognitive ability identified in Chapter 7 to identify variability in the structure of prefrontal cortex that correlates with metacognitive ability.

# Chapter 3

# Methodology

## 3.1 Overview

This chapter provides general background on the specific methods used in each experiment contained in the thesis. Signal detection theory (SDT) is used to analyse behavioural data in Chapters 4, 5 and 7, and a brief primer on the rationale behind the theory and measures of SDT is presented in section 3.2. The following sections review the principles behind magnetic resonance imaging, and its application to identify functionally relevant changes in both blood-oxygen level and the structure of grey and white matter.

## 3.2 Signal detection theory

As we saw in Chapter 2, a crucial problem in decision-making is dealing with uncertainty. Imagine an experiment where we repeat random states (present or absent) and record the value of an arbitrary internal state variable $X$; by doing so we can build up a picture of the conditional probability distributions over $X$ given a particular state of the world $h_1$ or $h_2$. These distributions overlap; it is only on average that the signal + noise distribution is higher than the noise-alone distribution (figure 2.1). The observer's job is to categorise a given value of $X$ as either arising from $h_1$ or $h_2$, signal or noise. The point at which the distributions cross, or $log(\beta) = 0$ (equation 2.3), is the optimal point to place a criterion (assuming uninformative priors).

Following placement of a criterion, the observer's job is simple – when $x > \beta$ choose $h_1$, else if $x < \beta$, choose $h_2$. Nevertheless, there will still be times when the observer makes mistakes, even for the optimal criterion – for '$h_1$' responses, these mistakes are termed false alarms, and are determined by the magnitude of the area under the $h_2$ curve to the right of the criterion. Correspondingly, we can splice up the other areas under the two curves to form a contingency table (table 3.1).

The performance of our decision-maker is dependent on how faithfully their cat-

| | | ‘$h_1$’ | ‘$h_2$’ |
|---|---|---|---|
| State of the world | $h_1$ | Hit | Miss |
| | $h_2$ | False alarm | Correct rejection |

**Table 3.1:** Type 1 signal detection theory contingency table.

egorisation of true states of the world is made. If hits are high and false alarms are low, then we can say the observer has a high sensitivity. Indeed, this difference between hits and false alarms is the basis of the measure of sensitivity inherent to signal detection theory (SDT):

$$d' = z(H) - z(FA) \tag{3.1}$$

where $z$ is the inverse of the cumulative normal distribution function (Green & Swets, 1966). The insight of SDT is that an observer's criterion placement is independent of his sensitivity. This can be immediately seen from figure 2.1: as the criterion is swept to the left, hit rate increases to a point at which all $h_1$ states are correctly classified. But this increase in hits is paralleled by a detrimental increase in false alarms, where $h_2$ states are incorrectly classified as $h_1$. The observation that the relationship between hits and false alarms is invariant across placements of the criterion is a cornerstone of SDT. If we assume a Gaussian generative model for hits and false alarms, $d'$ will increase as the separation between the means of the distributions increases (as the signal strength increases) and decrease as the variance of the distributions increases. Indeed, the use of $z$-transforms in equation 3.1 specifies $d'$ as the distance in units of standard deviation between the means of the assumed underlying signal and noise distributions.

As for sensitivity, SDT provides us with both a generative model and empirical description of bias. By assigning values to each of the outcomes of table 3.1, SDT specifies the optimal likelihood ratio criterion as being a function both of the payoff matrix and the *a priori* probabilities of each state of the world (equation 2.4; Dayan & Daw 2008; Green & Swets 1966). We can specify an empirical measure of bias by noting that the bias towards reporting $h_1$ independent of the state of the world is proportional to the sum of hits and false alarms. In SDT, this measure is scaled such that a value of $c = 0$ is indicative of a neutral criterion ($\beta = 1$):

$$c = -0.5[z(H) + z(FA)] \tag{3.2}$$

This treatment of sensitivity and bias in SDT allows one to go from a point estimate of $d'$ and $c$ in a psychophysical experiment (using table 3.1 and equations 3.1 and 3.2) to an assumed picture of the probability distributions upon which the decision is based (figure 2.1). However, this transformation relies on a leap of faith, in that point estimates of $d'$ and $c$ only connect to the underlying model under conditions in which the conditional probability distributions over $x$ given $h_1$ and

$h_2$ are normally distributed with equal variance. We can go one step further here by noting that the relationship between the two hypothetical distributions is fully specified by the (continuously valued) pairs of hit and false alarm rates generated by sweeping the criterion from left to right across the decision axis (figure 7.2).

The theoretical function generated by plotting hit rate against false alarm rate is known as a receiver operating characteristic (ROC) function (figure 2.1). As the only thing that changes when generating a ideal observer's ROC function is the position of $c$, it follows that $d'$ is constant for all points on an ROC curve, known as iso-sensitivity points. To measure an empirical ROC, all that is required is for an observer to adopt multiple criteria for a given level of sensitivity (e.g. through use of multiple confidence ratings), thus generating multiple hit-false alarm pairs which can be plotted in ROC space. The area under the empirical ROC is a measure of sensitivity that is assumption free with respect to underlying distributions (Kornbrot, 2006). Further, the assumptions of SDT can be checked by plotting the ROC in $z$-transformed space (Macmillan & Creelman, 2005). If the hypothetical $p(x|h_1)$ and $p(x|h_2)$ probability distributions are Gaussian, this transformation will yield a straight line whose slope is specified by the relative variance of the two distributions.

The same logic underlying Type 1 SDT measures can be applied to metacognitive reports about decision-making performance. This 'Type 2' SDT analysis was first devised by Clarke *et al.* (1959), but recently there has been a resurgence of interest in the method, in part spurred on by an in-depth derivation of the relevant probability distributions by Galvin *et al.* (2003). In Type 2 SDT, the 'evidence' which is being discriminated is the subject's own decision, rather than the state of the world. For post-decision confidence, a 'hit' is then a high-confidence correct judgment, and a 'false alarm' is a high confidence incorrect judgment (table 3.2). As for Type 1 SDT, if continuous confidence ratings are used, an ROC function relating how increasing confidence discriminates between correct and incorrect judgments can be derived (figure 7.1).

| | | 'High confidence' | 'Low confidence' |
|---|---|---|---|
| State of the world | Correct | Hit | Miss |
| | Incorrect | False alarm | Correct rejection |

**Table 3.2:** Type 2 signal detection theory contingency table.

## 3.3 Functional magnetic resonance imaging (fMRI) methods

Functional magnetic resonance imaging (fMRI) provides an elegant, non-invasive method for defining the neural mechanisms underlying human behaviour and subjective experience in healthy individuals. fMRI measures changes in local cerebral blood flow, providing an indirect picture of changes in brain activity over time. The specific nature of the linkage between blood flow and underlying neural activity is still an open area of investigation (Logothetis, 2008); however, recent surveys of the field indicate replicable mappings between changes in fMRI signal in particular brain regions or networks and particular cognitive functions (e.g. Poldrack *et al.* 2009). In this manner, fMRI, in tandem with cognitive psychology, has the potential to refine our concepts of the structure of the mind (Yarkoni *et al.*, 2010).

Initial studies in the field were concerned with functional localisation: identifying mappings from particular brain areas to particular cognitive functions in healthy subjects. More recent work has also focussed on functional integration (e.g. Friston *et al.* 2003), creating models of how one area's activity relates to, or putatively causes, another area's activity over time. The end-goal of this research is to identify a mechanism tying cognitive function to the large-scale interaction of multiple specialised brain regions. In this thesis, fMRI is used primarily in functional localisation mode, enabling answers to questions about which stage in the decision pathway (sensation, decision or action) contextual variables such as value and uncertainty become integrated into the decision process.

## 3.3.1 Principles of fMRI

### 3.3.1.1 Nuclear magnetism

In quantum mechanics subatomic particles such as protons have a property known as spin. Some atomic nuclei, because of the numbers of protons and neutrons they contain, have a non-zero spin and consequently a magnetic moment. When such nuclei are placed in a strong magnetic field they align around an axis along the direction of the field. Because this alignment is not perfect, nuclei 'precess' around the external field at a frequency known as the Larmor frequency. The Larmor frequency is determined both by the magnitude of the field (increasing as the field strength increases) and the nature of the nucleus itself (see figure 3.1).

Hydrogen nuclei (protons) align in two states, parallel and anti-parallel to the direction of the magnetic field. The anti-parallel state is a higher-energy state than the parallel one, so a slightly greater proportion of nuclei align themselves parallel to the field. This results in a net longitudinal magnetisation parallel to the external field, which increases with the field strength. Since the brain contains a large number of hydrogen nuclei, many of them in water, this kind of magnetisation occurs when it is placed in a magnetic field. The quantity of mobile protons in a tissue relative to water is referred to as its proton density.

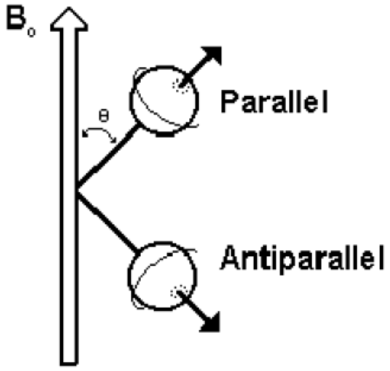

**Figure 3.1:** Hydrogen atoms align either parallel or anti-parallel to a magnetic field B0. Because alignment is not perfect (angle $\theta$) they 'precess' around the direction of B0 at the Larmor frequency.

#### 3.3.1.2 Generating an MR signal

If an oscillatory radio frequency (RF) pulse is applied perpendicular to the direction of the static magnetic field, resonant absorption of energy will occur in protons of any nuclei with a Larmor frequency matching those present in the pulse. In the case of hydrogen nuclei this means that some of the low-energy state protons absorb energy and move to the high-energy state, which tips the net magnetisation in the transverse plane. A sufficiently large pulse can tip the net magnetisation sideways (a 90° pulse) or even reverse it (a 180° pulse). This transverse angle is called the flip angle of the RF pulse, and all experiments reported here used a 90° flip angle.

After the RF pulse nuclei tend to recover their original orientation, giving off radio waves (photons) as this occurs. These emissions form the basis of MR images. This phenomenon, known as relaxation, occurs in two ways, longitudinal or T1 relaxation, and transverse or T2 relaxation, both of which are important in functional brain imaging. An image can be constructed because the protons in different tissues return to their equilibrium state at different rates.

#### 3.3.1.3 Generating an MR image

To produce a three-dimensional MR image it is necessary to distinguish between different spatial locations. This is achieved using three additional magnetic fields containing spatial gradients (gradient fields). These fields, aligned orthogonally to one another (usually on the superior-to-inferior ($z$) axis, left-to-right ($x$) axis, and posterior-to-anterior ($y$) axes) are known as the slice-select gradient, the readout or frequency-encoding gradient, and the phase-encoding gradient, respectively. Typically the gradient fields are discretely stepped, allowing the user to partition the image into small cubed elements (volume elements or voxels). All protons within a voxel are combined in the reconstructed image, and voxel size thus determines the maximum resolution of the image.

The slice-select gradient is switched on briefly at the same time as the RF pulse, creating a gradient in the magnetic field. Since the Larmor frequency of hydrogen nuclei is proportional to the field strength, it will vary along the slice-select gradient. This means that by confining the RF pulse to specific frequency ranges it is possible to affect the magnetisation vector in only a specific region along the slice-select axis (in a two-dimensional brain slice). To differentiate between locations within a 2D slice requires two additional gradient fields. The frequency-encoding gradient is applied during measurement of the signal (hence 'readout gradient'), and alters the precession frequency within a slice. Finally, phase-encoding is applied briefly between the RF pulse and measurement. Phase-encoding generates a gradient of precession frequencies, resulting in the nuclei at different locations along the axis of the phase-encoding field becoming out of phase. When the phase-encoding field is shut off they return to precessing at the same frequency, but maintain their altered phase. Both of these gradients allow the emitted radio wave to be differentiated in the Fourier spectrum of the MR signal, as the frequency of the released photons depends on position in a predictable manner.

#### 3.3.1.4 Different kinds of scan

Although the fundamentals remain the same, altered scanning parameters and procedures are used to optimise acquisition of different kinds of data. Two key parameters are the repetition time (TR) between two consecutive pulses, and the echo time (TE) between the RF pulse and the measurement of the signal. T1-weighted scans (which maximise the T1 difference between tissues) typically have a short TE and TR (e.g. 20ms and 500ms). T2/T2*-weighted scans usually have a long TE and TR (e.g. 80ms and 2000ms).

Echoes are signals produced by additional 180° RF or gradient pulses, and are used to resynchronise the precession of the nuclei to allow collection of further signal. Echo-planar imaging (which is used in all the studies reported in this thesis) makes use of echoes to acquire images with a tolerable signal-to-noise ratio fast enough for fMRI (typically once every 2-3 seconds; Stehling *et al.* 1991). Here an initial RF pulse is followed by a series of 180° refocusing pulses, each sandwiching separate phase and gradient-encoding steps.

#### 3.3.1.5 BOLD fMRI

Haemoglobin, the oxygen-transporting protein found in red blood cells, has different magnetisation properties depending on whether it is oxygenated (oxyHb) or not (deoxyHb). DeoxyHb is paramagnetic due to the presence of unbound-iron containing haem-groups. The presence of deoxyHb in red blood cells induces a difference in magnetic susceptibility between blood and surrounding tissue. In the large homogenous magnetic fields used in MRI, compartmentalised susceptibility differences induce small magnetic field distortions in the blood. Water protons in these areas are affected by the field distortion, altering the T2* relaxation time. When the content of deoxyHb changes in the blood, the relaxation process of water protons is modified, producing differences in the T2* signal. Pioneering work both in animals (Ogawa *et al.*, 1990) and humans (Ogawa *et al.*, 1992) showed that these changes can be reliably measured with fMRI.

Changes in blood oxygenation is related to brain activity due to metabolic activity by neurons requiring oxygen (Heeger & Ress, 2002; Logothetis, 2008; Ogawa *et al.*, 1992). The time course of the BOLD response to transient neuronal activity is now fairly well characterised (Heeger & Ress 2002; see figure 3.3). Neuronal activity increases metabolic demand, transiently increasing the concentration of deoxyHb in local vasculature. This is followed after a delay of ∼1-2 seconds by a large increase in local blood flow which leads to an oversupply of oxygenated blood, leading to a decrease in deoxyHb and consequent increase in BOLD signal peaking at around 6 seconds after the onset of activity. This is followed by an undershoot which lasts several seconds. Although measuring the initial dip in blood oxygenation might be the most sensitive way of exploring BOLD changes, these changes are very small, to the extent that their existence remains somewhat controversial (Heeger & Ress, 2002). Consequently, fMRI analysis tends to concentrate on clearly identifiable BOLD peaks. The time course of the BOLD response provides a natural limit to the temporal resolution of fMRI.

There is less consensus about how to relate BOLD changes to specific patterns of neuronal activity, although one finding is that BOLD changes are more tightly coupled to synaptic activity (inputs to cortical regions) rather than cell firing (Heeger & Ress, 2002; Logothetis, 2008). In particular it is not clear how BOLD changes relate to inhibitory modulatory activity, or situations where both inhibition and excitation are increased at the same time within a confined region of the brain (Logothetis, 2008). Furthermore, the interpretation of a decrease in BOLD signal relative to a resting baseline remains controversial (Lin *et al.*, 2010). These concerns do not invalidate fMRI as a powerful investigative tool, but they should be borne in mind when designing and interpreting fMRI studies.

### 3.3.2 fMRI preprocessing

In order to transform the acquired functional images into a format suitable for analysis, a number of preprocessing stages are necessary. Before preprocessing the first few images are typically removed from analysis to allow for T1-equilibration effects. Analyses reported here were carried out in SPM5 (Chapters 4 and 5) or SPM8 (Chapter 6) (Wellcome Trust Centre for Neuroimaging, London, `www.fil.ion.ucl.ac.uk/spm`), with each stage described in the sections that follow.

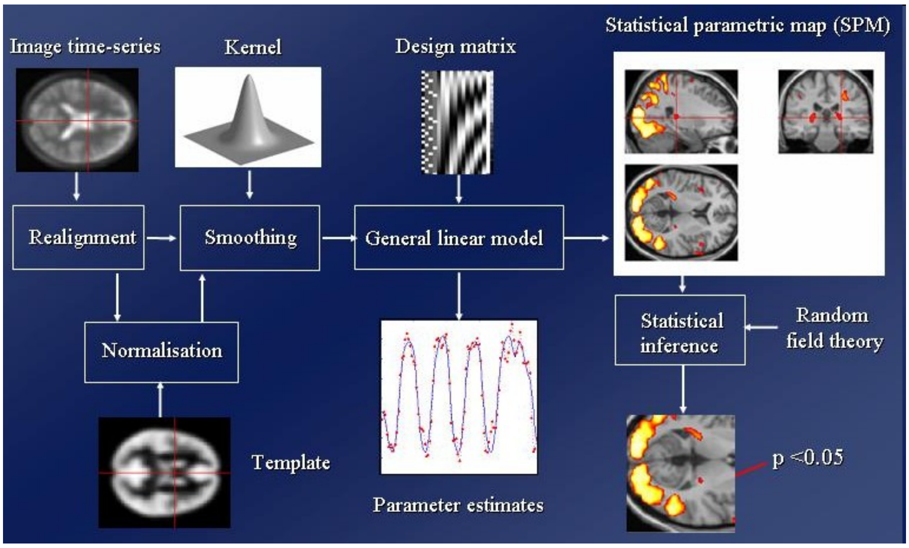

**Figure 3.2:** Schematic depicting the processing stages that start with a raw imaging data sequence and end with a statistical parametric map (SPM). Voxel-based analyses require that data are in the same anatomical space: this is achieved through realignment of the data. After realignment, the images are normalised and smoothed. The general linear model is then employed to estimate the parameters of an analysis model (depicted here by a design matrix) and derive the appropriate univariate test statistic at every voxel. The test statistics (usually $T$ or $F$-statistics) constitute the SPM. Statistical inferences are made based on the SPM and Random Field Theory and characterise the responses observed using the parameter estimates (reproduced from Flandin & Friston 2008).

#### 3.3.2.1 Realignment and unwarping

The image time series first needs to be realigned to a common reference frame to correct for any head movements during scanning. This is performed using a rigid-body affine transformation, with the reference frame taken as the first image in the acquired time series. Even once realignment has been performed, considerable movement-correlated variance is found in the time series due, among other things, to the existence of magnetic field inhomogeneities, and consequent movement-by-inhomogeneity interactions (Andersson *et al.*, 2001). In other words, the BOLD contrast in a particular region will be subtly different if that part of the brain is moved into a different part of the magnetic field, considerably reducing the statistical power of a subsequent analysis of task-related activation. One method to alleviate this issue is to include the movement parameters as covariates in the statistical model used to analyse the data. A drawback is that if movements are task-correlated, activations of interest may be removed. An alternative method (unwarping) is to acquire fieldmaps, images which are used to estimate inhomogeneities in the magnetic field and then generate a forward model of the movement-by-inhomogeneity interactions (Andersson *et al.*, 2001). The imaging studies reported in Chapters 4-6 incorporate a combination of unwarping and realignement covariates to deal with movement-correlated activity.

#### 3.3.2.2 Normalisation

To aid anatomical interpretation and standardisation between studies, images are typically transformed into a standard space. This is performed using a three-stage process. First the mean realigned/unwarped image is coregistered with the subject's T1-weighted structural image using a rigid-body transformation estimated by maximising the mutual information between the two images. The structural image is then 'segmented' (Ashburner & Friston, 2005) into separate grey- and white-matter images using a nonlinear deformation field to map it onto template tissue probability maps. This mapping is then applied to both the structural and functional images to create spatially normalised images. Normalisation is to Montreal Neurological Institute (MNI) space.

#### 3.3.2.3 Smoothing

For analysis under Gaussian Random Field Theory assumptions (described below), functional images must be spatially smoothed with a Gaussian filter. In addition, smoothing improves the signal-to-noise ratio, though at the cost of spatial resolution. This is done by convolving the images with Gaussian kernels with full-width at half-maximum (FWHM) of 8mm (Chapters 4 and 5) or 6mm (Chapter 6).

### 3.3.3 Statistical testing

The most common way to analyse fMRI time series is with a mass-univariate approach. This involves performing a statistical test separately at each voxel in the image. The resulting statistics can then be assembled into an image – a statistical parameteric map (SPM) – and improbable patterns attributed to an experimental factor (figure 3.2). Typically, this statistical testing is done using a General Linear Model (GLM). All statistical tests reported in this thesis were carried out in SPM5 (Chapters 4 and 5) or SPM8 (Chapters 6 and 8; Wellcome Trust Centre for Neuroimaging, London, `www.fil.ion.ucl.ac.uk/spm`).

#### 3.3.3.1 The General Linear Model

The General Linear Model (GLM) is a broad framework of which most of the statistical tests typically used in neuroimaging (e.g. t-test, ANOVA, linear correlation) are special cases (Flandin & Friston, 2008). At the core of the GLM is an equation relating a matrix $Y$, containing observations of BOLD signal, to a linear combination of explanatory (predictor) variables, contained in the design matrix $X$:

$$Y = X\beta + \epsilon \tag{3.3}$$

where $\beta$ is a vector containing the parameters to be estimated and $\epsilon$ is a residual term (Friston *et al.*, 1994a). (This is also often described as multiple regression analysis, see Howell 2009). The GLM approach assumes that the residuals are independently and identically distributed (IID) (a condition often described as ‘sphericity’, or, equivalently, as having a multivariate normal distribution). This condition is not met by fMRI time series, so a correction must be applied to impose sphericity (Glaser, 2004).

In analysing fMRI data the design matrix consists of the experimental manipulations, confounds and covariates of no interest, each entered as a separate column. These columns are referred to as regressors. Experimental events are typically modelled as either stick (delta) or boxcar functions and then convolved with a haemodynamic response function (HRF) before being entered into the design matrix (see figure 3.3). In SPM, the $\beta$ parameters are then estimated using a restricted maximum likelihood (ReML) algorithm.

Statistical inferences about parameter estimates are made using their estimated variance. Two sorts of test are possible: an $F$-statistic testing the null hypothesis that the parameter is zero, and a $T$-statistic testing whether some linear combination of estimates is significantly different from zero (and is therefore directional). Applied at each voxel, an image of $T$- or $F$-statistics across the entire brain volume (the SPM) is produced. Regressors can be categorical, such as a 1 or 0 to index the onset of a particular stimulus condition, or parametric, such that the height of the regressor is modulated by the quantity associated with the current trial. An example of the latter would be modulating a regressor’s height by the subject’s reaction time (see also Grinband *et al.* 2008), isolating activity that covaries with reaction time. The experiments in this thesis make use of both categorical and parametric regressors.

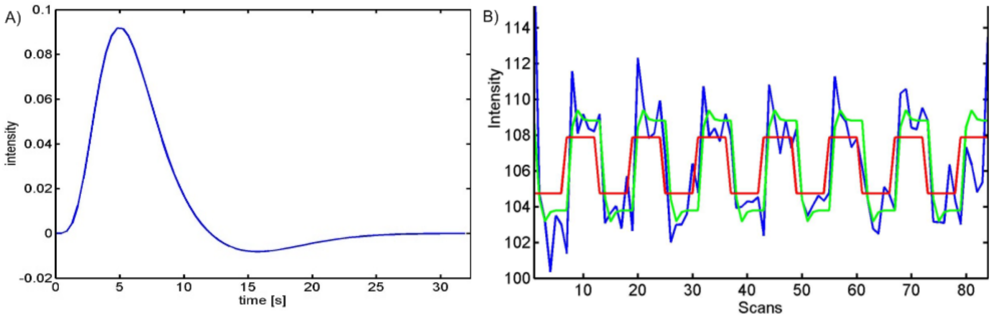

**Figure 3.3:** Left panel: canonical haemodynamic response function. Right panel: time series showing the model of the stimulus (red), model after convolution with canonical HRF (green), and observed data from a single voxel (blue). The latter two terms form single columns of each of the $X$ and $Y$ matrices in equation 3.3.

#### 3.3.3.2 Correction for multiple comparisons

In fMRI the mass-univariate approach involves thousands of separate tests across the whole-brain volume, and this presents a large multiple-comparison problem that must be addressed before interpreting regions as being significantly ‘active’. One way around this problem is to specify a priori a specific brain structure of interest (such as the STN in Chapter 6), and extract contrast estimates only from this region for further analysis. However, if the analysis is more exploratory, or is used as a first step to define regions for subsequent analysis, large-scale corrections for multiple comparisons are required. One common approach is to use family-wise error (FWE) correction to control the probability of making one or more Type 1 (false positive) errors. In classical statistics with multiple independent tests FWE correction is often implemented through a Bonferroni correction (dividing the acceptable Type 1 error rate, $\alpha$, by the number of statistical tests being carried out). However, because of the large number of comparisons and spatial correlations between neighbouring voxels, Bonferroni correction is extremely conservative for neuroimaging data. Alternative approaches are therefore used to refine this issue and achieve an appropriate balance between sensitivity and specificity.

One method for increasing the sensitivity of FWE-corrected tests is to appeal to Random Field Theory (Adler, 2009; Friston *et al.*, 1995). Functional images show a strong degree of spatial correlation (indeed, this is ensured by the smoothing step in preprocessing; see section 3.3.2.3), leading to covarying clusters of voxels ('resolution elements' or 'resels'). It therefore makes sense to control false positives at the level of these clusters, rather than individual voxels. Moreover, we are interested in functional brain *regions*, not individual voxels in the image. Since there are always fewer clusters than voxels this allows for more sensitive FWE-corrected tests. A cluster-level inference tests if the number of activated voxels in a particular cluster (the cluster volume) is greater than would be expected by chance, whereas voxel-level inference tests only the peak value in that cluster. Cluster-level inference tends to be more sensitive to extended activations that may not contain a definite peak. In contrast, voxel-level inference permits greater spatial localisation (Poline *et al.*, 1997). Cluster-level inference requires a 'cluster-defining threhsold', such as $P < 0.001$, uncorrected. While the cluster-defining threshold is somewhat arbitrary, simulations show that the assumptions behind cluster-level correction hold down to a defining threshold of around $T = 2.5$, and allow inferences to be based on a combination of peak height and spatial extent (Friston *et al.*, 1994b). The statistical inferences made in this thesis at the whole-brain level are carried out at the cluster-level (see section 3.4.3 for specific considerations regarding voxel-based morphometry analyses).

A different (and often complementary) approach is to restrict the search volume to a particular region of interest (ROI). This can be defined either as an anatomical region, or by specifying a volume around a particular location, usually the peak activation found in a previous study. Once such regions have been defined, a small volume correction (SVC) can be implemented using one of the previously described multiple comparisons techniques. An alternative approach for defining ROIs is to use orthogonal contrasts to restrict the search volume to task-relevant activations. For example, in Chapter 4, I identify regions in ventral visual cortex that are active in proportion to the amount of 'face' information in an image. These active voxels are then used as ROIs for further analysis of whether face-sensitive regions are affected by the (orthogonal) effect of changes in the cost structure of the task. This procedure is valid to the extent that ROI selection is independent of contrasts subsequently applied to analyse the signal from this ROI (Kriegeskorte *et al.*, 2009; Vul & Kanwisher, 2010).

#### 3.3.3.3 Group analyses

The above procedure describes how parameter and contrast estimates are generated at the single subject level. Typically fMRI studies consider data from several subjects. In order to generalise the results of this group analysis to the population, one has to incorporate the between-subject variance of the estimates. This can be implemented using a two-stage 'summary statistics' procedure (Friston *et al.*, 2005) in which contrast estimates at the first-level (individual subjects) are treated as a new response variable $Y$ in a second-level (group) analysis (equation 3.3). (This is called a 'random effects' analysis, and makes the assumption that first-level parameter estimates are normally distributed in the population). This second-level model can then be analysed in precisely the same way as the first-level models via the GLM. This approach provides a good approximation to the ideal, but computationally expensive, mixed effects analysis (Friston *et al.*, 2005).

## 3.4 Voxel-based morphometry methods

Voxel-based morphometry (VBM) proceeds in a very similar fashion to analysis of functional MRI timeseries, except now the image is static in time (there is a single observation). The goal of VBM is to isolate differences in structure, either between groups (e.g. schizophrenics and controls; Kubicki *et al.* 2002), timepoints (e.g. before and after learning a motor skill; Draganski *et al.* 2004) or different levels of a psychological profile (e.g. structure covarying with language learning; Carreiras *et al.* 2009). The analysis proceeds by identifying a particular tissue type (grey or white matter) using segmentation algorithms, and warping these segmented images to a common anatomical space. The original methods proposed for VBM relied on the fact that these warps would be necessarily imperfect (low-dimensional), and thus the residual differences in structure in a region would remain despite macroscopic differences (such as changes in head size) being eliminated by the normalisation routine (Ashburner & Friston, 2000). With the advent of high-dimensional warping routines that can be carried out using readily available computing power, current approaches scale the pre-processed data such that the total volume of tissue in each structure is preserved (Ashburner & Friston, 2001). 'Modulated' VBM thus measures the tissue volume per unit volume of spatially normalised image. The steps in this analysis are outlined next.

### 3.4.1 Segmentation

Segmentation algorithms rely on a high-contrast between grey and white-matter in the MRI image. The 'unified segmentation' routine in SPM5/8 assumes that every voxel belongs to one of four tissue classes: grey matter, white matter cerebrospinal fluid and everything else (Ashburner & Friston, 2005). Image intensity inhomogeneity is modelled out of the data using a linear combination of low-frequency 3D cosine transform spatial basis functions. A 'mixture of Gaussians' model is fit to the variation in intensity values across the brain to map each pixel into one of the four tissue classes. Assigning a voxel to a tissue class also relies on tissue priors, incorporating knowledge of how common a particular tissue is at a particular location in the image. These prior probabilities are encoded in template tissue probability maps, which are combined with the likelihood of the model given the data (the mixture of Gaussians fit) in a Bayesian fashion to complete the segmentation. More broadly, the SPM5/8 segmentation approach combines inhomogeneity correction, tissue classification and nonlinear registration (warping) within a single probabilistic generative model.

### 3.4.2 DARTEL registration

The unified segmentation routine discussed in the previous section generates warps to MNI space as part of the generative model. However, these warps are relatively low-dimensional ($\sim 1000$ parameters), and modulated VBM requires very accurate warps to be able to localise regional differences in volume. As a solution to this approach, a separate registration model has been developed using in the order of $6,000,000$ parameters (Ashburner, 2007). A recent empirical validation showed that the accuracy of DARTEL registration is much higher than that achieved by other intersubject registration approaches (Klein *et al.*, 2009). In this study, measures of accuracy were based on the amount of overlap achieved for regions of the brain manually delineated by human experts.

DARTEL involves an iterative procedure that alternates between computation of a cohort-specific template, and warping all subjects' tissue probability maps (produced by segmentation) into increasingly good alignment with this template. This optimisation is again carried out in a Bayesian fashion, by minimising the negative log posterior probability of the model given the data. DARTEL creates a 'flow-field' for each subject, which encodes how the individual grey- and white-matter tissue probability images should be warped, or deformed, to best-match the average shape of the template. While this template (created from a large number of subjects) tends to reflect the 'common average' of multiple brains, it may not be well-aligned to MNI space, and a further spatial transformation can be applied to permit reporting of results in MNI coordinates.

As mentioned above, a high-dimensional warp 'throws away' information about fine-grained volumetric differences. In other words, in the ideal case where image registration is exact, there would be no volume differences left to examine. In contrast, by multiplying the warped images by the Jacobian determinants of the deformation, information about volume is retained. For example, if one subject's temporal lobe is much larger than that of the group, the large deformations required to warp this lobe to a common template will multiply the signal intensity of this region. Thus the signal from any region following modulation can be interpreted as the tissue volume per unit of spatially normalised image.

### 3.4.3 Statistical analysis

As for functional brain imaging data, the warped, modulated tissue-class images are blurred through a smoothing step. Smoothing alleviates problems caused by inaccuracies in inter-subject registration, and allows the use of Gaussian random-field theory correction during statistical analysis, as outlined in section 3.3.3.2. The study reported in Chapter 8 uses a 8mm FWHM Gaussian kernel for smoothing.

Statistical analysis proceeds using the same GLM framework outlined in section 3.3.3.1 for fMRI data. Here one image per subject (or per time-point, in a longitudinal study) is entered into a multiple regression model to identify regions of volume change that covary with the parameter or group membership of interest. If there were $N$ subjects in the study, the design matrix contains $N$ rows. It is also important to model covariates of no interest that may affect the statistical interpretation of changes in regional volume (Barnes *et al.*, 2010). For instance, it is known that men have systematically larger brains with proportionally more white matter than women (Luders *et al.*, 2006); in the statistical model used in Chapter 8, a covariate indicating gender group membership is included to account for this effect. Similarly, 'global' differences in brain volume are taken into account by scaling each voxel's value to be a proportion of the total volume of a particular tissue class.

Inference as to whether regional volume significantly correlates with one or more regressors of interest is performed in the manner outlined in section 3.3.3.2. Gaussian random field theory is used to correct for multiple comparisons across the whole-brain volume. One exception is the use of cluster-based statistics – during the original development of VBM, it was observed that cluster-based inference produces an unreasonable number of false positives (Ashburner & Friston, 2000), due to the calculation of the expected number of clusters depending on local variations in the smoothness of the image. In fMRI, 'stationarity' of smoothness across the whole image is assumed, and this assumption is usually a good approximation to reality. However, SPMs generated from tissue probability data violate the stationarity assumption (Hayasaka *et al.*, 2004). Implementing cluster-based inference (which is preferred for the reasons outlined in section 3.3.3.2) requires correction for non-stationarity of smoothness (Hayasaka *et al.*, 2004). Computational simulations show that for designs with high degrees of freedom and sufficient smoothness (as used in Chapter 8), using a cluster-defining threshold of $P < 0.001$ with correction for non-stationarity provides adequate control ($P < 0.05$) over the family-wise false positive rate (Hayasaka *et al.*, 2004).

It is important to note that statistically significant regions in a VBM analysis only reflect regional differences in volume to the extent that the segmentation and registration steps are effective. Furthermore, real increases in volume in a particular region may be due to increases in folding, increases in cortical thickness, or a combination of the two. These considerations are illustrated in figure 3.4, showing examples of how differences in tissue shape and registration affect the total volume encoded by a voxel value.

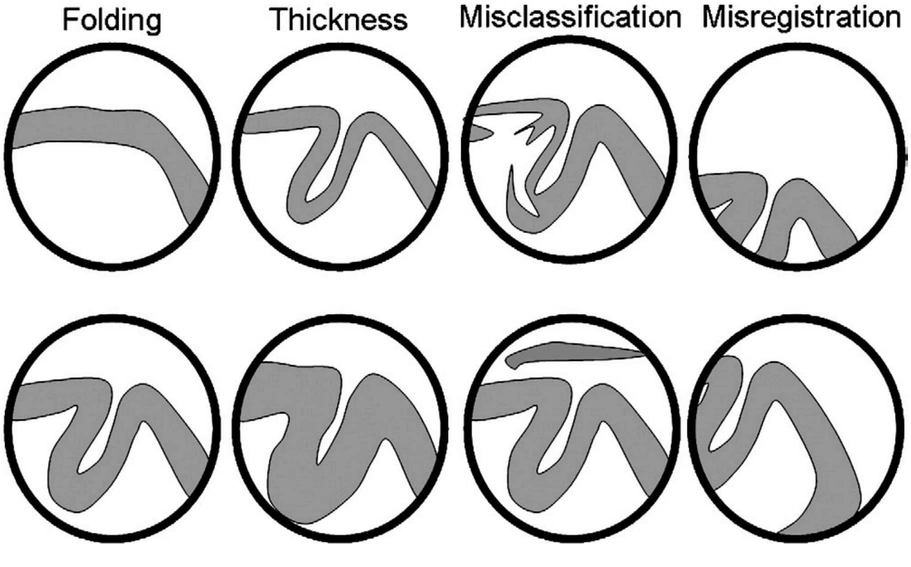

**Figure 3.4:** This illustrates how findings from a VBM study of grey matter could be interpreted. The top row shows situations where there would be less grey matter in a cortical region compared to the situation shown below it. From left to right, differences could be attributed to folding, thickness, misclassification or misregistration. Generally, the objective is to interpret differences in terms of thickness or folding. Figure reproduced with permission from Ashburner (2009).

## 3.5 Diffusion-tensor imaging methods

Water molecules in the brain move over distances of around 20$\mu$m every 50ms, bouncing off, traversing or interacting with other tissue components such as cell membranes and fibres (Bihan, 2003). Diffusion occurs in three dimensions, but molecular mobility in tissues is not necessarily isotropic (i.e. not necessarily the same in all directions). For example, the diffusion constant is lower across compared to along white-matter fibre bundles. In MRI, the effect of diffusion in the sample of interest is to reduce the signal, and is considered an artifact. However, it has been determined that the extent of reduction in the signal can be related to the diffusion constant of the sample. Because this constant can be different in different spatial directions (i.e. anisotropic), several directionally-specific measurements are needed. These measurements are encoded in the MRI signal by using large, bipolar magnetic field gradient pulses which are played out along the direction of interest. The effect of the tendency for diffusion to prefer one direction over another – anisotropy – can thus be quantified through observation of variations in the diffusion measurements when the direction of the gradient pulses is systematically changed.

Diffusion anisotropy in white matter originates from its organisation in bundles of fibres running in parallel. Original diffusion imaging studies demonstrated that this feature could be used to track the orientation in space of the white matter tracts in the brain (Douek *et al.*, 1991). By fitting a diffusion tensor at each voxel (Basser *et al.*, 1994), anisotropy effects in diffusion MRI data can be quantified. The methods behind DTI scanning and analysis are covered below, with a focus on the measure employed in the Chapter 8, fractional anisotropy (FA).

### 3.5.1 Diffusion-tensor imaging acquisition

Diffusion acts to attenuate the MRI signal in a manner dependent on the $b$-value, which characterises the gradient pulses used in the DTI sequence, and the diffusion coefficient $D$. A diffusion tensor describes the attenuation in each direction of diffusion at each voxel, and the covariance between these attenuations:

$$D = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yz} \\ D_{xz} & D_{yz} & D_{zz} \end{bmatrix} \tag{3.4}$$

In order to estimate the diffusion tensor, one must acquire diffusion-weighted images along several gradient directions. As the diffusion tensor is symmetric (equation 3.4), determination of its 6 unique elements requires a minimum of 6 measurements with non-collinear directions. In practice sampling more than 6 directions improves tensor fitting and signal-to-noise ratio; in the study reported in Chapter 8 68 gradient directions are measured. In addition, DTI data was acquired twice to correct for image distortions introduced by magnetic field inhomogeneities; these methods are covered in more detail in Chapter 8.

### 3.5.2 Extracting fractional anisotropy

Once diffusion-weighted images are collected, the tensor is fitted using multiple regression at each voxel (see Bihan *et al.* 2001) given knowledge of the $b$-value and gradient encoding directions. In this thesis I use algorithms implemented in FSL v 4.1 (`http://www.fmrib.ox.ac.uk/fsl/`) for DTI preprocessing and tensor fitting. The diagonal components of the diffusion tensor (the eigenvalues; $\lambda_1 = D_{xx}, \lambda_2 = D_{yy}, \lambda_3 = D_{zz}$) are then used to define fractional anisotropy through the following equation:

$$FA = \sqrt{\frac{3}{2}} \frac{\sqrt{(\lambda_1 - \lambda_2) + (\lambda_1 - \lambda_3) + (\lambda_2 - \lambda_3)}}{\sqrt{(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)}} \tag{3.5}$$

This calculation assigns each voxel a single FA value, which is a measure of the fraction of the magnitude of $D$ that can be ascribed to anisotropic diffusion. In other words, FA indicates the asymmetry in the directionality of diffusion, where this asymmetry can be oriented in any direction.

### 3.5.3 Statistical analysis

Inference as to whether regional FA significantly correlates with one or more regressors of interest is performed in the manner outlined in section 3.3.3.2. FA maps can be analysed using voxel-based techniques. In Chapter 8, warps (DARTEL flowfields) calculated during VBM preprocessing are applied to the FA images (after coregistration), and voxel-based analysis is carried out in a similar manner to the analysis of grey matter (including correction for smoothness non-stationarity). The one difference between the grey-matter and FA analyses is that the latter images are not modulated, as FA is not a volumetric measure.

FA is thought to underlie the microstructural integrity of a particular white-matter tract, although the precise functional implications of changes in FA remain to be determined (Alexander *et al.*, 2007). Care must therefore be taken when interpreting statistical differences in FA. Furthermore, similar caveats to those flagged for VBM analysis apply to voxel-wise analysis of FA (figure 3.4): imperfections in image registration and partial volume effects can lead to spurious differences in FA. In Chapter 8 I attempt to militate against these factors through high-resolution DARTEL warps and use of a mask to restrict analysis to white-matter tracts. Finally, one additional caveat is that regions where two or more fibres cross will tend to have low FA, despite each individual fibre having high FA. Thus knowledge of the expected white-matter crossing profile in a particular region is required to correctly interpret FA values, and, in regions with significant crossing, changes should be interpreted with care.

Assuming FA reflects the microstructural integrity of a white matter tract, any changes in this property may affect brain function and behaviour through affecting nerve conduction and signalling along the tract. Pathological conditions, such as optic neuritis and multiple sclerosis, provide extreme examples of the behavioural consequences of a loss of white matter integrity. In healthy populations, changes in white matter function have been observed following training on particular motor skills. For example, white matter FA in several tracts was seen to correlate with the amount of time spent practicing the piano in professional musicians (Bengtsson *et al.*, 2005). Similarly, increases in FA following juggling practice (compared to a control group) were found in tracts underlying primary motor cortex (Scholz *et al.* 2009; see Draganski *et al.* 2004 for a similar study reporting changes in grey matter volume). The relevance of FA changes for alterations in functional connectivity was investigated by Boorman *et al.* (2007): here, the magnitude of paired-pulse TMS effects measured between premotor and motor cortices correlated with FA variation along pathways believed to mediate the stimulation pulses.

# Chapter 4

# Effects of category-specific costs on neural systems for perceptual decision-making

> ‘How many fingers, Winston?’
> ‘Four. I suppose there are four. I would see five if I could. I am trying to see five.’
> ‘Which do you wish: to persuade me that you see five, or really to see them?’
> ‘Really to see them.’
>
> *Nineteen Eighty-Four*, George Orwell

## 4.1 Introduction

In Chapters 1 and 2 we saw how the loss function in Bayesian decision theory could be theoretically applied at one of several stages of visuomotor processing. Signal Detection Theory (SDT) makes it clear that decisions, including those involved in simple sensory judgements, necessitate decision thresholds (Green & Swets, 1966). These thresholds provide a means to splice up noisy sensory input and recover the most likely causes in the environment of a signal. The relatively simple solution provided by SDT is that a decision criterion is applied to the fixed, unchanging sensory evidence on any given trial. Any changes to this criterion, for instance brought about by the influences of prior expectation and reward, are said to occur downstream from the accumulation of sensory evidence. However, while this theoretical dissociation between the compilation of evidence and incorporation of utility is inherent to SDT (and several more complex models of perception), there is no *a* *priori* reason to expect that its neural implementation should neatly reflect such a division of labour. Thus, the issue of where stimulus value exerts its effects within the sensorimotor transform remains an unresolved empirical question (figure 1.2).

Evidence from psychophysics suggests that prospective costs have strong effects on human perceptual decision criteria (Green & Swets, 1966; Whiteley & Sahani, 2008; Landy *et al.*, 2007). Changes in value linked to particular regions of space are thought to alter intermediate representations between sensory coding and motor planning (Liston & Stone, 2008), and to modulate spatially selective regions of early visual (Serences, 2008) and somatosensory (Pleger *et al.*, 2008) cortex, potentially via recruitment of fast attention-like mechanisms (Maunsell, 2004; Serences, 2008). However, it is unclear whether costs associated with particular categorical outcomes, such as deciding between the presence of a face or house in a noisy input, are similarly mediated via category-sensitive visual areas (the left-most arrow in figure 1.2).

An alternative suggestion is that potential losses and gains exert their bias at a decision stage, either in frontoparietal regions thought to accumulate category evidence (Heekeren *et al.* 2004; Tosoni *et al.* 2008; Ho *et al.* 2009; Philiastides *et al.* 2006; Philiastides & Sajda 2007; Ploran *et al.* 2007; Thielscher & Pessoa 2007; Pleger *et al.* 2006; but see McKeeff & Tong 2007), or via cortico-basal ganglia loops important for controlling the threshold for action (Bogacz *et al.*, 2010; Lo & Wang, 2006). This suggestion would map onto the middle arrow in 1.2, and is supported by recent single-unit recording evidence showing that inducing shifts in decision criteria through changing a learnt category boundary (the speed of moving dots) modulates neural firing in frontal eye fields (Ferrera *et al.*, 2009).

To examine how prospective losses bias perceptual categorisation, I manipulated the costs associated with visual categories (faces and houses) while obtaining brain data using functional magnetic resonance imaging (fMRI)[1]. I predicted that if biases are expressed through changes in classically defined object representations, I should observe asymmetries in the activity of face- and house-selective regions located in fusiform and parahippocampal gyri in inverse proportion to the loss associated with a particular category. Alternatively, if losses solely bias evidence accumulation, effects of category-specific cost may be restricted to fronto-parietal regions known to compare evidence for perceptual decisions. A third possibility is that payoffs are integrated at a post-decision stage, at the level of response coding (the right-most arrow of 1.2). This possibility would predict that biases towards or away from a particular category (here, face or house) are expressed at the level of the very effector used to indicate a response to that category (here, the left or right hand). By design, response hand was orthogonal to the perceptual category, allowing us to specifically examine such effector-specific biases in activity induced by an asymmetric loss function.

[1] This study was carried out in collaboration with Louise Whiteley, Oliver Hulme and Maneesh Sahani. See page 14 for details of contributions.

## 4.2 Methods

### 4.2.1 Subjects

Nineteen right-handed subjects participated in the psychophysics session (7 male; 19 – 44 years of age; mean age, 25.0 years). All had normal or corrected-to-normal vision, and no history of psychological or neurological illness. Of these participants, sixteen were scanned. One participant was excluded at this stage due to a change of response strategy in the scanner that led them to disregard the face/house image, leaving fifteen subjects (5 male; 19 – 27 years of age; mean age, 23.9 years) in the analysis. The study was approved by the Institute of Neurology (University College London) Research Ethics Committee.

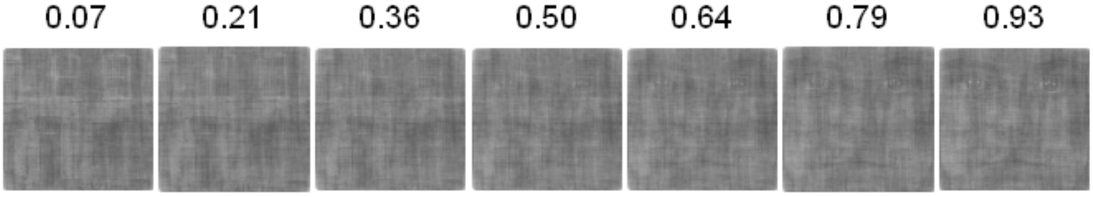

**Figure 4.1:** Example Fourier phase transition from a single house image to a single face image.

### 4.2.2 Stimuli

I used 10 neutral faces (5 male, 5 female) from the KDEF face set (Lundqvist et al. 1998) and 10 houses (photographed by the author). The stimuli were all cropped to be of equal size and converted to grayscale. To create a stimulus continuum, I adapted a technique used by Heekeren et al. (2004). Fourier transforms (FT) of each image were computed, producing 20 magnitude and 20 phase matrices. The average magnitude of all house and face stimuli was then stored. On each trial, a linear combination of one randomly selected house and face phase matrix was computed, plus a constant proportion (0.35) of a stored white noise matrix (figure 4.1). The resulting phase matrix was then recombined with the average magnitude matrix of the whole stimulus set using an inverse FT. Finally each image was normalised to have average luminance equal to that of the screen background and constant RMS contrast.

Face/house images were presented for 100ms on a grey background using Cogent 2000 (`www.vislab.ucl.ac.uk/cogent.php`) running in MATLAB. In the psychophysics experiment, stimuli were presented using a 20.1 inch Dell 2001FP monitor running at a refresh rate of 60 Hz, situated in a dimly lit room. All images subtended 4 degrees of visual angle at a viewing distance of 60cm. During the fMRI experiment, stimuli were presented using an NEC LT157 LCD projector, viewed by subjects via an adjustable mirror. At the beginning of each scanning session, a custom-written Cogent routine adjusted stimulus size and position to match that used in the psychophysics.

### 4.2.3 Psychophysics

The experiment was divided into two separate sessions. The first session involved acquiring psychophysics data outside the scanner; the second session repeated the same task during fMRI data acquisition. Participants were not informed of the image continuum, but were instead asked to categorise each briefly presented stimulus as either a noisy face or house. Participants found this task natural and were unaware of any blend between the two image categories when debriefed. Before introducing a monetary component to the task, each participant completed 540 trials of simple face/house discrimination using the same stimulus timings as in the main experiment.

Face and house responses were made using left and right-hand key presses respectively. There were 15 stimulus levels, spaced in equal steps from 100% house to 100% face phase, enabling us to plot out each individual's psychometric function. The point of subjective equality (PSE) for each subject was then used to define face and house categories for the category-specific cost task.

The category-specific cost task involved further face/house discrimination under asymmetric losses for incorrect responses. There were three levels of the cost factor: face value (FV; -50p for an incorrect 'house' response, -10p for an incorrect 'face' response), neutral value (NV; -30p for an incorrect 'house' response, -30p for an incorrect 'face' response) and house value (HV; -10p for an incorrect 'house' response, -50p for an incorrect 'face' response). Before each block of trials, subjects were given an endowment of £10, and informed that they would keep any money they did not lose on the task. I used losses as losses are hypothesised to engender a greater behavioural impact on decision criteria than gains (Kahneman & Tversky, 1979). Cumulative feedback screens displaying the current total were provided only every 15 trials, to avoid incremental learning of decision strategy via trial-by-trial adjustments (Whiteley & Sahani, 2008).

Image phase (15 levels) was randomised and orthogonal to the cost factor, which was signalled to participants prior to the face/house stimulus on every trial (figure 4.2). The cost level changed every two trials. Subjects completed nine experimental blocks of 140 trials each, spanning a single session lasting around 3 hours including breaks. Note that when the penalty for answering 'house' incorrectly is greater, a reasonable strategy is to answer 'face' more often when uncertain of the answer.

## 4.3 fMRI experiment

The fMRI experiment took place within a week of the psychophysics, and employed the same task with minor alterations. Subjects completed four runs of 105 trials. The initial endowment for each block was £12, and feedback was given every 10 trials. Each trial began with a cost cue presented for 1820ms, followed by a variable interval of 100ms – 3000ms during which a fixation cross was presented. The face-house image was then presented for 100ms, and subjects were able to respond immediately following the onset of the image. Following the offset of the face-house image, a fixation cross was presented for a variable interval of 1625ms – 3625ms prior to the start of the next trial. The buttons indicating face and house responses were switched halfway through the session, so that each subject made face and house decisions with both left and right button presses. To avoid switch costs, a short training run was given with the new response mapping without any imaging data being collected.

Stimuli were presented in a permuted randomised fashion, so that the full phase range was covered every $\sim$ 7 trials. Similarly, the three cost levels were cycled every 6 trials (changing every two trials, as in the out-of-scanner psychophysics), while keeping stimulus phase and cost orthogonal. This cycling over $\sim$ 30s was chosen to match the filter properties of the canonical haemodynamic response function (HRF), maximising power for estimating the cost- and stimulus-related parameters in our event-related analysis.

### 4.3.1 fMRI acquisition

Images were acquired using a 3T Allegra scanner (Siemens, Erlangen, Germany). BOLD-sensitive functional images were acquired using a gradient-echo EPI sequence (48 transverse slices; TR, 3.12s; TE, 65ms; 3 × 3mm in-plane resolution; 2mm slice thickness; 1mm gap between adjacent slices; z-shim, +0.6 mT/m; positive phase encoding direction; slice tilt, -45 degrees) optimised for detecting changes in the parahippocampal region and fusiform gyrus (Weiskopf *et al.*, 2006). Four runs of 213 volumes were collected for each subject, followed by a T1-weighted anatomical scan and local field maps.

### 4.3.2 Behavioural data analysis

Subjects' psychophysical responses outside the scanner were modelled using a cumulative normal psychometric function incorporating a random lapse term (Wichmann & Hill, 2001), assuming binomial response counts (see Whiteley & Sahani 2008, for full details of the mathematical model). The curve for each cost condition (indexed by $j$) had three free parameters: the mean ($\mu_j$) reflecting the PSE, the slope ($\rho_j$) reflecting a subjects uncertainty over the whole stimulus range, and the lapse rate

($\epsilon_j$) reflecting motor errors and lapses of attention. In equation 4.1 below, $CP_{ij}$ gives the probability of answering 'face' for each given stimulus phase combination, $x_i$, in the $j$th cost condition:

$$\begin{aligned} CP_{ij} = \ & (1-\epsilon_j).\frac{1+f(\sqrt{\pi}.\rho_j.(x_i-u_j))}{2}+\frac{1}{2}\epsilon_j \\ & \text{where} \quad f(z)=\frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt \end{aligned} \tag{4.1}$$

I used gradient descent algorithms to find the parameter values that produced optimal curve fits to the observed data. I additionally implemented a Bayesian model comparison to determine whether sharing each of the parameters $\mu$, $\rho$ and $\epsilon$ between cost conditions gave better fits to the data than allowing each to be optimised separately.

To define decision difficulty for a given stimulus phase, I rectified each individual's psychometric function under the neutral cost condition around the PSE and normalised the result such that the range varied from 0 to 1 (equation 4.2).

$$U_i = \frac{||0.5 - CP_i| - 0.5}{0.5} \tag{4.2}$$

Following Grinband *et al.* (2006), this procedure defines the PSE as having maximal decision difficulty and 100 % house/face phase as having minimal decision difficulty. Note that this use of the term 'difficulty' refers to the difficulty of categorisation for a particular face-house phase composition, as opposed to the overall uncertainty about the task expressed in the slope of the psychometric function across the full phase axis. The latter might be expected to change with, for example, practice, stimulus duration, or lighting conditions.

The psychometric function fits to the in-scanner data were not a robust basis for inference, given the lower number of trials per data point compared to the psychophysics session. Consequently, I conducted further behavioural analysis using the framework of signal detection theory (SDT; Green & Swets 1966). Stimuli were classified as being faces or houses, depending on which side of the PSE they fell, yielding a classic 2 × 2 stimulus-response table for each cost condition. This approach implicitly approximates the stimulus continuum as being drawn from two overlapping Gaussian distributions, one for each category. This allowed us to compute subject-specific measures of sensitivity ($d'$) and criterion ($c$) separately for each cost condition. Despite being a cruder measure of behavioural performance than the psychometric function fitting described above, this method provides a useful index of whether value primarily affects sensory discrimination or decision/response criteria (Macmillan & Creelman, 2005), and circumvents the problem of fewer trials in the scanner leading to unreliable psychometric function fits.

### 4.3.3 fMRI data preprocessing and analysis

Functional data were analysed using SPM5 as outlined in Chapter 3. The first five volumes of each run were discarded to allow for T1 equilibration. Using the FieldMap toolbox (Andersson *et al.*, 2001), field maps were estimated from the phase difference between the images acquired at the short and long TE. The EPI images were then realigned and unwarped using the created field map, and slice-timing correction applied to align each voxel's timeseries to the acquisition time of the middle slice. Each subject's T1 image was segmented into grey matter, white matter and CSF, and the segmentation parameters were used to warp the T1 image to the SPM MNI template. The resulting normalisation parameters were then applied to the functional data. Finally, the normalised images were spatially smoothed using an isotropic 8mm full-width half-maximum Gaussian kernel.

fMRI timeseries were regressed onto a composite GLM containing delta (stick) functions representing the onsets of the cost cue, stimulus, response and cumulative feedback. These delta functions were convolved with the canonical HRF, and low-frequency drifts were excluded with a high-pass filter (128s cutoff). Short-term temporal autocorrelations were modeled using an AR(1) process. The stimulus delta functions were separated into three regressors dependent on the cost condition on each trial (face value – FV, neutral value – NV and house value – HV).

Each stimulus onset was parametrically modulated by two subject-specific functions. The first was the choice probability ($CP$) curve fitted to the out-of-scanner psychophysics data in the neutral value condition. The second was the decision difficulty function ($U$), again derived from the out-of-scanner psychophysics data, and orthogonalised with respect to choice probability (see equations 4.1 and 4.2 for mathematical definitions). The cumulative feedback stick function was also modulated with the amount of money lost on the previous 10 trials. To investigate interactions of value and response hand, the response delta function was separated by cost, decision and response hand, giving a 3 (cost; FV vs. NV vs. HV) × 2 (decision; face vs. house) × 2 (response; left vs right) factorial combination. Motion correction regressors estimated from the realignment procedure were entered as covariates of no interest.

### 4.3.4 Statistical inference

Statistical significance was assessed using linear compounds of the model parameters (regression coefficients of the trial-specific stimulus functions above), for each subject. These contrast images were then entered into a second-level random effects analysis using a one-sample t-test against zero to assess group-level significance. Cluster-based statistics (Friston *et al.*, 1994b) were used to define significant activations based both on their intensity and spatial extent. Clusters were defined using a height threshold of $P < 0.001$ and corrected for multiple comparisons across the whole brain using family-wise error correction (FWE) and a threshold of $P < 0.05$. Images are displayed at the cluster-defining threshold of $P < 0.001$ using MRIcron (`http://www.sph.sc.edu/comd/rorden/mricron/`). Small-volume correction (SVC) was applied to category-specific responses by using anatomical masks for fusiform and parahippocampal gyri as specified in the PickAtlas toolbox (Maldjian *et al.*, 2003). Percent signal change was extracted from clusters of interest for further analysis by averaging over subjects and sessions using MarsBar (Brett *et al.*, 2002). Estimated time courses within clusters are plotted at seven TRs following stimulus onset using a finite impulse response (FIR) model. I note that timecourses are plotted for illustration purposes only, inference having first been carried out using appropriate adjustments for multiple comparisons within SPM.

## 4.4 Results

### 4.4.1 Behavioural results

Subjects' average point of subjective equality (PSE) in the face-house discrimination pre-test was 53.9 ± 9.15 % face phase. Categorisation probability data from a representative subject's psychophysical results are shown in figure 4.2. To explore the effects of asymmetric cost on choice probability, I fit psychometric functions to the data with either shared or separate mean ($\mu_j$), slope ($\rho_j$), and lapse rate ($\epsilon_j$) parameters for the three cost conditions (indexed by $j$). I then carried out Bayesian model comparison, thereby revealing which of eight possible parameter structures (single vs. shared mean × single vs. shared slope × single vs. shared lapse rate) best accounted for the effects of manipulating asymmetric cost on choice probabilities.

All subjects consistently shifted their responses towards the category carrying lower cost, as expected (figure 4.3). Paired-sample t-tests confirmed that average shifts were significantly different from NV for both FV (t(18) = 5.95, $P < 0.0001$) and HV trials (t(18) = 4.98, $P < 0.001$). There were also small, but significant, differences in psychometric function slope between value conditions (white markers in figure 4.3b; one-way ANOVA, significant effect of value: $F_{(2,36)} = 4.61, P < 0.05$). Consistent with these results, figure 4.3a shows that the model with both separate means and slopes provided the best model of the data, despite an Occam's razor-like penalty for greater model complexity inherent in Bayesian model comparison. However, the magnitude of the difference between the summed log model evidences for shared and separate slopes is rather small, rendering definitive conclusions about differences in slope between conditions difficult.

As fewer trials precluded fitting reliable psychometric curves to the choice data in the scanner, I carried out a signal detection analysis (Green & Swets, 1966) to characterise in-scanner behaviour, collapsing stimuli into either face or house

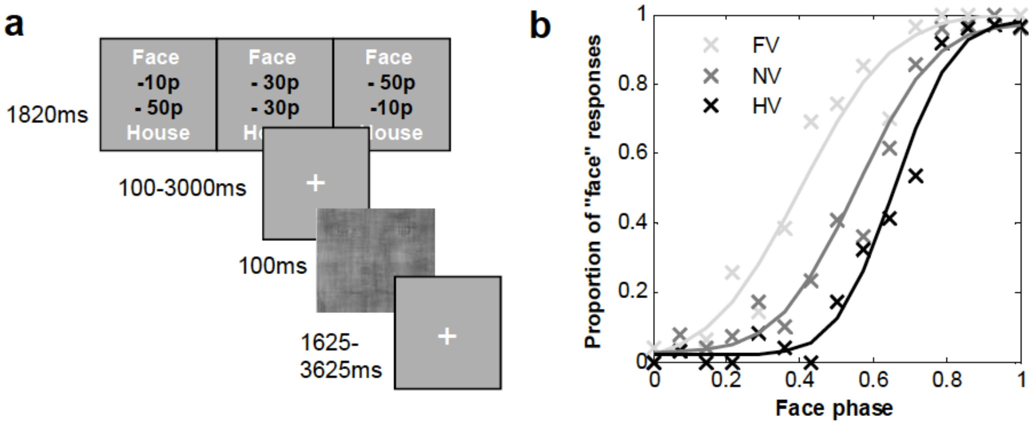

**Figure 4.2:** Perceptual decision task and example behavioural data. (a) Experimental procedure. Subjects viewed a cost signal screen informing them of the potential losses for an incorrect face or house categorisation at the start of each trial. They were then asked to categorise an image randomly drawn from the face-house phase continuum as a 'face' or 'house'. Timings shown are for the fMRI experiment. (b) Illustrative psychophysics data from one subject (LB). Crosses show choice probabilities for each stimulus phase and cost combination; lines show psychometric curves fit to the data.

categories based on each individual subject's PSE. This analysis confirmed that asymmetric cost led to deviations of the decision criterion in the predicted direction, relative to the neutral value condition ($c$; FV, t(14) = 5.82, $P < 0.0001$; HV, t(14) = 5.78, $P < 0.0001$) but did not change category discriminability ($d'$; $F_{(2,28)} = 0.41, P > 0.5$; figure 4.3c).

Importantly, mean reaction times (RT) did not differ across value conditions (psychophysics, $F_{(2,36)} = 0.70, P > 0.4$; in-scanner, $F_{(2,28)} = 1.67, P > 0.2$), suggesting that any bias-related differences I find in brain activity are not driven by systematic differences in task difficulty (figure 4.3). RT was however significantly correlated with decision difficulty (figure 4.4; psychophysics, mean $r = 0.79 \pm 0.092, N = 19$; in-scanner, mean $r = 0.56 \pm 0.21, n = 15$).

## 4.4.2 fMRI results

### 4.4.2.1 Cost-selective regions

I first identified regions involved in processing the extra demand of integrating asymmetric cost by computing the COST > NEUTRAL contrast (FV + HV > 2NV). A frontoparietal network (figure 4.5) was consistently active for both types of asymmetric cost condition compared to neutral ($P < 0.05$, whole-brain corrected), suggesting its involvement in the biasing of perceptual decisions as a function of category-specific cost. In addition to frontoparietal areas, I also found increased activity in a cluster spanning the subthalamic nucleus (STN) region, thalamus and caudate nucleus ($P < 0.05$, whole-brain corrected).

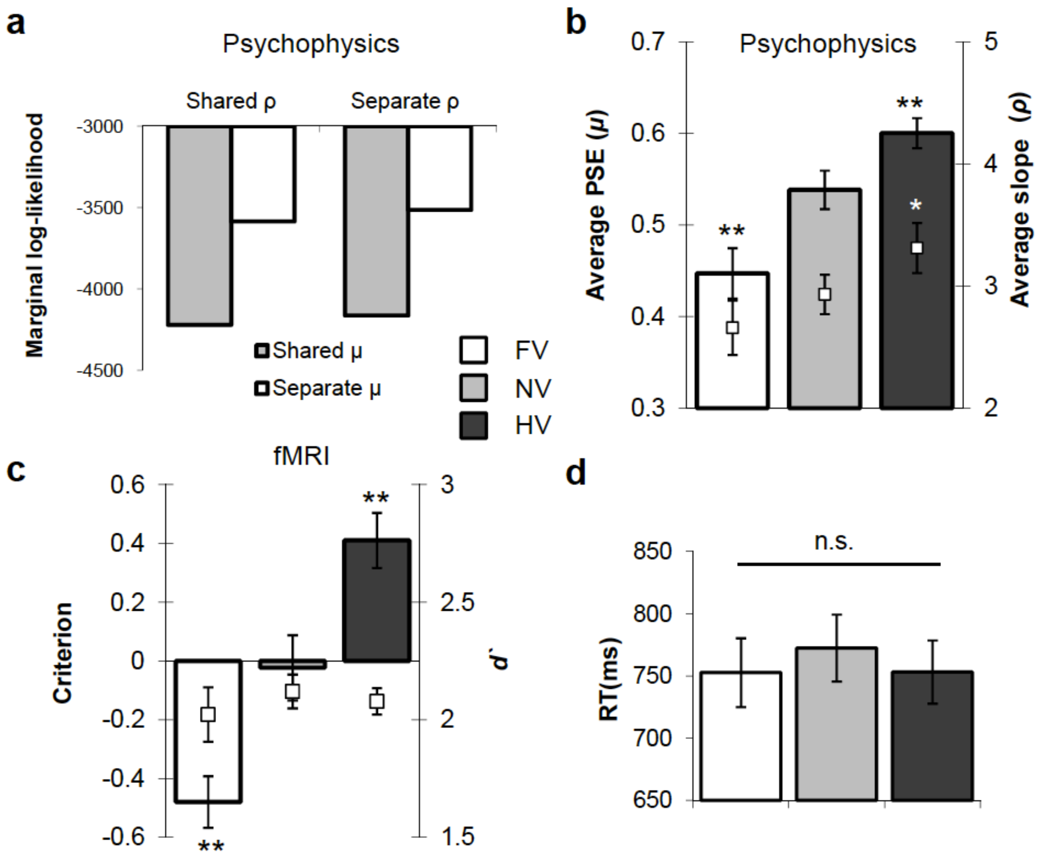

**Figure 4.3:** Behavioural results. (a) Bayesian model comparison was used to show that the best model for the psychophysics data was one with separate mean and slope parameters for each cost condition. The chart shows Laplace approximation to the total log marginal likelihood across subjects and across shared and separate error parameters (it seems possible that attentional lapses would vary with value condition, which does not bear on the hypothesis of interest) – a smaller negative value indicates a better model. Note that each unit difference in log likelihood corresponds to an $e$-fold ratio of model probabilities. (b, c) Average parameters of the psychometric function fits to the psychophysics data, N = 19 (b), and a corresponding signal detection analysis of the in-scanner data, N = 15 (c). Bars represent the point of subjective equality (PSE)/criterion in FV, NV and HV conditions. White markers indicate the average slope/$d'$ parameter in each value condition for comparison. (d) Mean reaction times (RT) averaged across changes in stimulus information for each cost condition. In all panels, error bars denote SEM; two asterisks (**), $P < 0.005$; one asterisk (*), $P < 0.05$ in comparison with NV.

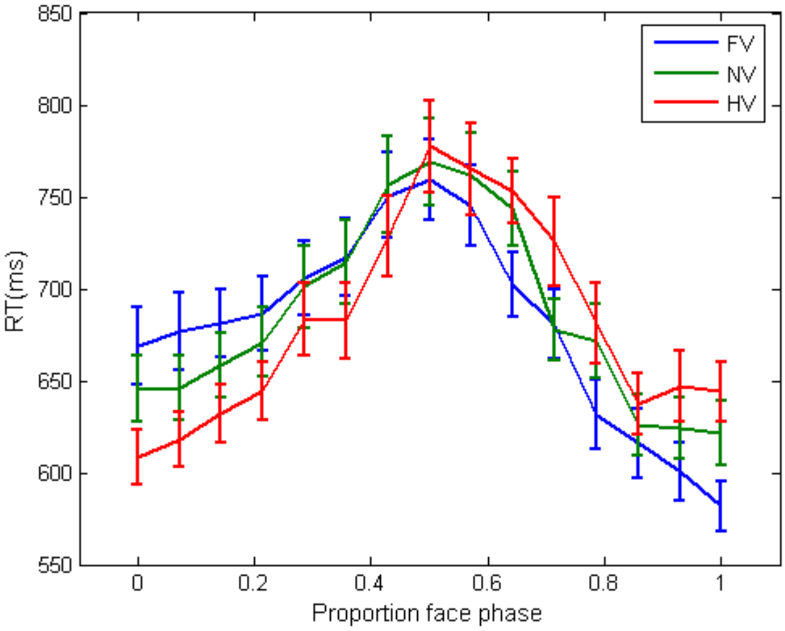

**Figure 4.4:** Mean reaction times from the psychophysics experiment as a function of cost condition and stimulus phase.

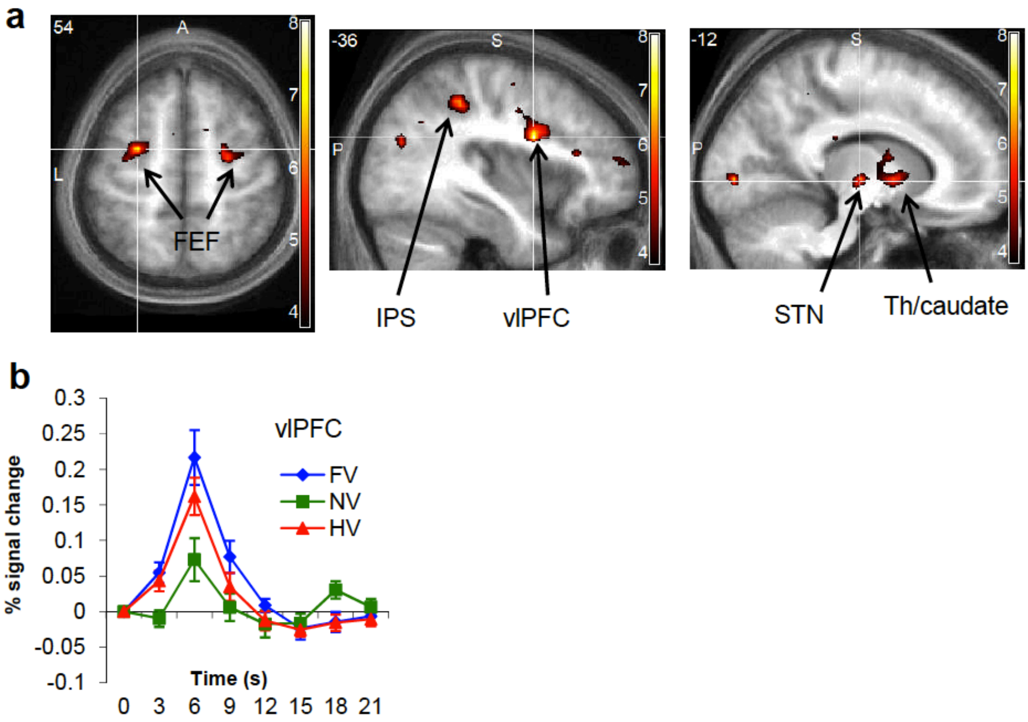

**Figure 4.5:** (a) Axial ($z$ = 54) and saggital ($x$ = -36 and -12) sections showing brain activations reflecting the main effect of asymmetric cost [(FV + HV) > 2NV], averaged over category. Shown are significant clusters in left ventrolateral prefrontal cortex (vlPFC), intraparietal sulcus (IPS), bilateral frontal eye fields (FEF) and subcortical regions (STN – subthalamic nucleus region; Th – thalamus; see also table 4.1). Labelled activations are significant at $P < 0.05$, cluster FWE whole-brain corrected. (b) Haemodynamic response time courses aligned to stimulus onset for the three different cost conditions, plotted for the significant cluster in vlPFC.

| Label | Voxels | $Z$-score | Cluster-FWE $P$-value | MNI coordinate | Laterality |
|---|---|---|---|---|---|
| IFG (p. opercularis) | 106 | 4.92 | $< 0.001$ | -36 3 24 | L |
| FEF | 36 | 4.56 | 0.004 | -27 3 54 | L |
| Caudate / thalamus / STN | 93 | 4.28 | $< 0.001$ | -15 -18 0 | L |
| FEF | 42 | 4.10 | 0.002 | 26 -9 54 | R |
| IPS | 49 | 4.26 | 0.001 | -36 -39 45 | L |
| IFG (p. triangularis) | 31 | 3.94 | 0.009 | -45 21 0 | L |
| Insula/Putamen | 33 | 3.90 | 0.007 | 36 15 -6 | R |
| post. MTG | 43 | 4.04 | 0.001 | -36 -72 21 | L |

**Table 4.1:** COST > NEUTRAL activations. Clusters were defined using a threshold of $P < 0.001$, uncorrected.

#### 4.4.2.2 Stimulus-selective regions

To test our first hypothesis, that category-specific costs directly affect responses in the ventral visual stream, I computed the signal change for each cost condition in each stimulus-selective ROI identified above. One-way ANOVAs (FV vs. NV vs. HV) revealed effects of cost on right PHG ($F_{(2,28)} = 4.80, P = 0.002$), but not left PHG or right FG ($F_{(2,28)} < 2.5, P > 0.1$). Further investigation of the pattern of differences in right PHG revealed increases in the HV condition compared to NV (t(14) = 3.09, $P = 0.008$), a trend for increases in the FV condition (t(14) = 1.77, $P = 0.10$), but no evidence for differences between the category-specific FV and HV conditions (t(14) = 1.26, $P = 0.23$). A similar trend for non-specific increases under asymmetric cost conditions can be seen in all three ventral visual areas (figure 4.6c and d), and an omnibus ANOVA in which region was included as a separate factor indicated an overall significant effect of cost ($F_{(2,28)} = 5.95, P = 0.007$). Together, these analyses indicate that asymmetric cost has a general driving effect on ventral stimulus-selective regions, but in a manner that does not appear to discriminate between stimulus categories.

I also considered the possibility that any biasing effects of cost in extrastriate visual areas may impact on the face- or house-selectivity of these regions, rather than a tonic level of activity. To test this I computed beta estimates for the choice probability-modulated regressor at peak object-selective voxels for each different cost condition. As expected, the activity of these regions showed significant correlations with either face or house stimulus information in each cost condition (seen as consistently above-zero parameter estimates in figure 4.7); however, one-way cost ANOVAs showed no significant influence of value on this selectivity (all $F_{(2,28)} < 0.95, P > 0.40$).

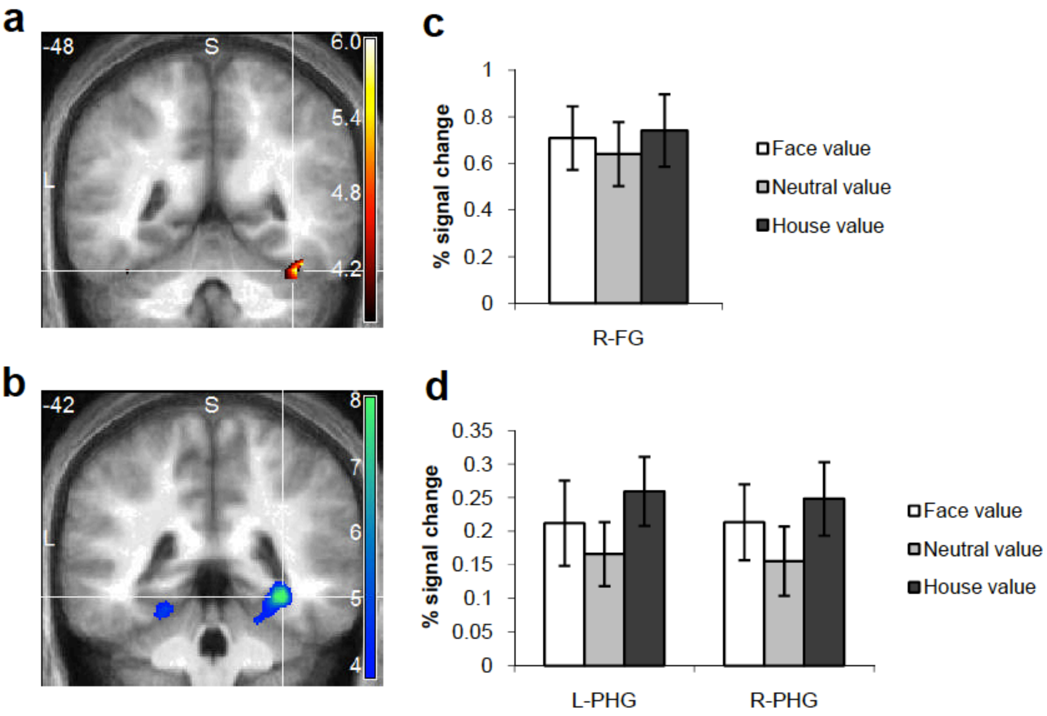

**Figure 4.6:** Effects of cagetory-specific costs on stimulus-selective brain activity. (a) Coronal (y = -48) section showing parametric effects of the probability an image was classified as a face in right fusiform gyrus (FG; 39, -48, -24; $Z$-score = 4.07; $P < 0.05$, small-volume corrected). (b) Coronal (y = -42) section showing parametric effects of the probability an image was classified as a house in bilateral parahippocampal gyrus (PHG; left: -21, -42, -15; $Z$-score = 3.68; right: 33, -42, -9; $Z$-score = 5.26; both $P < 0.05$, small-volume corrected). (c, d) Percent signal change as a function of value condition in stimulus-selective ROIs defined from clusters shown in (a) and (b).

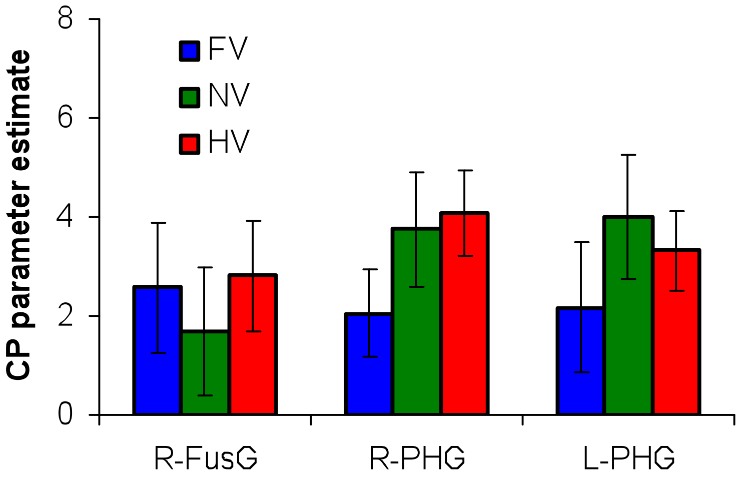

**Figure 4.7:** Parameter estimates for the correlation with face or house choice probability ($CP$) in neutral and category-specific cost trials, plotted at the peak voxel of each object-selective ROI. Error bars denote s.e.m.

#### 4.4.2.3 Category-specific effects of cost

The account presented thus far indicates that category-selective stimulus information is to some degree represented independently of category-specific biases induced by changes in the payoff matrix. In other words, payoff asymmetries lead to only general, not category-specific, increases in the signal in voxels sensitive to stimulus (face or house) information. However, the mechanism by which asymmetric value information brings about a change in perceptual decision (such as a bias towards responding face) remains unclear. To address this question, I computed the category-specific value contrasts (FV > HV; HV > FV). A cluster in left parahippocampal gyrus (-33, -36, -15; $P < 0.05$, SVC) responded specifically to increases in house value (decrease in cost for responding house), lateral and anterior to the stimulus-selective cluster I characterised previously (figure 4.8a). No significant clusters were evident in the opposite (FV > HV) contrast even at a liberal ($P < 0.01$) defining threshold. Together, these findings suggest that category-specific costs exert effects on the ventral visual pathway (at least for the bias towards responding house).

I noted that while the direction of decision criterion shifts under asymmetric cost was consistent across subjects (figure 4.3), individual differences in the size of this shift were evident in the behavioural data (figure 4.8b). To explore whether subjects who exhibit greater decision criterion shifts also show greater activity within regions that are the putative sources or targets of these shifts, I regressed the category-specific value contrasts (FV > HV; HV > FV) against between-subjects covariates encoding the amount of behavioural bias in the relevant asymmmetric value condition (FV criterion shift; HV criterion shift). Subjects who displayed greater criterion shifts in the HV condition tended to activate the anterior cingulate cortex (ACC; 6, 36, 30) more than subjects who shifted to a lesser degree (figure 4.8c, d; $P < 0.05$, whole-brain corrected). Again, as for the simple main effect of FV > HV, no significant correlations were found with individual differences in the FV criterion ($P > 0.001$, uncorrected).

#### 4.4.2.4 Decision difficulty

To look for brain regions responsive to decision difficulty, I regressed a parameter which essentially measures how close to chance the subject is in deciding whether the stimulus is a face or a house (see equation 4.2) onto the fMRI signal at the time of choice. Dorsal medial frontal (paracingulate) cortex (dMFC; 6, 12, 51) and right anterior insula (42, 24, -3) showed positive correlations with decision difficulty (both $P < 0.05$, whole-brain corrected; figure 4.9). I next established whether this difficulty-related BOLD signal was independent or overlapping with the frontoparietal regions found to be active under conditions of asymmetric cost. By exclusively masking the COST > NEUTRAL contrast for regions correlating with decision difficulty at a liberal ($P < 0.05$, uncorrected) threshold, I found that

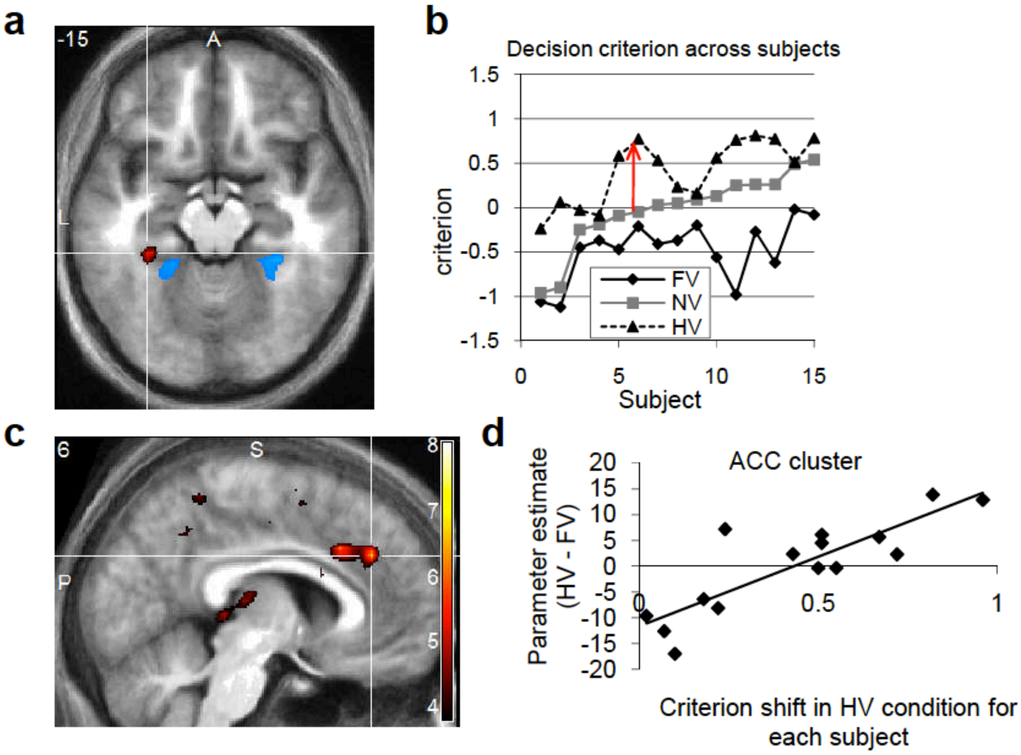

**Figure 4.8:** (a) Axial (z = -15) showing the region in left parahippocampal gyrus (red) active during decreased cost (increased value) for houses (HV > FV; -33, -36, -15; $Z$-score = 3.84; $P < 0.05$, small-volume corrected). Shown in blue are clusters selective for house stimulus information (figure 4.6b) for comparison. (b) Intersubject variation in decision criteria, with subjects ordered by their decision criterion in the NV condition. The arrow shows the difference (extent of behavioural shift under HV) used as a covariate for the relevant contrast testing for HV-specific effects of asymmetric value shown in (a). (c) Saggital (y = 6) and axial (z = 30) sections showing a region in the anterior cingulate (ACC) that shows greater activity in subjects who show greater behavioural shifts in the HV condition (6, 36, 30; $Z$-score = 4.29; $P < 0.05$, whole-brain corrected). (d) Averaged HV > FV beta within the ACC cluster shown in (c) plotted against the criterion shift in the HV condition, across subjects. Inference was carried out using appropriate corrections for multiple comparisons in the SPM framework; this plot is simply provided for illustration purposes.

left vlPFC, left caudate/thalamus/STN and bilateral FEF were specifically active under conditions of asymmetric cost, independent of changes in decision difficulty (table 4.2). Conversely, dMFC activity was independent of changes in category value, indicating a partial dissociation in the brain between regions encoding changes in decision difficulty and prospective costs during perceptual categorisation. The negative effect of the decision difficulty regressor (testing for greater activity in 'easy', certain decisions) was seen in ventromedial prefrontal cortex (vmPFC; 9, 33, -6; $P < 0.05$, whole-brain corrected; figure 4.9).

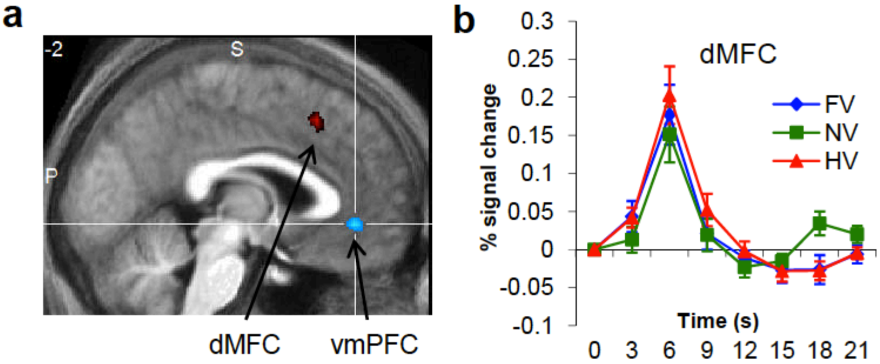

**Figure 4.9:** (a) Saggital section (y = -2) showing positive (red) and negative (blue) correlations with a regressor indexing decision difficulty (see equation 4.2). Dorsomedial frontal cortex (dMFC; 6, 12, 51; $Z$-score = 3.70; $P < 0.05$, whole-brain corrected) and insula (not shown; 42, 24, -3; $Z$-score = 4.11; $P < 0.05$, whole-brain corrected) showed positive correlations with difficulty. A cluster in ventromedial prefrontal cortex (vmPFC) showed increased activity for easier decisions (6, 36, 30; $Z$-score = 4.20; $P < 0.05$, whole-brain corrected). (b) Haemodynamic response timecourses for the three different cost conditions, plotted for the significant cluster in dMFC. While showing strong correlations with the categorical difficulty regressor, this region was insensitive to changes in category value (cf. figure 4.5).

#### 4.4.2.5 Interaction of cost with motor planning

The previous analyses identified brain regions that responded preferentially to a particular direction of bias (towards responding 'house'). I next asked whether any bias effects are expressed at the level of the motor system, given that response hand (left or right) was orthogonal to decision (face or house). Interactions of cost asymmetry with response hand were computed by coding each trial as to whether the left or right hand was assigned to a high or low cost response (face or house), and examining the interaction of cost condition with response hand. No effects were found ($P > 0.001$, uncorrected), suggesting that the biasing effects of asymmetric value occur upstream of effector-specific response planning.

| Contrast | Voxels | $Z$-score | Cluster-FWE $P$-value | MNI co-ordinate | Laterality | Label |
|---|---|---|---|---|---|---|
| [(FV+HV) > NV] ex. masked by $U$ | 35 | 4.56 | 0.013 | -27 -3 54 | L | FEF |
| | 93 | 4.28 | $< 0.001$ | 15 -18 0 | L | Caudate / thalamus / STN |
| | 42 | 4.10 | $< 0.001$ | 27 -9 54 | R | FEF |
| | 43 | 4.04 | $< 0.001$ | -36 -72 21 | L | post. MTG |
| | 29 | 3.90 | 0.035 | 36 15 -6 | R | Insula |
| $U$ ex.masked by [(FV + HV) > NV] | 35 | 3.54 | 0.027 | 9 12 48 | R/L | dMFC |

**Table 4.2:** Activations following exclusive masking for either cost- or difficulty-related activity.

## 4.5 Discussion

Here I investigated the brain mechanisms that integrate prospective costs and sensory evidence during decisions under uncertainty. The behavioural manipulation systematically biased the perception of a noisy image using asymmetric costs, leading to shifts in decision criteria. These shifts functioned to reduce monetary losses, by biasing decisions toward the category with lower cost when the participant was unsure of the answer. Using fMRI, I then asked whether category-specific shifts were reflected by changes in frontoparietal areas known to accumulate evidence leading to perceptual categorisation (e.g. Heekeren *et al.* 2004), in ventral visual regions known to encode category-specific information about faces and houses, and/or through changes in effector-specific mechanisms. Our data best fit the former, 'decision stage' hypothesis. The requirement to integrate asymmetric cost information into perceptual decisions robustly activated a frontoparietal network, despite conditions being closely matched for expected value and reaction time. In addition, a cluster in the thalamus/caudate was active under asymmetric cost, consistent with subcortical loops being important for the setting of decision criteria (Lo & Wang, 2006; Simen *et al.*, 2006). A specific effect on ventral visual areas (parahippocampal gyrus) was found under decreasing cost for houses, although this effect was anatomically separate to that shown to be sensitive to bottom-up stimulus information. Finally, subjects who showed greater shifts in decision criteria towards houses demonstrated greater activation of the anterior cingulate cortex (ACC), a region thought to be pivotal in the adjustment of decision strategy (Behrens *et al.*, 2007; Botvinick *et al.*, 2001).

### 4.5.1 Sources of category-specific bias

The dorsal frontoparietal network active under asymmetric cost is similar to that commonly activated in studies of transient allocation of attention (Corbetta *et al.*, 2008; Corbetta & Shulman, 2002; Yantis *et al.*, 2002), and has been implicated in the modulation of early visual cortical activity by rewards tied to particular locations in visual space (Serences, 2008). It is plausible that changes in category-specific costs co-opt a similar network. Low-level changes in arousal or task difficulty are unlikely to be explanations for this widespread increased activity, as RTs and potential gains/losses were matched across conditions. Instead, our findings indicate that frontoparietal (ventrolateral prefrontal cortex, insula, intraparietal sulcus, frontal eye fields) and subcortical (anterior thalamus/STN) regions are recruited when an asymmetric loss function affects the perceptual decision. Bilateral activation was found at the junction of the precentral and superior frontal sulci, consistent with the location of the frontal eye fields (FEF; Lobel *et al.* 2001), which have been shown to encode shifts in decision criteria (Ferrera *et al.*, 2009). In addition, activation in ventrolateral prefrontal regions including insular cortex is consistent with involvement both in the accumulation of sensory evidence (Romo *et al.*, 2004; Ho *et al.*, 2009) and the incorporation of gains and losses in decision-making (Leon & Shadlen, 1999; Watanabe & Sakagami, 2007). Indeed, the network outlined above may be recruited more generally when shifts in decision criteria are induced by manipulations other than asymmetric payoffs. This view is supported by previous findings of modulation of subcomponents of this network when decision criteria are shifted through changes in category boundary – specifically, BOLD signal in anterior thalamus/caudate and insula/vlPFC (Grinband *et al.*, 2006; Li *et al.*, 2009) and single-unit activity in FEF (Ferrera *et al.*, 2009).

The present analyses cannot pin down the source of the bias towards houses and faces, as effects of category-specific bias were not observed in the network discussed above. However, it is possible that local neural subpopulations within these areas encode biases towards face and house categories. This suggestion is supported by a recent study by Rorie *et al.* (2010) in monkeys, demonstrating that asymmetric payoffs in a perceptual decision task bias the initial firing rate of individual neurons in the intraparietal sulcus coding for a saccadic response to one of two particular targets. Similarly, using fMRI, distributed patterns in the inferior frontal gyrus/insula have been found to discriminate the direction of criterion shifts in a visual categorisation task (Li *et al.*, 2009). Our finding that the ACC tracks individual differences in the extent of a signed criterion shift is also consistent with category-specific payoff information being represented in frontal cortex.

### 4.5.2 Differential effects of house and face cost

When the signal change in stimulus-selective ROIs was calculated, general but not selective effects of asymmetric cost were observed. In contrast, a parahippocampal region anterior to the stimulus-selective areas was significantly more active when houses rather than faces were more valuable. Due to the potential impact of smoothing and thresholding procedures, I am cautious in attributing separate locations to the stimulus- and cost-sensitive regions of the PHG. However, I note evidence suggesting local separation of task- and stimulus-driven regions using neural stimulation coupled with high resolution imaging (Ekstrom *et al.*, 2008); similarly, differences between stimulus- and task-driven localisations have been reported in the fusiform gyrus during perceptual decision making (Philiastides & Sajda, 2007).

Category-specific biases were seen in ventral visual regions for increases in house, but not face, value. Furthermore, across subjects, a correlation with decision criteria was seen in the ACC for bias towards houses, but not faces. While I am cautious about over-interpreting null results, I note that previous studies examining attentional and decisional biases towards faces and houses have also found asymmetric effects of the two categories. Specifically, Summerfield *et al.* (2006a) found that mistaken categorisations of houses as faces were accompanied by increases in fusiform gyrus activity, but that the opposite mistake did not increase PHG responses. In contrast, Serences *et al.* (2004) found that shifts in object-based attention towards houses recruited parietal and frontal regions to a greater degree than shifts towards faces. Both these findings and the asymmetry in the results of the present study can be reconciled by assuming that subjects have a dominant prior to respond ‘face’. This hypothesis is supported by informal debriefing – some subjects in our study commented that they performed the task by responding house whenever evidence for a face was scant (see also Summerfield *et al.* 2006b). In the case of the present data, increases in value for houses would lead to top-down shifts in PHG activity to overcome this implicit prior towards responding face, but the converse may not be necessary. Whether visual phenomenology also changes under such shifts in decision criteria is an open question, one that could potentially be addressed by eliciting detailed reports from subjects under biased and unbiased conditions (Jack & Roepstorff, 2002).

Finally, one limitation of our results is that I was unable to specify subject-specific face- and house-sensitive ROIs. Parametric correlations with stimulus phase were not robust at a single-subject level, precluding the identification of individual ROIs in the ventral visual cortex and potentially obscuring inter-individual variability in regions of peak stimulus sensitivity (e.g. Spiridon *et al.* 2006). However, I note that if this analysis was underpowered to detect inter-individual differences in peak locations of FFA/PPA, this limitation applies equally to the category-specific contrasts (FV > HV; HV > FV). The lack of robust ventral visual activity in these contrasts, despite detection of parametric correlations with face/house information, supports a view that asymmetric costs modulate decision-making downstream of extrastriate visual regions.

### 4.5.3 Responses to decision difficulty

Consistent with previous reports, I found that activity in dorsal medial frontal (paracingulate) cortex (dMFC) and anterior insula correlates with decision difficulty (Grinband *et al.*, 2006; Preuschoff *et al.*, 2008; Philiastides & Sajda, 2007). There has been recent debate about the functional role of the medial frontal/paracingulate cortex in perceptual decisions (Heekeren *et al.*, 2008). Here I report preliminary evidence for segregation of networks responding to changes in decision difficulty and category value. Dorsal paracingulate activity correlated with increases in decision difficulty, independent of changes in value; conversely, the frontal eye fields and caudate/thalamus/STN were specifically active during decisions requiring integration of loss function information. As decision difficulty was correlated with reaction time (RT), I was unable to dissociate the contributions of decision time to these activations (cf. Grinband *et al.* 2006, 2008). Interestingly however, the dMFC region lies just dorsal to the ACC, which responded to the degree of decision criterion shift across individuals. Given that such shifts are only required when subjects are uncertain about the sensory data (Maddox & Bohil, 2003), the close anatomical relationship between these regions may be optimal for integration of decision uncertainty during shifts in decision criteria.

Conversely, the opposite contrast (examining brain activity that increases for 'easy', certain choices), revealed a cluster in ventromedial prefrontal cortex. This finding supports recent evidence that the ventromedial prefrontal cortex may signal a perceptual 'match' between observed and predicted stimulus information (Summerfield & Koechlin, 2008), and accords with suggestions that perceptual accuracy itself may act as a reinforcer (Bohil & Maddox, 2001). However, in the present experiment this signal could also be related to the ongoing assessment of the expected value of the current decision (Boorman *et al.*, 2009; Gottfried *et al.*, 2003), as both perceptual accuracy and potential rewards are highly correlated on any given trial.

### 4.5.4 Summary

To conclude, the findings in this chapter extend previous reports that costs attributed to perceptual decision outcomes have consistent effects on stimulus categorisation, with subjects acting to minimise prospective losses. I show that this effect of cost on perceptual decisions is robustly associated with BOLD signal increases in a frontoparietal network, in keeping with the hypothesis that loss functions affect a decision stage of processing. When the cost for responding 'house' decreased, I additionally observed selective activation within the parahippocampal gyrus. Across subjects, greater shifts in decision criteria were associated with greater activation of the medial frontal cortex (ACC). These findings are consistent with the hypothesis that asymmetric costs alter an intermediate representation between perception and action, albeit with possible recurrent effects upon extrastriate cortex.

# Chapter 5

# Effects of an inaction default on perceptual decision criteria

To do nothing is in every man's power.

Samuel Johnson (1709-1784)

## 5.1 Introduction

In Chapter 4 we saw that activity in inferior frontal and parietal cortex, and subcortical regions including thalamus, caudate and STN, was increased when an asymmetric loss function is incorporated into a perceptual decision. This result suggests that the loss function adjusts a decision stage, as opposed to early coding of category beliefs (evidenced by weak modulation of extrastriate visual activity). However, I found no modulation of category-specific biases in the active fronto-basal ganglia network. One reasonable explanation for this null result is that the experiment in Chapter 4 was optimised to investigate the effects of asymmetric bias on *category* coding, rather than on *response* coding. In other words, because the response in both cases was a single button press, we might expect category-specific biases in a cortico-basal ganglia network to be irresolvable with the available spatial resolution of fMRI.

In the present study I take an alternative tack to address this question. I set up a default response option, requiring no action on the part of the subject. On each trial, this manipulation associates half the decision axis with inaction (figure 5.1). This allows investigation of the neural signature for crossing a threshold for action. Furthermore, I expect the default to itself act as an asymmetric prior: under equivocal sensory input, the default should be favoured. I can thus set up a 2 × 2 design crossing decision difficulty with whether the default is rejected or accepted. As rejecting the default requires action, I expected a markedly different neural signature in cortico-basal ganglia loops for rejecting compared to accepting the default. I asked participants to make sensory judgements in the context of a tennis 'line-judgement' game (figure 5.2) while undergoing functional magnetic resonance imaging (fMRI). This game was selected on the basis of its natural default option – line judges remain silent to indicate that the ball was in, but make an overt response by shouting 'out' to reject the default. Further, such a task involves graded perceptual difficulty (Mather, 2008). To examine brain mechanisms for overcoming this bias, I therefore implemented a simple factorial design by crossing high and low decision difficulty with rejection or acceptance of the default.

The neural mechanisms involved in adjusting thresholds for action in the face of difficulty are unknown. Decision difficulty activates the dMFC (Botvinick *et al.* 2001; Heekeren *et al.* 2008; see also Chapter 4); and the dMFC is strongly interconnected with the lateral frontal cortex (Koski & Paus, 2000), which may implement behavioural adjustments in control (Buch *et al.*, 2010; Kouneiher *et al.*, 2009). In addition, conflict is thought to in part recruit STN (Frank, 2006), which is known to play a key role in response inhibition (Aron & Poldrack, 2006; Aron *et al.*, 2007; Li *et al.*, 2008), and is strongly interconnected with medial and lateral frontal cortex via an anatomical hyperdirect pathway (Aron *et al.*, 2007; Nambu *et al.*, 2000). Furthermore, despite the beneficial effects of deep-brain stimulation (DBS) of the STN in Parkinsons disease (Bergman *et al.*, 1990; Limousin & Martinez-Torres, 2008; Gradinaru *et al.*, 2009), DBS can lead to impairments in cognitive control (Alberts *et al.*, 2008; Ballanger *et al.*, 2009; Hershey *et al.*, 2004; Frank *et al.*, 2007), suggesting a core function of the STN is to modulate cortico-basal ganglia circuits involved in decision making (Frank, 2006; Gurney *et al.*, 2001).

An alternative perspective on this prediction can be drawn from behavioural economics. When faced with a complex decision people tend to accept the status quo, as reflected in the old adage 'When in doubt, do nothing'. Indeed, across a range of everyday decisions, such as whether to move house or trade in your car, or even whether to flip the TV channel, there is a considerable tendency to maintain the status quo, and refrain from acting (Samuelson & Zeckhauser, 1988). One factor driving this status quo bias is the difficulty of the decision process. In supermarkets, for example, there is often an overwhelming choice of different brands for the same product, and consumers may leave the store empty handed because of a difficulty-induced bias towards inaction (Anderson, 2003; Iyengar & Lepper, 2000).

As the previous paragraph indicates, the default bias can be shaped by a number of complex and interacting factors including economic costs involved in making the transition (Johnson & Goldstein, 2003; Samuelson & Zeckhauser, 1988), aversion to losing what you presently own (DeMartino *et al.*, 2009; Kahneman & Tversky, 1979), and the potential for regretting a change (Anderson, 2003). Here I restrict investigation to the ubiquitous factor of decision difficulty, minimising the influence of other, potentially confounding psychological variables. In my simple visual detection task, the choice set size remains constant (two-alternative forced-choice) and outcomes are omitted. This simple factorial design allows investigation of the effect of impact of an asymmetric loss function – the default – independently of the difficulty of the decision.

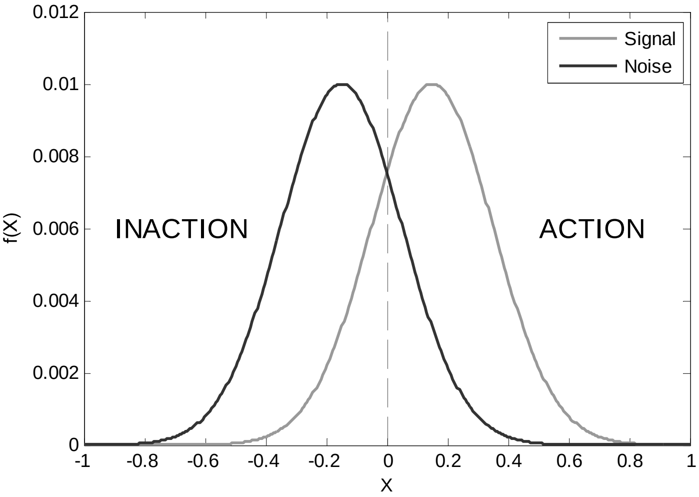

**Figure 5.1:** Signal detection theory schematic of an action-asymmetric decision. On each trial, either the area to the left or right of the criterion is mapped to inaction (here, it is the noise response).

## 5.2 Methods

### 5.2.1 Subjects

Seventeen healthy right-handed subjects who provided informed consent took part in the study. All had normal or corrected-to-normal vision, and no history of psychological or neurological illness. One participant was excluded due to poor behavioural performance (33 % errors on low difficulty trials). The study was approved by the Institute of Neurology (University College London) Research Ethics Committee. Participants received a fixed reimbursement plus a small bonus payment calculated from their best-scoring block of trials.

### 5.2.2 Task and procedure

Stimuli were presented on a grey background using Cogent 2000 (`www.vislab.ucl.ac.uk/cogent.php`) running in MATLAB. The court consisted of two white tramlines presented either side of fixation, the outer edge of which was viewed at an eccentricity of 12.4 degrees of visual angle (figure 5.2). The ball was a filled yellow circle subtending 3.7 degrees. Stimuli were presented using an NEC LT157 LCD projector running at a refresh rate of 60 Hz, viewed by subjects via an adjustable mirror.

Each trial began with a central fixation cross flanked by two longitudinal white tram lines, presented for a variable interval (750-3000ms) in peripheral vision. Participants were asked to maintain fixation, and instructed that not doing so would compromise their performance on the line judgment task. The target ball was presented at either tramline for 66ms, either overlapping the line (in) or outside the line (out). The difficulty of the decision was manipulated by altering the distance of the stimulus from the outside edge of the tramline. An inter-stimulus interval of 750ms followed the offset of the target.

Responses were made using an optical keypad and consisted of a go/nogo decision. Specifically, during each run of trials, participants were required to depress one key with the index finger of their right hand, designated as the default. Response options (in/out) were presented for 2000ms. One of these options was defined as the default by a surrounding black square. Participants continued to depress the default key to select the default option (accept); this key was released and a second key pressed to select the alternative (reject). On half of trials the target offset was defined as low difficulty and on the other half high difficulty by drawing the offsets from two separate Gaussian distributions defined on the basis of pilot data. The random draw of offsets was further constrained to produce half 'out' and half 'in' ball positions. The default option was balanced over in/out and over low/high difficulty trials, giving a fully factorial design.

Each participant was given both written and verbal task instructions, before being familiarised with the task format by a short practice block (16 trials). The task involved 3 runs of 80 trials, with a short break between runs. Participants were informed that they would earn 20p per correct decision and lose 10p for every incorrect decision. This asymmetry in wins and losses was designed to ameliorate the effects of loss aversion on the status quo bias, given previous findings of losses looming around twice as large as gains (Tversky & Kahneman, 1991). Feedback, in terms of cumulative money earned and lost, was given every 10 trials; trial-by-trial feedback was not given. At the end of the task participants received a bonus payment equivalent to their earnings in their highest scoring run.

### 5.2.3 Behavioural analysis

Behavioural responses were classified according to whether the trial led to a rejection or acceptance of the default, and whether the trial was high or low difficulty. An inaction bias was assessed by comparing the proportion of trials leading to an acceptance response on high and low difficulty trials, using a two-tailed paired t-test. Each participant's decision criteria ($c$) and sensitivity ($d'$) were estimated from the data using signal detection theory (SDT; see Chapter 3), where the hit rate

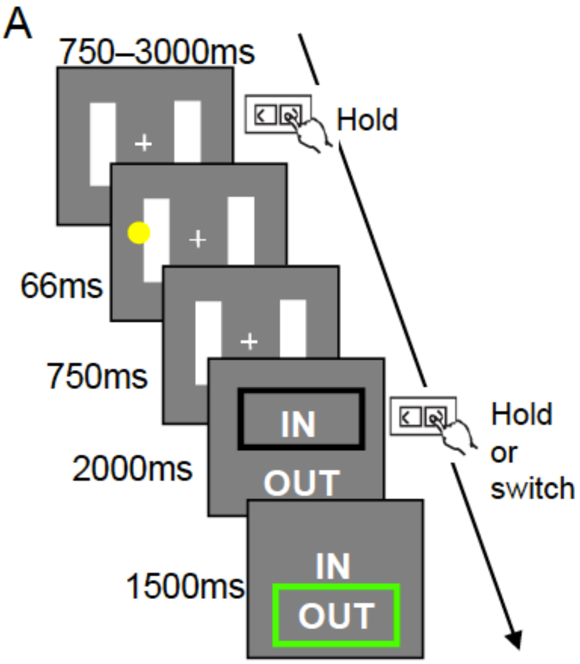

**Figure 5.2:** Participants played a 'tennis line-judgment' game in which the default was systematically manipulated in a balanced factorial design. At the beginning of each trial, participants were asked to depress the 'default' key and fixate on the cross between two tramlines. They then saw a ball land on the court, before being asked to make a decision on whether it was in (overlapping the line) or out. This decision was indicated by continuing to depress the key to accept the default, or releasing it and switching to the opposite key to reject. Uncertain and certain trials were randomly interleaved within a block, and balanced across whether the correct response was to accept or reject the default.

($H$) was defined as $p(\text{`in'}|\text{ball} = \text{in})$ and false alarm rate ($F$) as $p(\text{`in'}|\text{ball} = \text{out})$. SDT parameters and error rates were analysed using repeated-measures analysis of variance (ANOVA).

### 5.2.4 fMRI analysis

Functional data were analysed using SPM5 in the manner reported in Chapter 3. The first five volumes of each run were discarded to allow for T1 equilibration. EPI images were realigned and unwarped using field maps (Andersson *et al.*, 2001), and slice-timing correction applied. Each subject's T1 image was segmented into grey matter, white matter and CSF, and the segmentation parameters were used to warp the T1 image to the SPM MNI template. These normalisation parameters were then applied to the functional data. For one subject, normalisation parameters were estimated from the SPM EPI template due to the unavailability of a T1 image. Finally, the normalised images were spatially smoothed using an isotropic 8mm full-width half-maximum Gaussian kernel.

fMRI timeseries were regressed onto a composite GLM containing delta (stick) functions representing the onsets of the lines, ball, choice screen, button press (if any), response screen and cumulative feedback. These delta functions were convolved with the canonical HRF, and low-frequency drifts were excluded with a high-pass filter (128s cutoff). Short-term temporal autocorrelations were modeled using an AR(1) process. Stimulus delta functions were separated into two regressors depending on the perceptual uncertainty on each trial (high/low). Choice screen delta functions were separated into six regressors dependent on whether the trial was high/low difficulty, whether it led to an accept/reject response, and, on high difficulty trials, whether this response was correct or incorrect (reject_high_correct, reject_high_incorrect, reject_low, accept_high_correct, accept_high_incorrect, accept_low). Response accuracy (correct/incorrect) was not modeled as a separate factor on low difficulty trials given the relative rarity of incorrect responses ($4.9 \pm 1.0\%$, s.e.m.). Additionally, the reject stick functions were parametrically modulated by the reaction time on each trial, and the cumulative feedback stick function was modulated by the amount of money won on the previous 10 trials. Motion correction regressors estimated from the realignment procedure were entered as covariates of no interest.

Statistical significance was assessed using linear compounds of the regressors in the GLM, generating statistical parametric maps (SPM) of $T$-values across the brain for each subject and contrast of interest. These contrast images were then entered into a second-level random effects analysis using a one-sample t-test against zero. Our critical contrast of interest (the interaction of default rejection and difficulty, collapsing across correct/incorrect) was computed as follows: [reject_high_correct = +0.5; reject_high_incorrect = +0.5; reject_low = −1; accept_high_correct = −0.5; accept_high_incorrect = −0.5; accept_low = +1].

Cluster-based statistics were used to define significant activations both on their intensity and spatial extent (Friston *et al.*, 1994b). Clusters were defined using a threshold of $P < 0.005$ and corrected for multiple comparisons within a given search volume using family-wise error correction (FWE) and a threshold of $P < 0.05$. Small volume correction was applied to a priori regions of interest (ROIs) in the STN and dmPFC. Right and left STN ROIs were defined as 10 x 10 x 10 mm boxes centred on $\pm 10, -15, -5$, following Aron & Poldrack (2006); the dmPFC ROI was defined as a 12mm sphere centred on 0, 27, 30. This volume is representative of coordinates reported in a recent meta-analysis of conflict-related activity in the dmPFC (Ridderinkhof *et al.*, 2004). In order to quantify the interaction effect, percent signal change within each STN ROI was extracted for each condition and averaged across subjects and sessions using MarsBar (Brett *et al.* 2002; `http://marsbar.sourceforge.net/`).

## 5.3 Results

### 5.3.1 Behaviour

In line with theoretical predictions, there was a greater tendency to accept the default on high compared to low difficulty trials ($t(15) = 2.51, P < 0.05$; figure 5.3). Post-hoc paired t-tests confirmed that this interaction was driven by a significant increase in error rates when the default was accepted on high difficulty trials relative to when it was rejected ($t(15) = 2.45, P < 0.05$), with no differences in low difficulty default acceptance and rejection errors ($t(15) = 0.58, P = 0.57$). These behavioural effects were replicated in a separate experiment (N=18) outside the scanner (figure 5.3).

Judgment accuracy on low difficulty trials was 95.1 ± 1.0 % (s.e.m.). By design, accuracy on high difficulty trials was reliably lower ($t(15) = 24.3, P < 0.0001$), but remained significantly above chance (58.0 ± 1.3 % s.e.m., one-sample t-test against 50%, $t(15) = 5.71, P < 0.001$). As expected, rejection response times (RTs) were greater on high compared to low difficulty trials ($t(15) = 5.28, P < 0.001$). The distribution of RTs in the two difficulty conditions is shown in figure 5.4.

We next computed signal detection theory (SDT) measures from our data (see Chapter 3). This analysis confirmed shifts in criteria ($c$) as a function of default position (in/out) on high difficulty trials ($c_{in} = 0.31, c_{out} = -0.48$) but not low difficulty trials ($c_{in} = 0.049, c_{out} = 0.0052$), leading to a significant interaction of default and difficulty level ($F_{(1,15)} = 9.84, P < 0.01$). Changes in sensitivity ($d'$) due to difficulty level did not interact with default position (in/out; $F_{(1,15)} < 1, P = 0.69$).

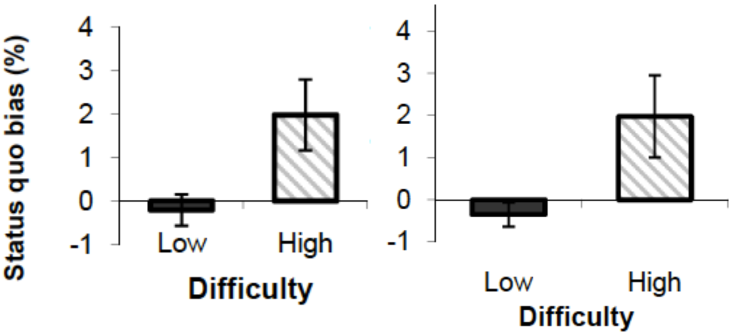

**Figure 5.3:** Inaction bias was calculated as the percentage of default acceptance greater than 50% on both high and low difficulty trials. A significant bias towards accepting the default was seen on high, but not low, difficulty trials. Left panel, fMRI experiment (N=16); right panel, behavioural replication (N=18). Error bars reflect ± s.e.m. Two subjects in the latter cohort also provided data in the main fMRI experiment. The task and experimental protocol were identical to the fMRI design, except the experiment was carried out seated in front of a computer monitor. The head was stabilised using a chin rest at a distance such that stimulus size and eccentricity was matched to that reported in the main text.

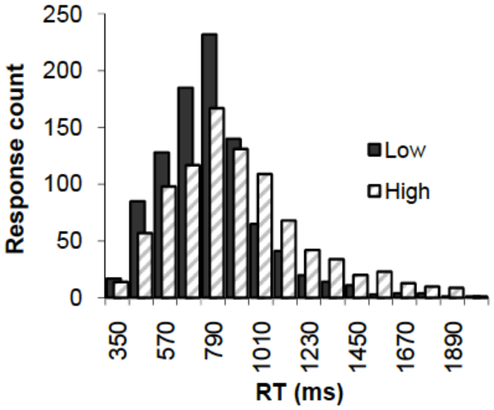

**Figure 5.4:** Histogram of reaction times split by decision difficulty (high and low).

### 5.3.2 fMRI

Our behavioural findings of an inaction bias on high but not low difficulty trials motivated us to explore the neural basis of this interaction. Crucially, we were interested in regions showing differential activity for rejection of the default under high difficulty compared to low difficulty. To isolate such regions, we computed an interaction contrast [reject_high - accept_high] - [reject_low - accept_low]. In this interaction we found activation in the right STN region that survived correction for the whole brain ($P < 0.05$, family-wise error (FWE) corrected; figure 5.5). Similar activation was found in left STN ($P < 0.05$, small-volume corrected (SVC)). No other brain regions survived whole brain correction, and the reverse contrast did not reveal any other significant interaction effects.

To further explore the observed interaction, we computed percent signal change for each trial type, averaging over all voxels within anatomically defined STN regions of interest (ROIs) (Aron & Poldrack, 2006) and entered these values into a repeated measures analysis of variance [ANOVA; factors STN_side (left/right) × decision (accept/reject) × difficulty (high/low)]. We found a significant interaction between decision difficulty and default rejection ($F_{(1,15)} = 17.70, P < 0.001$) that was consistent across both left and right STN (no three-way interaction with STN_side; $F_{(1,15)} < 1, P = 0.80$). A main effect of decision was also present (greater activity on reject trials; $F_{(1,15)} = 18.04, P < 0.001$). Specifically, the interaction effect is driven by increases in STN activity on trials where the default is rejected in the face of high decision difficulty, as shown in figure 5.5b. Importantly, this difference is similar for both correct and incorrect responses (no difference between grey and white bars in figure 5.5b), suggesting that the behavioural difference in accuracy for accept_high and reject_high responses cannot explain the signal change we observe in the STN.

As expected, we found a widespread motor network (table 5.1) when contrasting reject > accept responses, with greater activity on the left side consistent with

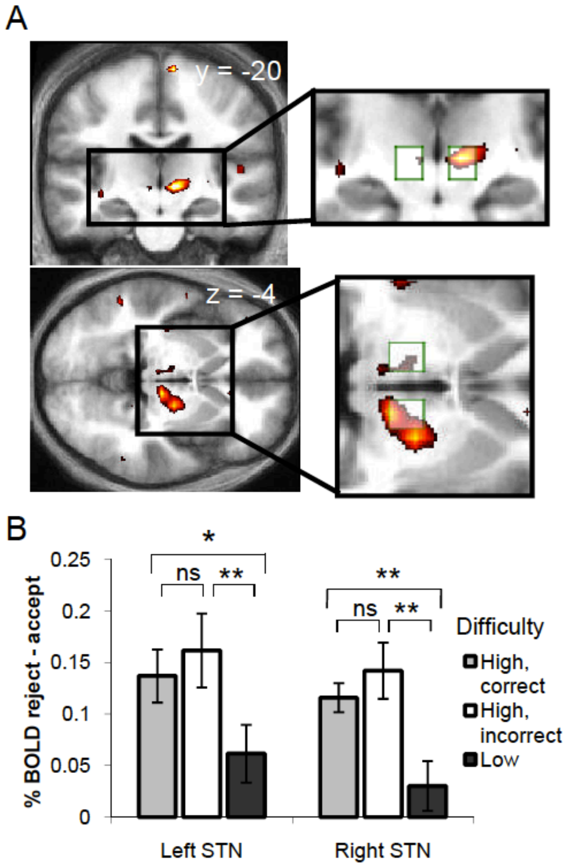

**Figure 5.5:** (A) $T$-map for the interaction contrast [(reject_high - accept_high) - (reject_low - accept_low)], shown in coronal and axial sections (right $P < 0.05$, whole-brain corrected, left $P < 0.05$, SVC; shown at $P < 0.005$, uncorrected). Activity is seen bilaterally in the region of the STN (peak voxels; left, -6,-24,-3; right, 12,-18,0). The insets (right) show the overlap between the active clusters and the anatomically defined STN ROIs (10 x 10 x 10 mm boxes centred on $\pm10, -15, -5$). (B) Average difference in percent signal change (reject accept) calculated from an unbiased average of all voxels within each STN box ROI. Events are split as a function of difficulty level. High difficulty trials were further split into correct and incorrect (the relative rarity of an incorrect, low difficulty response precluded the same split on low difficulty trials). The interaction effect was driven by a greater STN response for rejecting the default on high compared to low difficulty trials. Post-hoc paired t-tests; $^*P < 0.05$, $^{**}P < 0.005$. Error bars reflect ± s.e.m.

rejection responses being made with the contralateral (right) hand. The reverse contrast, accept > reject, did not reveal any significant activations.

Activity in bilateral inferior frontal cortex (IFC; $P < 0.05$, FWE whole-brain corrected) and bilateral dorsomedial prefrontal cortex (dmPFC; both $P < 0.05$, SVC) correlated with increasing reaction time for rejecting the default (figure 5.6 and table 5.2). We saw additional main effects of decision difficulty in both dmPFC ($P < 0.05$, FWE whole-brain corrected) and IFC ($P < 0.001$, uncorrected) (table 5.5).

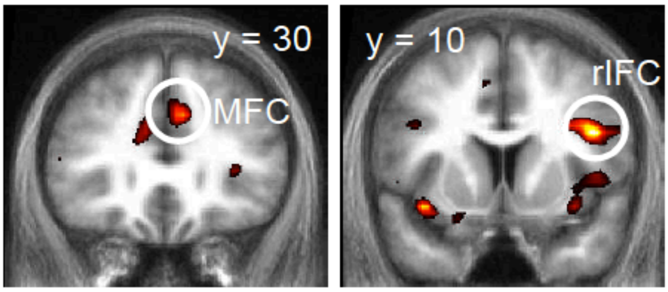

**Figure 5.6:** Coronal sections are shown through the group $T$-map for positive correlations with the reaction time (RT) regressor ($P < 0.005$, uncorrected).

### 5.3.3 Anatomical localisation of the interaction effect

To explore the anatomy of our interaction effect, group level clusters were projected onto the averaged structural from the same subjects in MNI space. With the aid of the atlas of Duvernoy (Duvernoy, 1999), the STN was localised as lying lateral and slightly anterior to the high-signal red nucleus when viewed on an axial slice. On a coronal section, the STN is separated from the grey matter of the thalamus by the zona incerta and the lenticular fasciculus. Using these landmarks, the group maximum for the interaction of decision difficulty and response type (12, -18, 0) was identified as lying ventral to the border of the thalamus, overlapping with the zona incerta/STN. The right-side cluster may extend dorsally into the body of the thalamus (figure 5.7), thus we cannot rule out a contribution of ventral thalamic motor nuclei to the interaction effect. However, the percent signal changes shown in figure 5.5 were calculated by averaging over all voxels within a priori STN ROIs, and are thus directly comparable to previous 'STN region' activations seen in recent studies investigating response inhibition (Aron & Poldrack, 2006; Aron *et al.*, 2007; Li *et al.*, 2008).

## 5.4 Discussion

The results in this chapter show that participants are more likely to accept a no-action default when faced with difficult choices, leading to more errors. This result

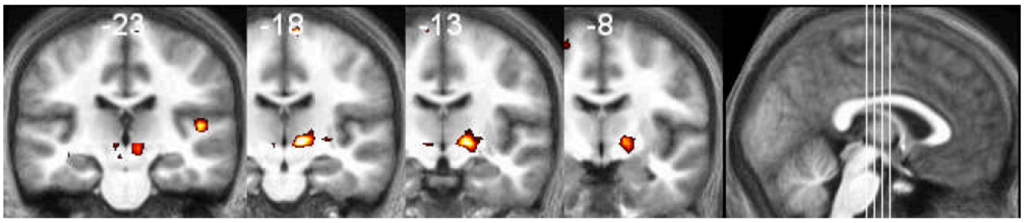

**Figure 5.7:** Multiple coronal views of the interaction effect displayed in 5.5

indicates that the default acted as an asymmetric loss function, leading to a bias towards acceptance when difficulty is high. In a signal detection analysis, participants showed criterion shifts towards the default on high, but not low, difficulty trials, suggesting that the threshold for action was modulated by difficulty. The brain imaging findings provide a neural perspective for how difficulty may modulate action initiation. In the fMRI data, rejection of the default on uncertain trials recruited bilateral regions encompassing the subthalamic nucleus (STN), a component of the basal ganglia thought to play a pivotal role in action selection (Bergman *et al.*, 1990; Frank, 2006). Specifically, BOLD signal increased in both left and right STN when the default was rejected on difficult, but not easy, trials. This effect was not explained by a change in decision accuracy. Instead, the interaction suggests a specific role for STN activity in switching away from a default when decisions are uncertain.

This context-dependence of STN activity is consistent with findings from deep-brain stimulation (DBS) studies which report a role for STN under conditions of high but not low difficulty in Parkinsons patients (Alberts *et al.*, 2008; Frank *et al.*, 2007; Hershey *et al.*, 2004). An alternative account might suggest that the activation we observe is epiphenomenal, rather than being causal in the amelioration of a default bias. I consider this possibility as less likely, for a number of reasons. First, the activity increase observed is specific to rejecting a difficult default, rather than rejection of the default per se, and is not easily explained through simple correlation with motor output or decision accuracy. Second, the effects we observe are anatomically specific and consistent across bilateral STN, a region proposed as a key node for control of decision making (Frank *et al.*, 2007; Gurney *et al.*, 2001). Finally, and perhaps most persuasively, deep brain stimulation in Parkinsons disease reveals a causal role for the STN in the modulation of decision making (Alberts *et al.*, 2008; Ballanger *et al.*, 2009; Hershey *et al.*, 2004; Ray *et al.*, 2009; van den Wildenberg *et al.*, 2006), while lesions to the STN in rodents produce impaired response selection under situations of high conflict (Baunez *et al.*, 2001; Eagle *et al.*, 2008).

The pattern of activity in the STN region can be further examined in the context of two influential models that address the broader role of the basal ganglia in decision making (Frank, 2006; Gurney *et al.*, 2001). In brief, it is proposed that activation of striatal neural populations by salient sensory stimuli drives selection of an appropriate response, releasing pallidal inhibition of the thalamus. A 'hyperdirect' pathway from frontal cortex to the STN (Nambu *et al.*, 2000) leads to modulation of pallidal-thalamic responses as a result of decision difficulty (Frank, 2006), adjusting basal ganglia output. Regions sensitive to task difficulty in the present study may represent putative sources of this hyperdirect signal.

Studies of the stop-signal reaction time (SSRT) task using fMRI have isolated both the right IFC and STN as critical nodes in the stopping of ongoing responses (Aron & Poldrack, 2006; Li *et al.*, 2008). Further, deep brain-stimulation of the STN in patients with Parkinsons disease directly modulates SSRTs (van den Wildenberg *et al.*, 2006; Ray *et al.*, 2009). In the present task, a simple inhibitory account of STN function would suggest greater activity when a difficult default is accepted (lack of action), whereas an account which emphasises a role for the STN in switching would predict greater activity when the default is rejected. The data in this Chapter favour the latter view, and together with related evidence (Aron *et al.*, 2007; Isoda & Hikosaka, 2008; Neubert *et al.*, 2010) implicate the STN in both outright response suppression and controlled slowing or switching.

In summary, I describe a potential neural mechanism for modulating action thresholds under uncertainty centred on STN. Difficult choice scenarios led to greater acceptance of the default, resulting in suboptimal decision-making. When the default was successfully rejected on difficult trials, a selective increase in STN region activity was found. I note however that the present design is unable to tease apart the influence of prior expectations from action costs as the primary drivers of the default bias. In other words, the pattern of STN activity I observed could either be due to a difficulty × action interaction, or a difficulty × prior interaction, as on each trial rejecting the default (a priori) response is synonymous with acting. In the following Chapter 6 I explicitly address this question with a focus on the STN.

## 5.5 Appendix

| Label | MNI coordinate | Voxels | Z-score | Voxel-FWE $P$-value |
|---|---|---|---|---|
| L postcentral gy. | -54 -15 51 | 575 | 6.12 | $< 0.001$ |
| R cerebellum | 24 -51 -30 | 610 | 5.95 | $< 0.001$ |
| L putamen | -33 0 -3 | 430 | 5.85 | $< 0.001$ |
| L precentral gy. | -57 3 30 | 46 | 5.79 | $< 0.001$ |
| L cingulate gy. | -6 -24 48 | 47 | 5.49 | 0.001 |
| R thalamus | 15 -15 3 | 115 | 5.45 | 0.001 |
| L SPL | -24 -54 51 | 25 | 5.35 | 0.002 |
| R MFG | 42 -6 57 | 67 | 5.34 | 0.002 |
| R IFG | 60 12 15 | 23 | 5.29 | 0.003 |
| R postcentral gy. | 63 -15 33 | 28 | 5.13 | 0.008 |
| R postcentral gy. | 36 -33 57 | 76 | 4.93 | 0.022 |
| R pre-SMA | 9 3 60 | 99 | 4.92 | 0.023 |
| R cerebellum | 30 -81 -21 | 6 | 4.89 | 0.028 |
| R cerebellum | 18 -60 -57 | 25 | 4.87 | 0.030 |
| L cerebellum | -36 -45 -30 | 57 | 4.85 | 0.033 |
| L cerebellum | -15 -63 -27 | 37 | 4.84 | 0.035 |
| R post. temporal | 42 -57 0 | 3 | 4.82 | 0.038 |
| L IFG | -39 33 27 | 6 | 4.79 | 0.045 |

**Table 5.1:** Significant activations for Reject $>$ Accept. Due to the widespread activity in this contrast, individual clusters were separated through adoption of a conservative height threshold of $P < 0.00001$.

| Label | MNI coordinate | Voxels | Z-score | Cluster-FWE $P$-value |
|---|---|---|---|---|
| R IFC | 45 12 24 | 140 | 4.94 | 0.002 |
| L insula | -39 9 -12 | 192 | 4.32 | < 0.001 |
| R insula | 33 18 9 | 296 | 5.79 | < 0.001 |
| L precentral gy. | -57 3 30 | 46 | 3.89 | < 0.001 |
| R cingulate gy. | 9 30 33 | 57 | 3.89 | 0.005, SVC |
| L cingulate gy. | -9, 24, 30 | 23 | 3.72 | 0.022, SVC |

**Table 5.2:** Significant activations correlating with reaction time. Clusters are defined using a threshold of $P < 0.005$, uncorrected.

| Label | MNI coordinate | Voxels | Z-score | Cluster-FWE $P$-value |
|---|---|---|---|---|
| L precentral gy. | -54 3 42 | 824 | 6.35 | < 0.001 |
| L/R MFC (paracingulate gy./pre-SMA) | 3 15 54 | 255 | 3.93 | < 0.001 |
| L MFG | -24 0 63 | 91 | 3.81 | 0.021 |

**Table 5.3:** Significant activations in the High > Low difficulty contrast. Clusters are defined using a threshold of $P < 0.005$, uncorrected.

# Chapter 6

# Control of the unexpected by human subthalamic nucleus

## 6.1 Introduction

We saw in Chapter 5 that the requirement to initiate non-default actions activates STN (see also Isoda & Hikosaka 2008; Coxon *et al.* 2010). A hallmark of human decision-making is its flexibility. When carrying out a well-practiced task, such as driving, an unexpected change in the environment  a cat running into the road, say - wrenches us into immediate evasive action. This situation results in response competition  between the ongoing process of driving, and the need to hit the brakes to avoid the cat. One influential account of these effects is that the need for cognitive control activates hyperdirect afferents to increase STN activity, allowing more time for the correct decision to be made, and resolving pre-response conflict (Frank, 2006). In support of this hypothesis, BOLD signal in the region of the STN increases during successful stopping in the stop signal paradigm (Aron & Poldrack, 2006; Li *et al.*, 2008), and correlates with the extent of slowing as task difficulty increases (Aron *et al.*, 2007). Furthermore, co-activation of prefrontal cortical loci, including the inferior frontal gyrus (IFG) and pre-supplementary motor area (pre-SMA)/anterior cingulate cortex (ACC) is often found during tasks requiring inhibitory control (see Aron 2010 for a comprehensive review); given that the STN receives monosynaptic input from these prefrontal structures via the so-called hyperdirect pathway (Monakow *et al.*, 1978; Nambu *et al.*, 2000; Aron *et al.*, 2007), cortico-STN circuitry may play a role in the resolution of response competition (Frank, 2006; Gurney *et al.*, 2001; Neubert *et al.*, 2010).

However, it is unknown how different factors contributing to response competition affect the STN. Control is thought to be necessary both when decisions are difficult, due to conflicting conditional action triggers (Botvinick *et al.*, 2001; Wendelken *et al.*, 2009), and when the environment is not in accord with our expectations (Holroyd & Coles, 2002; Braver *et al.*, 2001; Isoda & Hikosaka, 2008; Redgrave *et al.*, 2010). One view suggests that stimulus uncertainty is the main driver of STN activity. For example STN-DBS is known to have detrimental effects on performance when decisions are difficult (Jahanshahi *et al.*, 2000; Frank *et al.*, 2007), and activity in the nucleus increases when there is competition between putative controllers of decision-making (Fumagalli *et al.*, 2010). Alternatively, the history of responses (ones expectation) may be more important, rather than the difficulty of the current decision. For example, recent data suggest that updating prior action plans (see Mars *et al.* 2007) may engage similar mechanisms to inhibition of a single action (Aron & Poldrack, 2006; Mostofsky & Simmonds, 2008; Isoda & Hikosaka, 2008; Kenner *et al.*, 2010; Neubert *et al.*, 2010; Verbruggen *et al.*, 2010). In an important single-unit recording study, Isoda & Hikosaka (2008) showed that neurons in the STN are selectively active when a cue calls for a switch of prepotent response. We note that these influences on STN are not mutually exclusive; indeed, data from Chapter 5 is potentially consistent with both accounts: here, BOLD signal in the STN region was increased when the current default is rejected on difficult, but not easy, perceptual decisions.

I set about disambiguating the effects of difficulty and expectation on STN activity by analysing BOLD signal from anatomically-defined STN using rapid event-related fMRI. Our task dissociated, through factorial design, the effects of action, difficulty and expectation to changes in STN activity. I expected heightened difficulty and violation of expectations to slow down reaction times allowing for more controlled (less erroneous) responses to be made (Luce, 1991). More generally, the role of the basal ganglia in contextual modulation of this speed-accuracy tradeoff remains an open question (Bogacz *et al.*, 2010). For example, striatal activity, and structural projections between pre-SMA and striatum, closely correlate with individual differences in the lack of caution in a speeded decision task (Forstmann *et al.*, 2008, 2010). Given the potentially complementary roles of STN and striatum in cognitive control, here we compare results from the STN with a similar analysis of anatomically-defined striatum (caudate/putamen).

In Chapter 5, STN activation was localised using a 'box' region of interest at the group level (Aron & Poldrack, 2006). However, given the small size of STN and the unknown nature of inter-individual variability in location, I was only able to attribute changes in fMRI signal here to the STN region, not the STN proper. A more sensitive approach is to define regions based on individual anatomy. Here we isolate STN voxels using individual anatomy and high-resolution functional imaging[1], bypassing conventional normalisation and smoothing procedures to permit strong inferences on regionally localised activity.

[1] This study was conducted in collaboration with Christian Lambert at the Wellcome Trust Centre for Neuroimaging. See page 14 for details of contributions.

# 6.2 Methods

## 6.2.1 Participants

Twenty-one right-handed subjects gave informed consent to participate. All had normal or corrected-to-normal vision, and no history of psychological or neurological illness. Two participants displayed an unacceptable proportion of erroneous responses (> 50% omission errors to GO cues), and for one additional participant, STN proved difficult to isolate from anatomical scans (see below), leaving eighteen in the final cohort (9 female; 20 35 years of age; mean age, 25.11 years). The study was approved by the Institute of Neurology (University College London) Research Ethics Committee.

## 6.2.2 Task and procedure

The task was a modified version of the Eriksen flanker task (Eriksen & Eriksen, 1974). The flanker task has been showed to robustly modulate stimulus-bound difficulty, or conflict, dependent on whether the flankers are congruent or incongruent with the target (e.g. Botvinick *et al.* 2001). Two modifications to the classic flanker task were made to allow us to investigate our hypotheses of interest. First, the two response options were GO or NOGO, corresponding to a central 'Y' or 'X' cue respectively (figure 6.2). Second, the probability of having to make a GO or NOGO response was modulated in a blockwise fashion (p(go) = 0.3 or 0.7). This manipulation, coupled with 20% null (fixation-only) events per block, permitted analysis of changes in expectation ('surprise') of GO or NOGO events. Our design was thus 2 (difficulty) × 2 (action) × 2 (expectation) factorial.

Each trial lasted 2.5 seconds on average (figure 6.2), and each participant completed four blocks of 175 trials. GO, NOGO and null events were distributed randomly throughout the block according to their fixed prior probabilities, allowing effective event-related averaging despite the fast ITI (Burock *et al.*, 1998). The difficulty level (high or low) changed every 7 trials, and p(GO) was switched at the start of each block. Whether the experiment started with a high or low p(GO) block was counterbalanced between subjects. GO responses were made with a right-handed keypress on an MR-compatible button box. Prior to entering the scanner, participants completed one block of 175 trials (with equal GO and NOGO probabilities) on a standard PC to allow familiarisation with the task. Both speed and accuracy were emphasised, and a reaction time deadline of 550ms instigated during training to encourage a high level of action preparation. This deadline was removed in the scanner to avoid contamination of activity with feedback on response errors.

### 6.2.3 fMRI acquisition

Functional images were acquired using a 3T Trio whole-body scanner (Siemens, Erlangen, Germany) operated with a standard body transmit and 32-channel head receive coil. BOLD-sensitive functional images were acquired using a gradient-echo EPI sequence (30 transverse slices; TR, 2.55 s; TE, 85 ms; 2.3 × 2.3mm in-plane resolution; 2mm slice thickness; 1mm gap between adjacent slices; z-shim, +0.6 mT/m; positive phase encoding direction; slice tilt, - 30 degrees). The positioning of the slice block (field of view) was adjusted to cover the brainstem, striatum, dorsomedial and lateral PFC, and is shown for a single example subject in figure 6.4. Four runs of 187 volumes were collected for each subject. The first 5 volumes of each run were discarded before preprocessing to allow for T1 equilibration effects.

Anatomical images were collected using multiecho three-dimensional FLASH for mapping proton density, T1 and magnetization transfer (MT), and by T1-weighted inversion recovery prepared EPI sequences, all at $1\text{mm}^3$ resolution (Weiskopf & Helms, 2008). Additionally, field maps were acquired using a double-echo FLASH sequence (TE1 = 10ms, TE2 = 12.46ms, 3 × 3 × 2mm resolution, 1mm slice gap) to allow distortion correction of the EPI images. The midbrain nuclei are known to be particularly susceptible to physiological noise (e.g. D'Ardenne *et al.* 2008). To allow removal of this noise from the fMRI timeseries during analysis, I recorded participants' heart rate (using a pulse oximeter) and respiratory phase and volume (using a breathing belt).

### 6.2.4 Behavioural analysis

Reaction times (RT) to correct GO trials and the proportion of commission errors on NOGO trials were extracted for each difficulty/expectation condition and subjected to repeated-measures ANOVA using SPSS 17.0. There were very few omission errors on GO trials (mean 1.9 ± 1 % s.e.m.) and these trials were not analysed further.

### 6.2.5 Image preprocessing

Analaysis of fMRI data was carried out using Statistical Parametric Mapping software in Matlab2010b (SPM8 v4073, `www.fil.ion.ucl.ac.uk/spm`). The first five volumes of each run were discarded to allow for T1 equilibration. Using the FieldMap toolbox (Andersson *et al.*, 2001), field maps were estimated from the phase difference between the images acquired at the short and long TE. The EPI images were then realigned and unwarped using the created field map. These realigned and unwarped images were then used in the ROI analysis detailed below.

For exploratory whole-brain analysis, images were additionally normalized to Montreal Neurological Institute (MNI) space, and smoothed. Each subject's T1-weighted structural image was segmented into grey matter, white matter and cerebrospinal fluid, during which the transformation to the SPM MNI template space was estimated. The resulting normalization parameters were then applied to the functional data. Finally, the normalized images were spatially smoothed using an isotropic 6mm full-width half-maximum Gaussian kernel.

### 6.2.6 Anatomical ROI definition

The borders of the STN are defined by the zona incerta superiorly and immediately medially; prelemniscal radiations, lateral hypothalamus and red nucleus further medially and cerebral peduncle laterally. On its inferior-most lateral surface lies the superior aspect of the substantia nigra pars reticulata. The majority of the nucleus appears hypointense on T2-weighted images due to the presence of iron-containing neuromelanin (Dormont *et al.*, 2004; Marani *et al.*, 2008). Due to the variability in STN position and orientation, direct visualisation is the most accurate method to identify the structure (Hariz *et al.*, 2003; Ashkan *et al.*, 2007). We used R2* images, defined as $1/T2$, leading to the STN appearing hyperintense (Dormont *et al.*, 2004). The hyperintense region of the STN was manually segmented by Christian Lambert (see page 14) using ITK-SNAP software (`http://www.itksnap.org`; see figure 6.1). The anatomical guidelines listed above were used to aid identification of the superior and lateral boundaries of the STN, and where necessary ITK-SNAP's multi-session function was used to simultaneously visualise coregistered MT and proton density images. Finally, to allow comparison with previous ROI approaches (Chapter 5; Aron & Poldrack 2006; Aron *et al.* 2007; Li *et al.* 2008; Coxon *et al.* 2010; Neubert *et al.* 2010), we additionally applied MNI normalisation (as detailed above) to the STN images, and created an overlap map to reveal the extent of variability in ROI position across subjects (figure 6.1).

Striatal (caudate and putamen) ROIs were created on an individual basis using the automated subcortical segmentation routines implemented in FreeSurfer (`http://surfer.nmr.mgh.harvard.edu/`). Given previous studies implicating the anterior/dorsal striatum in modulation of inhibitory control (Aron *et al.*, 2003; Li *et al.*, 2008; Forstmann *et al.*, 2008, 2010), I combine caudate and putamen into a single striatal ROI.

### 6.2.7 Statistical modelling of the BOLD response

Statistical evaluation was performed using the general linear model. The design matrix consisted of delta (stick) functions aligned with the onset of the cues to correctly performed trials, and separated into four conditions dependent on whether the trial was high/low uncertainty and whether it required a GO or NOGO response. Null events were modelled with a separate regressor, and high and low p(go) blocks were modelled as separate sessions. GO trials were additionally para-

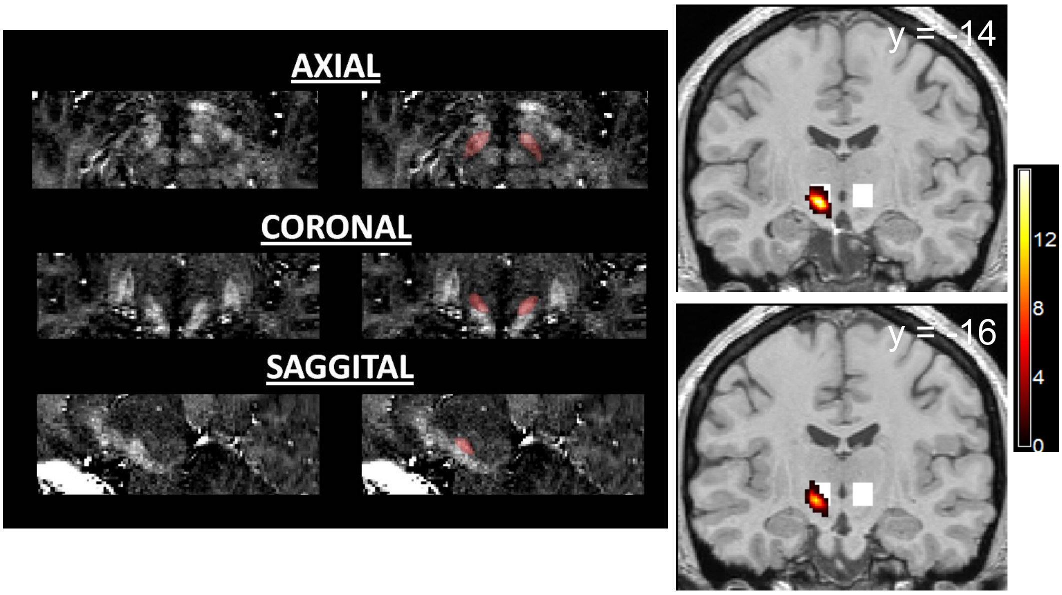

**Figure 6.1:** Left panels show sections through the midbrain and thalamus showing an example segmentation of the STN on R2* images for one subject. Right panels show the overlap obtained when normalising all individually-identified STN ROIs into MNI space. The colour bar indicates the number of subjects whose ROIs overlap at a given voxel. The box ROI used in Chapter 5 is reproduced here for comparison, demonstrating the advance in localisation achieved by accounting for inter-individual variability in STN anatomy.

metrically modulated by the z-scored reaction time (RT). These delta functions were convolved with the canonical HRF and its temporal derivative, and low-frequency drifts were excluded with a high-pass filter (1/128s cutoff frequency). Short-term temporal autocorrelations were modelled using an AR(1) process. Motion correction parameters estimated from the realignment procedure and physiological noise terms created from a Fourier expansion of heart rate and respiration timecourses were entered as covariates of no interest.

Contrast images for each condition in our factorial design were generated by subtracting the relevant within-session null event. A ‘surprising’ event was considered low-frequency for that block type. Thus high p(go) blocks contained surprising NOGO events, and expected GO events. Conversely, a low p(go) block contained surprising GO events, and expected NOGO events. Single-subject contrast images were estimated using both native-space, unsmoothed functional data, and normalised, smoothed functional data, for entry into the ROI and whole-brain analyses, respectively.

### 6.2.8 ROI analysis

To allow extraction of functional data from anatomically-defined ROIs, the T1-weighted structural was coregistered with the mean EPI image, and the same transformation applied to the ROI images. ROI images were then resliced into the same voxel dimensions (2.3 × 2.3 × 3mm) as the functional data. For each subject, the contrast estimates for each condition (created from a GLM analysis of native-space, unsmoothed functional data) were extracted for each voxel in a particular ROI and averaged. These contrast estimates were analysed using a 2 (hemisphere) × 2 (action) × 2 (difficulty) × 2 (expectation) repeated-measures ANOVA. Having established a lack of significant interactions with hemisphere, we collapsed across this factor in subsequent analyses. To explore whether the effect of surprise on STN activity was action specific, follow-up 2 (difficulty) × 2 (expectation) ANOVAs were carried out for GO and NOGO trials separately.

Having observed an effect of expectation on STN responses, we additionally tested whether this differential response was linked to behavioural slowing, by creating separate contrasts for the parametric correlation of activity with RT separately for surprising and expected GO trials. Positive correlations with RT were assessed through two-tailed one-sample t-tests against zero at the group level.

All statistical analysis was carried out using SPSS 17.0, using the default Type III sum-of-squares estimation for repeated-measures ANOVA.

### 6.2.9 Whole-brain analysis

Contrast images for each condition in our factorial design (created from a GLM analysis of normalised, smoothed functional data) were entered into a second-level ANOVA in SPM8. I examined the main effects of action, surprise and difficulty using T-contrasts. Cluster-based statistics were used to define significant activations both on their intensity and spatial extent (Friston *et al.*, 1994b). Clusters were defined using a threshold of $P < 0.001$ and corrected for multiple comparisons across the whole brain volume using family-wise error correction (FWE) and a threshold of $P < 0.05$.

## 6.3 Results

### 6.3.1 Behaviour

Participants tended to make few errors overall, with the main source of errors being commission errors on high difficulty, NOGO trials. Formal analysis of commission error rates showed a robust effect of both difficulty ($F_{(1,17)} = 20.8, P < 0.00001$) but no significant effect of expectation ($F_{(1,17)} = 2.8, P = 0.11$). As expected, increases in difficulty and violations of expectation both increased reaction times (RTs) on GO trials (difficulty; $F_{(1,17)} = 34.5, P < 10^{-5}$; expectation, $F_{(1,17)} = 114.9, P < 10^{-9}$). In addition, there was a significant interaction between difficulty and expectation ($F_{(1,17)} = 4.89, P = 0.04$) driven by a greater speeding effect of expectation on low compared to high difficulty trials.

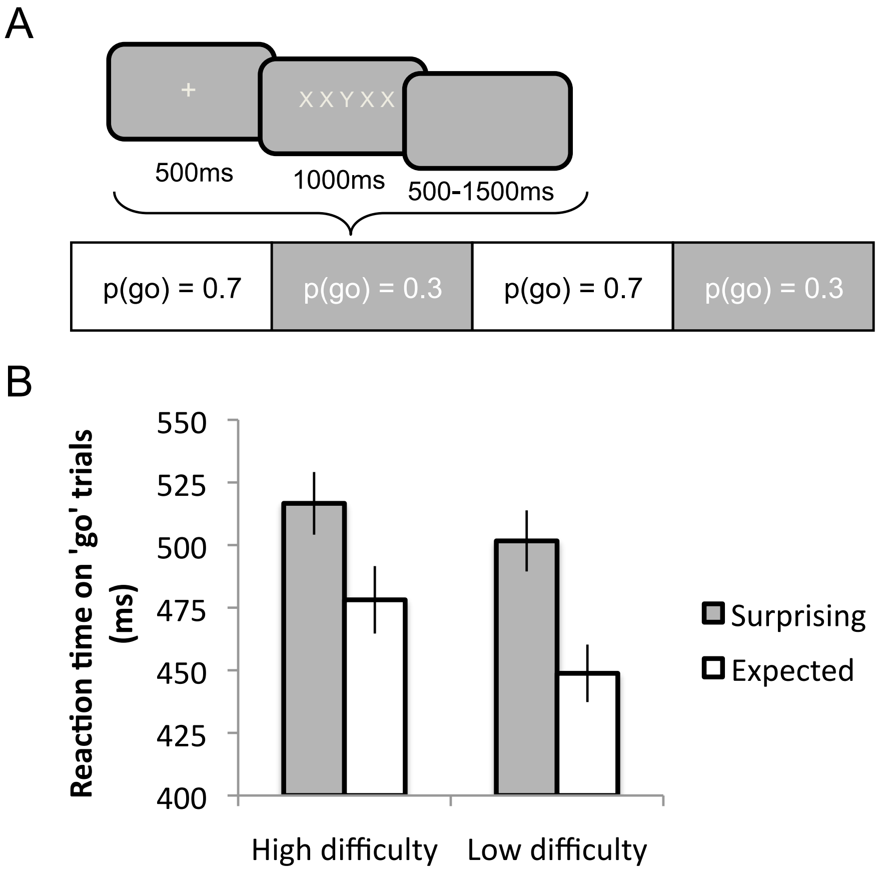

**Figure 6.2:** (A) Schematic of the stimuli and task. Participants fixated on a central cross, and were instructed to press a button with their right hand as quickly as possible when a central 'Y' was presented, and withhold the button press when a central 'X' was presented. Flanking stimuli could either be congruent (low difficulty) or incongruent (high difficulty). Whether the GO or NOGO response was expected was controlled by manipulating the probability of a GO response between blocks (p(go)). (B) Mean reaction times (excluding outliers > 3SD from the mean of each condition) for GO trials, conditional on whether the trial was high or low difficulty, and whether the GO response occurred on surprising (low frequency) or expected (high frequency) blocks. Significant effects of both expectation and difficulty were obtained (both $P < 10^{-4}$).

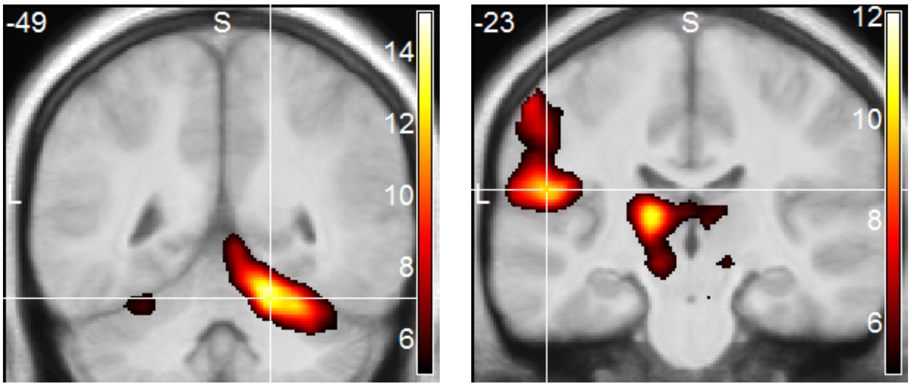

**Figure 6.3:** Whole-brain analysis of the main effect of action (GO > NOGO). As expected, increased activity was seen in cerebellum ispsilateral to the response hand, and in premotor cortex and thalamus contralateral to the response hand.

| Contrast | Voxels | Z-score | cluster $P$-value | Peak voxel | Laterality | Label |
|---|---|---|---|---|---|---|
| Surprising > Expected | 452 | 5.53 | $< 0.001$ | -32 16 13 | L | Insula/IFG |
| | 684 | 5.43 | $< 0.001$ | -7 19 31 | L | dmPFC |
| | 526 | 5.13 | $< 0.001$ | 37 19 7 | R | Insula/IFG |
| | 112 | 4.73 | 0.01 | -30 39 28 | L | BA46 |
| | 215 | 4.69 | $< 0.001$ | 30 41 22 | R | BA46 |
| | 110 | 4.39 | 0.011 | 66 -41 22 | R | post. STG |
| | 109 | 4.37 | 0.011 | -7 -20 31 | L | post. cingulate |
| | 165 | 4.23 | 0.001 | -34 -57 -14 | L | fusiform gy. |
| | 332 | 4.15 | $< 0.001$ | 14 -75 16 | R | V1 |
| High > low diff. | 38 | 4.05 | 0.326 | 25 -66 34 | R | SOG |
| | 22 | 3.93 | 0.680 | -23 -59 46 | L | Sup. parietal |
| | 11 | 3.71 | 0.938 | -30 -87 13 | L | MOG |
| | 10 | 3.50 | 0.952 | -25 -73 28 | L | MOG |
| Low > high diff. | 628 | 5.15 | $< 0.001$ | 14 -98 10 | L | V1 |
| | 125 | 3.92 | 0.006 | -12 -101 7 | R | V1 |
| GO > NOGO | 4555 | Inf | $< 0.001$ | 16 -50 -17 | R | Cerebellum |
| | 263 | 7.00 | $< 0.001$ | 39 2 7 | R | post. insula |
| | 300 | 6.40 | $< 0.001$ | -30 -55 -23 | L | Cerebellum |
| | 66 | 6.19 | $< 0.001$ | 57 -14 19 | R | Postcentral gy. |
| | 70 | 5.53 | $< 0.001$ | 14 -25 -5 | R | Thalamus/brainstem |
| | 126 | 5.29 | $< 0.001$ | 0 5 43 | L/R | ACC |
| | 10 | 4.87 | 0.019 | 7 -23 -20 | R | Brainstem |
| | 62 | 4.85 | $< 0.001$ | -5 -25 28 | L | post. cingulate |
| NOGO > GO | 706 | 6.23 | $< 0.001$ | 39 -20 52 | R | Postcentral gy. |
| | 108 | 4.62 | 0.012 | 11 -23 52 | R | SMA |

**Table 6.1:** Whole-brain results. Activations are reported for the contrasts discussed in the main text. Activations survive cluster-level correction for multiple comparisons across the whole brain volume ($P < 0.05$), except for the high > low difficulty contrast ($P < 0.001$, uncorrected, 10 voxel extent). The cluster-defining threshold is $P < 0.001$ uncorrected, except for GO > NOGO where a higher threshold of $P < 0.00001$ uncorrected was used to separate individual clusters of activation.

### 6.3.2 Functional imaging analysis

As an initial check analysis, I examined the simple GO > NOGO contrast at the whole-brain level. This contrast revealed contralateral ventral premotor cortex, and ipsilateral cerebellum, as expected (see figure 6.3). The reverse contrast, NOGO > GO, revealed activity increases in the ipsilateral premotor cortex and SMA (see table 6.1), although I note this activity could equally be interpreted as reflecting a decrease below baseline during motor execution.

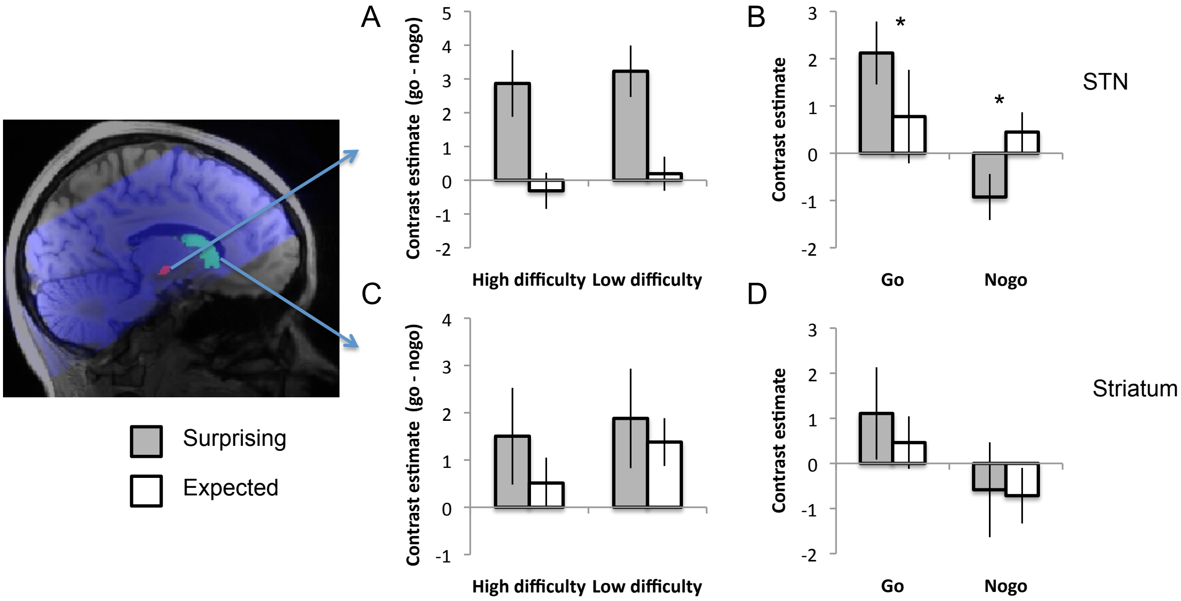

**Figure 6.4:** Left panel shows example ROIs from a single subject for STN (red) and caudate/putamen (green). The transparent blue region indicates the slice block used for functional data acquisition of this subject. (A, C) Contrast estimates examining the coding of action (GO - NOGO) as a function of each cell in our difficulty × expectation factorial design. STN shows a robust interaction between action (GO - NOGO) and expectation, such that action coding is seen on surprising but not expected trials ($P = 0.006$). No interactions were seen in striatum. (B, D) Effects of expectation collapsed across difficulty level, as a function of both GO and NOGO responses. STN is seen to differentiate between expected and surprising trials for both GO and NOGO responses (both $P < 0.05$).

#### 6.3.2.1 ROI analysis

Defining the STN on a single-subject basis revealed heterogeneity between subjects (figure 6.1). However, when normalised to a common template, most STN voxels lay within the box ROI commonly employed in previous studies (Aron & Poldrack, 2006; Aron *et al.*, 2007; Coxon *et al.*, 2010; Li *et al.*, 2008; Neubert *et al.*, 2010), indicating reasonable correspondence between a single-subject and group ROI approach. I extracted BOLD signal from each subjects anatomically-defined STN for each condition of our factorial design and subjected this signal to a 2 (action) × 2 (difficulty) × 2 (expectation) repeated-measures ANOVA. Our central question was whether violations of expectation or current-trial difficulty, or both, were the key modulators of STN activity. Consistent with the former hypothesis, a robust interaction between expectation and action was observed ($F_{(1,17)} = 10.0, P = 0.006$), driven by a coding of action (increase of GO relative to NOGO) on surprising but not expected trials (figure 6.4a). This effect of expectation on action was similar for both low and high difficulty trials (interaction with difficulty; $F_{(1,17)} = 0.03, P = 0.87$; figure 6.4a). No main effects of expectation ($F_{(1,17)} = 0.20, P = 0.66$) or difficulty ($F_{(1,17)} = 0.21, P = 0.65$) were found. A main effect of action was also present, driven by greater signal for GO than NOGO trials ($F_{(1,17)} = 19.9, P < 10^{-4}$), as reported in Chapter 5.

To further test whether the modulatory effect of expectation on STN activity held for both GO and NOGO trials, I carried out separate 2 (difficulty) × 2 (expectation) ANOVAs. Significant, albeit opposite, effects of expectation were observed for both GO ($F_{(1,17)} = 5.47, P = 0.03$) and NOGO ($F_{(1,17)} = 7.23, P = 0.02$) trials, shown in figure 6.4b. Importantly, observing a significant effect of expectation for NOGO trials obviates concerns that this effect is driven by subtle alterations in motor execution (such as RT differences; Yarkoni *et al.* 2009), as no motor response was made on these trials.

In contrast, a similar analysis of striatum did not reveal effects of expectation. Instead, caudate/putamen robustly coded for action, showing a greater response for GO than NOGO trials (figure 6.4c and d; $F_{(1,17)} = 17.2, P = 0.001$). In addition, we observed a marginal interaction between action and difficulty ($F_{(1,17)} = 3.39, P = 0.08$), driven by a greater GO response on low difficulty trials (6.4c).

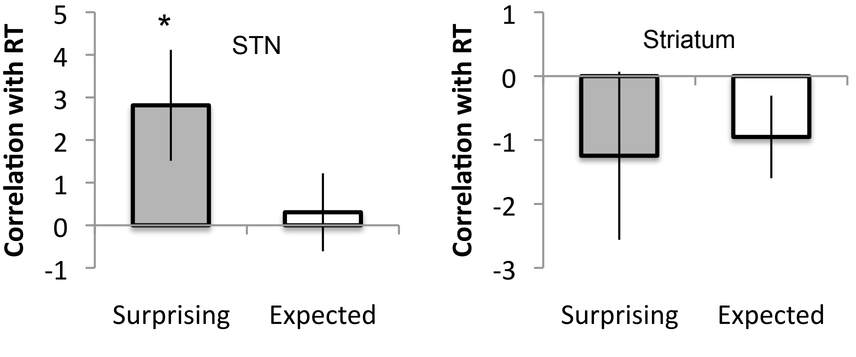

**Figure 6.5:** Group-level parameter estimates for the correlation between average STN response on GO trials and the associated reaction time, as a function of expectation. STN showed significant positive parameter estimates for the relationship between RT and activity on surprising, but not expected, GO trials. No significant relationship was observed between RT and activity in striatum.

#### 6.3.2.2 Linking surprise-induced slowing to STN activity

I hypothesised that the selective change in activity due to surprising events in STN may mediate the behavioural slowing manifest on these trials (Frank, 2006). If this hypothesis is correct, one would expect activity here to correlate with the extent of slowing on a trial-by-trial basis (cf. Aron *et al.* 2007). To test this hypothesis, I extracted the contrast estimate testing for the parametric correlation with reaction time separately for surprising and expected trials. I observed a significant positive parametric relationship between STN activity and RT on surprising (two-tailed one-sample t-test, $t(17) = 2.23, P = 0.04$), but not expected ($t(17) = 0.34, P = 0.74$), trials (figure 6.5). Reaction time parameter estimates in striatum tended to be negative, but did not reach significance (figure 6.5).

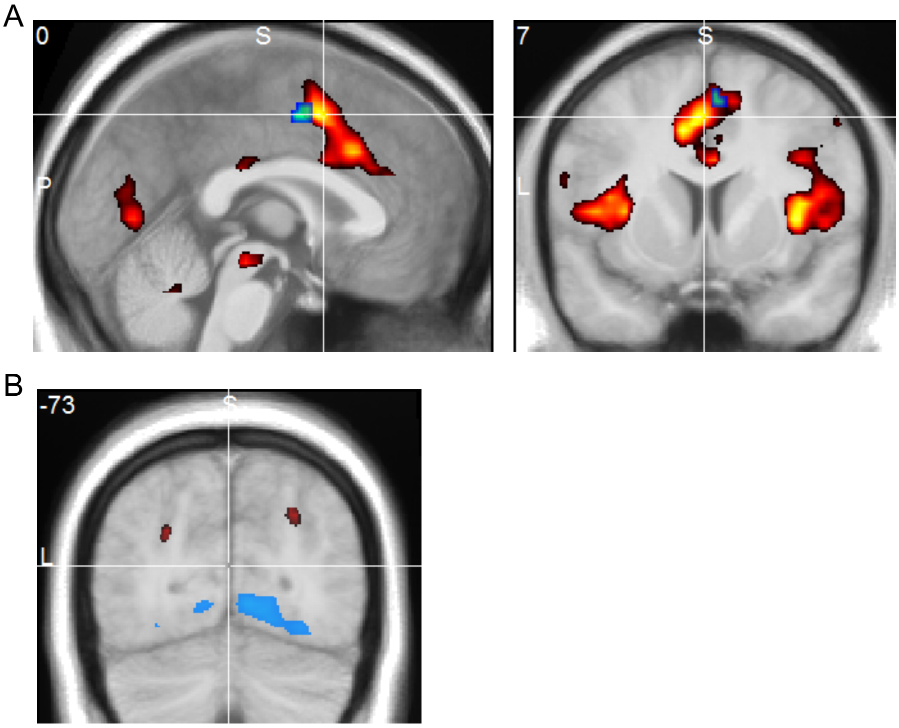

**Figure 6.6:** (A) Hot colours indicate activity (SPM $T$-map, shown at $T > 3$) showing a main effect of expectation (surprising > expected). Significant clusters were found in pre-SMA/ACC, insula, IFG, posterior cingulated and extrastriate cortex (not shown). Cool colours indicate activity (SPM $T$-map, shown at $T > 3$) correlating with GO-trial RT in each cell of my factorial design, obtained through a conjunction analysis. The SMA was the only region active in this contrast. (B) Hot (cool) colours indicate positive (negative) effects of difficulty (SPM $T$-map, shown at $T > 3$). Significant clusters were found in parietal cortex and primary visual cortex.

### 6.3.2.3 Whole-brain analysis

At a whole-brain level, I observed several regions in visual cortex and medial and lateral frontal cortex showing increased activity for surprising events (figure 6.6a and table 6.1), consistent with previous reports (Strange *et al.*, 2005; Rosa *et al.*, 2010). This network included ACC/pre-SMA, and inferior frontal cortex/insula, both known to share anatomical connections with STN (Aron *et al.*, 2007). The reverse contrast (greater activity for expected compared to surprising events) did not show significant activity. In contrast, the main effect of difficulty was associated with modulation of activity in visual and parietal cortex (figure 6.6b). Visual cortical activity (BA17) increased on low compared to high difficulty trials ($P < 0.05$, cluster-level corrected). Parietal cortex activity was increased for high compared to low difficulty trials, albeit at an uncorrected threshold ($P < 0.001$, uncorrected). Given that our restricted field of view was not optimised for detecting activity in the posterior brain (cf. figure 6.4), I do not interpret these activations further.

No significant effects of surprise or difficulty were observed on correlations with RT at the whole-brain level. A conjunction analysis of the parametric effect of slowing across each cell of the factorial design revealed pre-SMA as the only region correlating with behavioural slowing across all four conditions (figure 6.6a), consistent with a hypothesised role for this region in action inhibition (Sharp *et al.*, 2010).

## 6.4 Discussion

Previous studies have shown a link between STN function and inhibition of decision and action when the need for cognitive control arises (e.g. Aron & Poldrack 2006; Li *et al.* 2008; Coxon *et al.* 2010). Such functionality might explain why STN DBS in Parkinsons disease has subtle detrimental effects on executive function (e.g. Hershey *et al.* 2004). However, cognitive control is a multi-faceted construct (Ridderinkhof *et al.*, 2004), and it is unknown how these components map onto STN function. Here I dissociate, through factorial design, the effects of current-trial difficulty and changes in expectation, with both inducing separate main effects upon behaviour (RTs). I provide evidence that STN activity is modulated in response to unexpected events in a simple GO/NOGO task. This response (figure 6.4a) is similar both for high and low difficulty decisions, suggesting that a primary function of STN is to switch between action plans based on new and unexpected information (see also Isoda & Hikosaka 2008).

Importantly, I find a modulation of STN activity due to unexpected events for both GO and NOGO trials. This result refines the interpretation of results from my previous study (Chapter 5), where STN activity and connectivity with right IFC was found to be increased when a difficult ‘default’ response option was rejected, consistent with implementation of increased cognitive control through cortico-STN coupling. In this previous experiment, rejecting the default was synonymous with acting, preventing dissociation of these influences on STN function. In parallel to findings in the pre-SMA/ACC (Braver *et al.*, 2001; Nieuwenhuis *et al.*, 2003), the results of the current study suggests the STN does not show a ‘nogo-dominant’ response, but rather responds to events that are not in accordance with expectations (Isoda & Hikosaka, 2008). Indeed, Isoda & Hikosaka (2008) found both ‘go’ and ‘nogo’ switch cells in the STN, albeit a greater proportion of the latter. Functional imaging of elderly participants provides convergent data in support of this interpretation: here, deficits in switching between opposite directions of circular movement were associated with decreased modulation of bilateral STN (Coxon *et al.*, 2010).

The sign of STN modulation was opposite for unexpected GO and NOGO events. This interaction pattern is consistent with the view of STN as a cognitive modulator of motor output (Gurney *et al.*, 2001; Frank, 2006), potentially via distinct subnuclear territories (Hershey *et al.*, 2010; Mallet *et al.*, 2007). How the concepts of inhibitory control connect to action through the STN remains to be determined. The strong main effect of GO > NOGO may be due to reverberant effects in cortico-basal ganglia loops (Gradinaru *et al.*, 2009), and the increase in activity known to occur during movement execution itself (Wichmann *et al.*, 1994; Nambu *et al.*, 2000). Future studies could investigate this phenomenon by equating action, for instance through use of task-switching designs (Isoda & Hikosaka, 2008; Neubert *et al.*, 2010).

### 6.4.1 Indirect or hyperdirect pathway?

The STN is traditionally thought of as being part of the indirect pathway (Alexander & Crutcher 1990; figure 2.4). More recent studies have indicated a functional role for the cortical hyperdirect pathway into the STN, originating in prefrontal cortical sites known to be important for cognitive control (Nambu *et al.* 2000; Aron *et al.* 2007; Neubert *et al.* 2010; see Aron 2010 for a review). To the extent that the indirect pathway is considered inhibitory, our results linking STN to behavioural slowing may be consistent with activation of either pathway. However, given that changes in expectation are likely to be encoded in prefrontal cortex (e.g. Summerfield & Koechlin 2008), and that the dorsal striatum (also a component of the indirect pathway) did not show effects of expectation, I suggest the results are consistent with activation elicited by hyperdirect inputs into STN.

### 6.4.2 Effects of difficulty

In my previous study we found that the STN region was activated when non-default actions were initiated on difficult, but not easy, trials (Chapter 5). How can this finding be reconciled with the absence of difficulty effects here? First, the STN activity observed on non-default GO trials in my previous study can be attributed to increases in difficulty, the need to switch away from the default, or both. Here, through use of a factorial design, I dissociate these influences, demonstrating that STN activity is modulated on low-frequency trials requiring switches from the prepotent action plan, and that this modulation is similar for high and low difficulty trials. In addition, STN activity was correlated with behavioural slowing on unexpected, but not expected, trials. The absence of an interaction with difficulty could be due to the fact that the current task was in general more difficult than our previous task, due to the emphasis on speed and the relatively rapid stimulus presentation. Thus the 'low difficulty' condition in the current experiment may still require substantial cognitive control.

Similar to the present findings, Coxon *et al.* (2010) observed that activation in the STN in young people was increased during switches of action-plans, and that this increase was similar for low and high difficulty trials. However, there was a selective decrease in STN function solely for difficult switches in the elderly, potentially explaining the behavioural deficit observed in this condition. STN-DBS has also been reported to exert selective affects upon difficult decision-making (Frank *et al.*, 2007; Baunez *et al.*, 2001), although expectation was not manipulated as a factor in these designs. However DBS is unlikely to only affect activity within the nucleus, but also other interconnected areas in cortex (Ballanger *et al.*, 2009; Gradinaru *et al.*, 2009; Mallet *et al.*, 2007), regions that may be involved in the resolution of stimulus-response conflict. Further work is thus required to tease apart the role of the STN and interconnected regions during increases in task difficulty.

### 6.4.3 Relationship between switching and inhibitory control

Recent work has emphasised that the mechanism underlying action inhibition may be a special case of a broader circuit specialised for switching between different action plans (Aron *et al.*, 2007; Buch *et al.*, 2010; Kenner *et al.*, 2010; Mostofsky & Simmonds, 2008; Verbruggen *et al.*, 2010). The STN is reliably activated during successful stopping in the stop-signal reaction time (SSRT) task (Aron & Poldrack, 2006; Aron *et al.*, 2007; Li *et al.*, 2008). Assuming that our high/low p(GO) blocks differ to the extent that a GO response is prepotent, stopping on an infrequent NOGO trial may be similar to stopping on a stop trial in the SSRT task (Aron, 2010). In other words, to the extent that the stop-signal is relatively infrequent, the SSRT can be conceptualised as a task involving an occasional and rapid switch from the default GO response. Recordings from STN unit activity reveal selective increases on switch trials (Isoda & Hikosaka, 2008). Furthermore, the timing of the activity predicted whether the behavioural switch would be successful or not. These data are consistent with theoretical models where the STN temporarily inhibits the output structures of the basal ganglia to allow a new, correct response to be selected (Aron, 2010; Frank, 2006; Nambu *et al.*, 2002).

One further consideration is that the present design does not dissociate a selective inhibitory signal – for a particular effector – from a more global stopping requirement (Aron & Verbruggen, 2008), as participants knew that a single button press either would or would not be required on each trial. Thus the increase in signal I see for unexpected GO events may be attributable to a global stopping signal that permits subsequent selection of the correct action (Nambu *et al.*, 2002); or, alternatively, a selective 'switch' signal. Further studies using methodology designed to separate global and selective control are required to decide which hypothesis best accounts for STN function (Aron, 2010; Aron & Verbruggen, 2008). In addition, unexpected, or surprising, trials in the current design may have induced greater attentional orienting, due to their oddball status (Sharp *et al.*, 2010), consistent with increases in visual cortex activity in this contrast (table 6.1). I am unable to dissociate this attentional effect from the requirement for action reprogramming in the present design. However, I note the direction of the surprise effect was opposite for GO and NOGO trials, suggesting STN is involved in linking surprising events to the reprogramming of the motor response, rather than simply in detection of salient events. Experimental paradigms designed to dissociate attention and action control at a cortical level (Dodds *et al.*, 2010; Sharp *et al.*, 2010; Verbruggen *et al.*, 2010) could be usefully employed here to investigate the possible role of STN in attentional orienting over and above its putative role in action reprogramming.

### 6.4.4 Conclusions

Cognitive control is a multi-faceted construct, but is broadly invoked when the resolution of response competition is required. The anatomical position of the STN makes it well-situated to connect control signals to changes in motor output, but its precise role in this process is unknown. In this chapter I demonstrate that the BOLD signal in human STN responds to an unexpected change in action plan, in a manner dependent on whether action had to be initiated or withheld. Furthermore, STN activity was correlated with behavioural slowing when go responses were unexpected, but not expected. Together these findings indicate that STN is a key node in resolving the discrepancy between ongoing cognition and unexpected events through modulatory effects on action.

# Chapter 7

# Post-decision confidence during perceptual decision-making

## 7.1 Introduction

Knowledge of one's own uncertainty regarding an outcome plays a key role in determining decision strategy (Hampton, 2001; Kepecs *et al.*, 2008; Metcalfe, 1996; Smith *et al.*, 2004), and has been suggested to be a central property of higher-order consciousness (Cleeremans *et al.*, 2007; Kunimoto *et al.*, 2001; Lau, 2008; Persaud *et al.*, 2007). For example, knowing in advance that you are unlikely to pass a test might make you reluctant to take the test in the first place (Higham, 2007; Koriat & Goldsmith, 1996; Metcalfe, 2008). Up until this point I have considered how the observer uses context to optimise decisions under uncertainty, but the extent to which decision uncertainty is subjectively accessible after implementation of a decision is unclear. 'Subjectively accessible' here refers to overt reporting of confidence about a particular cognitive state, via metacognitive reports. Such reports are often taken to index conscious awareness of a particular state, a topic I will return to in Chapter 9. The goal of the present chapter is to extend the signal detection theoretic (SDT) approach outlined in Chapter 3 to metacognitive reports, and ask whether predictions derived from this account provide a good fit to metacognitive reports measured during perceptual decision-making.

The accuracy of metacognitive assessments can be intuited as how transparent the initial decision process is to a putative 'higher' level assessment. For example, if there is ambiguity in this decision process then the categorisation of one's own performance as being correct, or incorrect, will be subject to error. In standard applications of SDT (Type 1), detection performance is assessed by a comparison of the proportion of 'hits' and 'false alarms' in a stimulus detection task. By applying the logic of SDT to post-decision wagering, one can categorise a 'hit' as a high confidence response after a correct decision and a 'false alarm' as a high confidence response after an incorrect decision (see table 3.2), a type of analysis known as

Type 2 SDT (Clarke *et al.*, 1959; Clifford *et al.*, 2008; Galvin *et al.*, 2003; Kunimoto *et al.*, 2001). On the basis of these considerations, I derive a theoretical relationship linking Type 1 and Type 2 task performance by applying an ideal observer framework (Galvin *et al.*, 2003), enabling an exploration of the relationship between confidence and changes in decision performance.

I ask two questions[1]. First, I test the fit of a Type 2 SDT model to metacognitive confidence data collected during a perceptual task in which contributions of external noise (stimulus variability) are minimised. Second, I harness adaptive psychophysics procedures to test a prediction of the SDT model: that Type 2 performance (metacognitive ability) should scale with Type 1 performance (*d'*) across individuals. To pre-empt the results, I find that while a Type 2 model provides a good fit to post-decision confidence ratings, metacognitive ability is partially independent from perceptual performance, suggestive of a dissociable second-order stage of decision-making.

## 7.2 Type 2 signal detection model of metacognitive reports

### 7.2.1 Outline of model

In SDT, a Type 1 decision is based upon overlapping Gaussian probability distributions over a random variable $X$, conditional on the events signal ($S$) and noise ($N$) (figure 7.1). Assuming an unbiased response criterion, $c$, for the Type 1 detection decision, we can specify the distribution over $X$ for the probability of the Type 1 response being correct or incorrect:

$$\begin{aligned} f(x|C) &= \begin{cases} \dfrac{f(x|N)}{p(C)}, x \leq c \\ \dfrac{f(x|S)}{p(C)}, x > c \end{cases} \\ f(x|I) &= \begin{cases} \dfrac{f(x|S)}{p(I)}, x \leq c \\ \dfrac{f(x|N)}{p(I)}, x > c \end{cases} \end{aligned} \tag{7.1}$$

where $p(C)$ and $p(I)$ are the average probabilities of making a correct or incorrect response on any given trial (Macmillan & Creelman, 2005). Full derivations of equations 7.1 can be found in Galvin *et al.* (2003); the (constant) prior terms from their more general analysis are omitted here for clarity. The distributions specified by equations 7.1 are plotted graphically in figure 7.1. It is important to note that these Type 2 distributions are conditional transformations based on whether the

[1] The work presented in this chapter and in Chapter 8 was carried out in collaboration with Rimona Weil, Zoltan Nagy and Geraint Rees. See page 14 for details of contributions.

first decision was correct or not. That is, the shape of the $f(x|C)$ curve follows the signal distribution when $x > c$ (a Type 1 hit) and the noise distribution when $x < c$ (a Type 1 correct rejection). Similarly, the shape of the $f(x|I)$ curve follows shape of the noise distribution when $x > c$ (a Type 1 false alarm) and the signal distribution when $x < c$. The heights of both $f(x|C)$ and $f(x|I)$ are then scaled so that they sum to one.

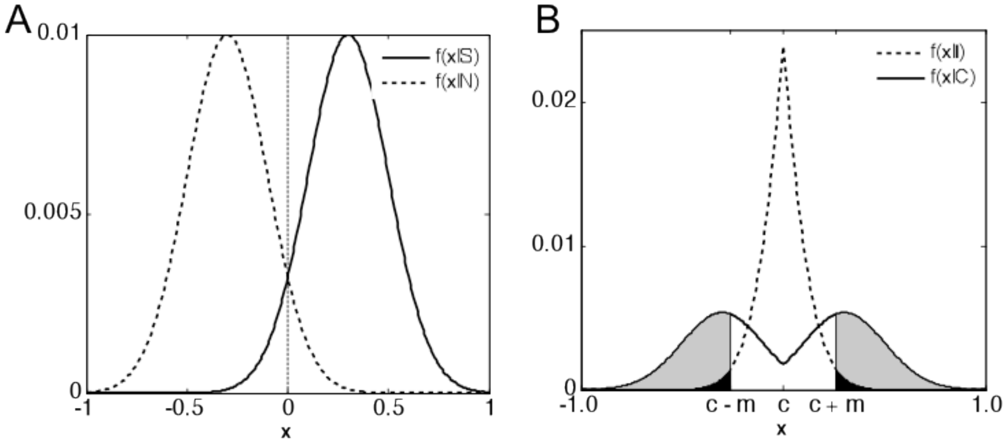

**Figure 7.1:** (A) Theoretical distributions over a random variable X (corresponding to an arbitrary stimulus axis) for signal ($S$, solid line) and noise ($N$, broken line). (B) Probability distributions over different values of $X$ for the probability of making a correct (solid line) and incorrect (broken line) categorisation. Shaded areas represent the integrals specified in equations 7.3 ($H$, grey; $FA$, black).

A correct Type 2 response is more likely towards the left or right-hand extremes of $X$ in figure 7.1 (high signal or high noise trials), whereas incorrect responses predominate where there is maximal overlap between signal and noise. The inherent assumption here is that the uncertainty associated with being sure of seeing something is the same as the uncertainty associated with being sure of not seeing something (the Type 2 distributions are symmetric around $X = 0$).

The log-likelihood of being correct on any given trial (likelihood ration; $\beta$) is the log of the ratio of equations 7.1:

$$\beta = \log\left(\frac{f(x|C)}{f(x|I)}\right) \tag{7.2}$$

I assume that confidence increases as the log-likelihood of being correct on the Type 1 task also increases. As the likelihood ratio is symmetric around $c$, there are thus two values of $x$ for each possible value of $\beta$, one for when $x < c$ and one for when $x > c$. This corresponds to being confident that a signal was or was not present. I define $c \pm m$ (as the likelihood ratio is symmetric about $c$) as values of $x$ that satisfy equation 7.2 for a given value of $\beta$.

Using the signal detection categories of 3.2, it is possible to compute theoretical hit and false alarm rates for a range of values of $\beta$ by integrating over the Type 2 probability distributions specified in equations 7.1:

$$\begin{aligned} H &= \int_{-\infty}^{c-m} f(x|C)\,dx + \int_{c+m}^{\infty} f(x|C)\,dx \\ FA &= \int_{-\infty}^{c-m} f(x|I)\,dx + \int_{c+m}^{\infty} f(x|I)\,dx \end{aligned} \tag{7.3}$$

These integrals are plotted graphically for a single arbitrary value of $m$ in figure 7.1.

### 7.2.2 Simulation results

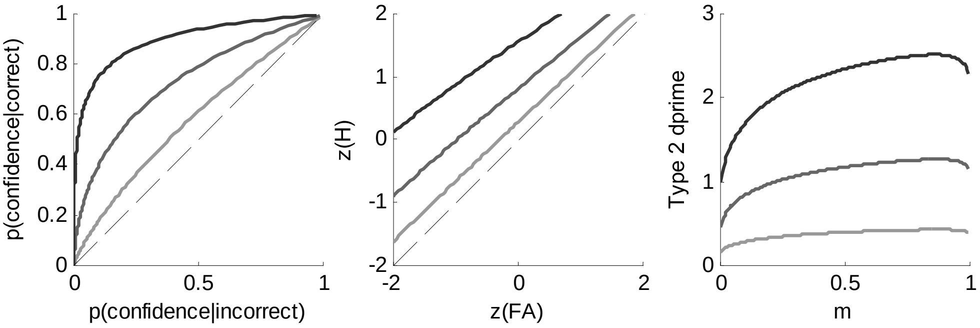

**Figure 7.2:** Computational simulations of Type 2 confidence functions for three values of Type 1 $d'$ (0.5 - light grey, 1.5 - medium grey and 3 - black). The left panel shows the predicted Type 2 ROC quantifying the ability to discriminate between correct and incorrect responses cumulated across levels of confidence. The middle panel shows the same ROC functions plotted in normal-normal ($z$) coordinates. The right panel shows the predicted Type 2 $d'$ $[z(H) - z(FA)]$ as a function of Type 2 criterion ($m$) and Type 1 $d'$. These functions would be flat if Type 2 sensitivity and bias were independent.

In figure 7.2 I simulate the Type 2 ROC for a range of values of Type 1 $d'$. Three qualitative results are of interest here. First, the area under the ROC ($A_{roc}$) is predicted to scale with Type 1 $d$ (leftmost panel). This prediction arises because Type 2 performance is generated from a linear transformation of Type 1 distributions; if there is greater signal for making the decision itself, there will also be greater signal for making a metacognitive judgment about this decision. In other words, metacognitive sensitivity should increase as performance increases (Kruger & Dunning, 1999).

Second, like Type 1 ROC functions, Type 2 ROCs are predicted to lie along a straight line in normal-normal (z) coordinates (middle panel), indicating that an equal-variance model is a good approximation of the relationship between Type 2 hits and false alarms (note that while the difference between Type 2 hits and false alarms is predicted to be normally distributed, it does not follow that the underlying distributions themselves follow normal distributions; indeed, examination of figure 7.1 indicates this is not the case). However, unlike Type 1 ROC functions, Type 2 ROCs are predicted to have a slope of less than unity on z coordinates, due to the 'long tails' of the $f(C|x)$ function. The form of this asymmetry predicts that there are categories of subjective confidence that are solely used when perception is veridical.

Finally, our simulations replicate a potentially counter-intuitive result obtained by Evans & Azzopardi (2007). Due to the sub-unity slope of the $z$-ROC, varying the Type 2 criterion ($m$) leads to an asymmetric effect on Type 2 hits and false alarms (rightmost panel in figure 7.2). This asymmetry predicts that as the Type 2 criterion becomes more conservative, $A_{roc}$ increases. In other words, the independence of sensitivity and bias assumed by SDT may not hold in Type 2 decision scenarios. This exact non-independence has been observed in empirical data (Evans & Azzopardi, 2007).

## 7.3 Comparison of Type 2 model with behaviour

### 7.3.1 Methods

#### 7.3.1.1 Participants

32 participants (15 males; aged 19 37 years; mean age 26.4 years) gave written informed consent to take part in the experiment. The study was approved by the local Research Ethics Committee.

#### 7.3.1.2 Stimuli

The perceptual decision display comprised six Gabor gratings (circular patches of smoothly varying light and dark bars) arranged around a central fixation point (figure 7.3). Each Gabor subtended 1.4 degrees of visual angle in diameter, and consisted of a luminance pattern modulated at a spatial frequency of 2.2 cycles per degree. Each 'baseline' Gabor had a contrast of 20% of maximum, and appeared at a mean eccentricity of 6.9 degrees. The fixation point comprised a black square measuring 0.2 degrees diameter, luminance 0.10 cd/m$^2$, with a central white square 0.1 degrees diameter, luminance 13.64 cd/m$^2$. The background was a uniform gray screen of luminance 3.66 cd/m$^2$.

Baseline Gabors were displayed with a contrast of 20% (where 0% is no difference between the luminance of the grating bars and 100% is maximum difference, i.e. black to white). The pop-out Gabors were drawn from a stimulus set in which contrast varied from 23 to 80% in increments of 3%. At the time of confidence ratings, the display consisted of a grey screen (luminance 3.66 cd/m$^2$) with the numbers 1 to 6 written left to right (luminance 13.64 cd/m$^2$, 0.7 degrees in height, centred around fixation).

Stimuli were presented on a gamma calibrated CRT display (Dell FP2001, 20.1 inch display; 800 × 600 pixels; 60 Hz refresh rate), at a viewing distance of approxi-

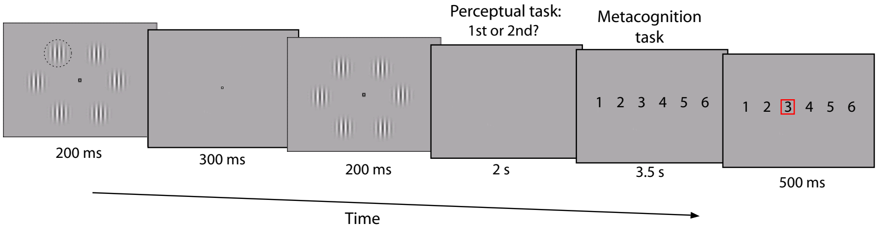

**Figure 7.3:** Subjects completed a two-alternative forced choice task that required two judgments per trial: a perceptual response followed by an estimate of relative confidence in their decision. The perceptual response indicated whether the first or second temporal interval contained the higher contrast (pop-out) Gabor patch (highlighted here with a dashed circle which was not present in the actual display), which could appear at any one of 6 locations around a central fixation point.

mately 60 cm, situated in a darkened room. Stimulus display and response collection were controlled by Matlab 7.8.0 (Mathworks Inc., Natick, MA, USA) using the COGENT 2000 toolbox (`http://www.vislab.ucl.ac.uk/cogent.php`).

#### 7.3.1.3 Task

The visual judgement comprised a temporal two-alternative forced choice pop-out task (see figure 7.3 for timings). All Gabors in one interval were of the same contrast, but in the other interval, one of the Gabors was of higher contrast than the others (the pop-out Gabor, illustrated by a dashed circle in figure 7.3 that was not present in the actual display). The temporal interval and spatial position of the pop-out Gabor varied randomly between trials. Participants were required to decide whether this pop-out Gabor had appeared in the first or the second interval. The perceptual judgement was indicated using the left hand with the numbers '1' (first interval) or '2' (second interval) on the QWERTY keypad of a standard PC keyboard. Participants then indicated their confidence in the perceptual decision they had just made on a scale of 1 (low relative confidence) to 6 (high relative confidence), using their right hand to press one of the numbers 1 to 6 on the numerical keypad. A square red frame (width 1 degree, thickness 0.1 degree) appeared around the selected rating (figure 7.3).

The contrast of the pop-out Gabor was chosen from the stimulus set of pop-out Gabors using a 1-up 2-down staircase procedure (Levitt, 1971) which, at the limit, results in convergence on 71% accuracy. The contrast of the pop-out Gabor at the end of each block was used as the starting contrast for the pop-out Gabor in the next block. Our aim in this staircase procedure was to equate objective perceptual performance across individuals, leaving quantification of metacognitive ability unconfounded by performance (Lau, 2010).

Participants were instructed to try to use the whole of the confidence scale in their responses, and to bear in mind that the scale represents relative confidence, as, given the difficult nature of the task, they would rarely be completely certain that their visual judgement had been correct. Participants performed a practice session to familiarise themselves with the stimuli and task. The main experiment consisted of 600 trials, split into 6 blocks of 100 trials. They were given no feedback about their performance until the end of the experiment.

#### 7.3.1.4 Type 2 signal detection analysis

Because the specific mathematical assumptions of conventional SDT may not hold for this new analysis (Evans & Azzopardi, 2007; Galvin *et al.*, 2003), I used nonparametric assessments of sensitivity and bias (Kornbrot, 2006). I constructed Type 2 ROC curves for each participant that characterised the probability of being correct for a given level of confidence. ROC curves were anchored at $[0,0]$ and $[1,1]$.

To plot the ROC, $h_i = p(\text{confidence} = i|\text{correct})$ and $f_i = p(\text{confidence} = i|\text{incorrect})$ were calculated for all $i$. These probabilities were then transformed into cumulative probabilities, and plotted against each other. Following Kornbrot, I computed distribution-free measures of sensitivity and bias from this ROC by dividing the area into two parts $K_B$ is the area between the ROC curve and the major diagonal to the right of the minor diagonal and $K_A$ is the area between the ROC curve and major diagonal to the left of the minor diagonal. From simple geometry (derived in the Appendix of Kornbrot 2006), these areas can be calculated as follows:

$$\begin{aligned} K_{\scriptscriptstyle A} &= \frac{1}{4}\sum_{k=1}^{k=3}[(h_{k+1}-f_k)^2-(h_k-f_{k+1})^2] \\ K_{\scriptscriptstyle B} &= \frac{1}{4}\sum_{k=4}^{k=6}[(h_{k+1}-f_k)^2-(h_k-f_{k+1})^2] \end{aligned} \tag{7.4}$$

Sensitivity ($A_{roc}$) is then the sum of these areas, and Type 2 bias ($B_{roc}$) is the log of the ratio:

$$\begin{aligned} A_{roc} &= K_{\scriptscriptstyle A}+K_{\scriptscriptstyle B} \\ B_{roc} &= \ln\left(\frac{K_{\scriptscriptstyle A}}{K_{\scriptscriptstyle B}}\right) \end{aligned} \tag{7.5}$$

Type 1 $d'$ and bias ($c$) were calculated in the standard manner (see Chapter 3) where $H = p(\text{response} = 1|\text{interval} = 1)$ and $FA = p(\text{response} = 1|\text{interval} = 2)$.

### 7.3.2 Results

#### 7.3.2.1 ROC model fits

I noted a practice effect in the staircase parameters (figure 7.4) reflected in a decrease in mean contrast and variability from block 1 to 2. A one-way ANOVA of mean contrast with block as a within-subjects factor revealed a significant effect of block ($F_{(5,155)} = 8.18, P < 0.001$) that was abolished on removal of block 1 ($F_{(4,124)} = 1.56, P = 0.19$). ROC analysis was therefore carried out on data from blocks 2 – 6 (indicated by the red box in figure 7.4), after stabilisation of psychophysical performance.

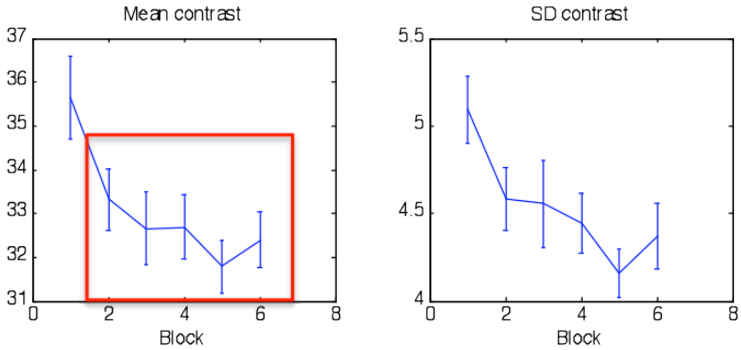

**Figure 7.4:** Mean and standard deviation (SD) of oddball Gabor contrast (percentage of maximum contrast) plotted for each block of the perceptual task, averaged over participants. Error bars represent one standard error of the mean. Because stimulus contrast and variability were significantly higher in block 1, indicating a period of gradual stabilisation of performance, only data from blocks 2-6 (indicated by the red surround) were used to calculate SDT measures.

To explore how well a Gaussian Type 2 SDT model accounted for the confidence rating data, I fit the following linear regression model:

$$z(h) = \beta_0 + \beta_1 . z(f) + \epsilon \tag{7.6}$$

where $z$ is the inverse of the cumulative normal distribution function. This simple linear model provided an excellent fit to the data (mean $R^2 = 0.98 \pm 0.016$ SD), indicating that the underlying $f(X|\text{correct})$ and $f(X|\text{incorrect})$ distributions are normal-like, lying along a straight line in $z$-coordinates. The $\beta_1$ parameter (slope) indicates the relative variance of the two distributions. This parameter was on average less than 1 within our sample ($0.87 \pm 0.09$ SD), indicating that the $f(X|\text{correct})$ distribution has greater variance than the $f(X|\text{incorrect})$. As can be seen from figure 7.2, this finding is in accordance with the direct-translation model, which also predicts slopes of less than 1 when the proportion of hits and false alarms are plotted against one another in $z$-coordinates.

#### 7.3.2.2 Individual differences in metacognitive ability

To test the prediction that $A_{roc}$ scales with Type 1 $d'$, I examined individual differences in metacognitive ability (figure 7.5). Through the staircase, I deliberately minimised variability in $d'$, thus isolating variability in $A_{roc}$ that might otherwise be obscured by significant covariation with Type 1 performance (see Lau *et al.* (2006) for a similar approach). I found considerable variation across individuals

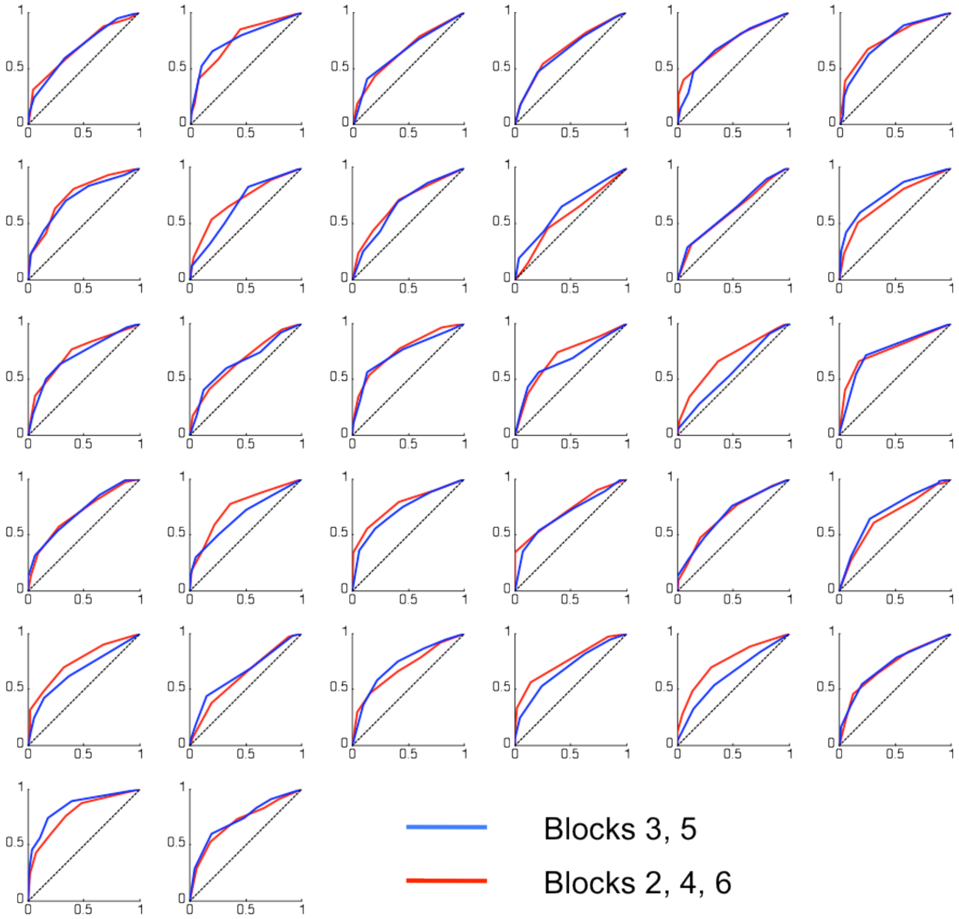

**Figure 7.5:** Individual ROC curves. Data is split into odd (blue; blocks 3 and 5) and even (red; blocks 2, 4 and 6) blocks.

in metacognitive ability ($A_{roc} = 0.55 - 0.75$) despite underlying task performance being held constant (proportion correct: 70 – 74%). Furthermore, both proportion correct (Pearsons $r = -0.21, P = 0.24$) and $d'$ (Pearsons $r = 0.08, P = 0.66$) were uncorrelated with $A_{roc}$ (figure 7.6). To establish whether this variability was stable, I split data from each participant into two halves, and computed the test-retest reliability of the two sets. This analysis revealed intrasubject consistency in $A_{roc}$ ($r = 0.69, P = 0.00001$; figure 7.5).

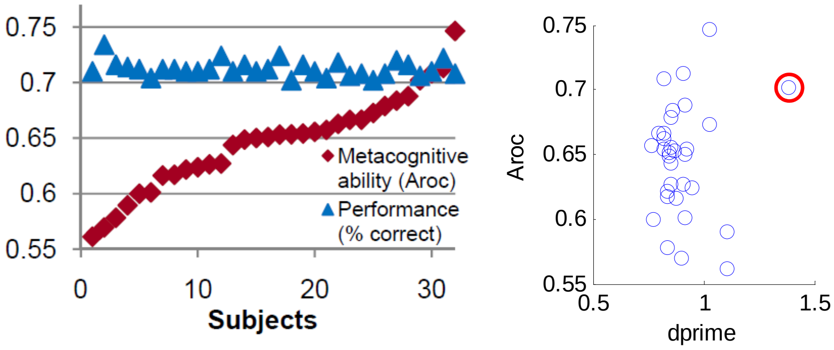

**Figure 7.6:** The left panel plots of the relationship between task performance (% correct) and $A_{roc}$, with subjects ordered by increasing $A_{roc}$. The right panel plots the relationship between $A_{roc}$ and Type 1 $d'$. The circled outlier ($d' > 3SD$ from group mean) was omitted from the analysis of brain structure reported in Chapter 8.

I next asked what might be driving these differences in $A_{roc}$. An increase in sensitivity can be driven either by a better sensitivity to correct decisions (hits vs. misses), incorrect decisions (false alarms vs. correct rejections) or a combination of the two. To examine this, I split participants into high and low $A_{roc}$ groups based on the median value, and collapsed across confidence ratings to generate a $2 \times 2$ factorial design crossing low (ratings 1-3) and high (ratings 4-6) confidence with decision accuracy. It can be seen that sensitivity in confidence ratings to incorrect decisions (false alarms and correct rejections) does not differ between groups (figure 7.7). However, the high $A_{roc}$ group demonstrate better monitoring of correct decisions; i.e. they are more likely to use higher than lower confidence ratings when they were actually correct on the Type 1 decision. This effect was reflected in a significant 3-way interaction between group, Type 1 outcome (correct/incorrect) and confidence ($F_{(1,30)} = 8.76, P = 0.003$). Follow-up one-sample t-tests of hit rate against 0.5 revealed that that this interaction was driven by significant sensitivity to correct decisions in the high $A_{roc}$ group ($t(15) = 2.26, P = 0.039$), but not the low $A_{roc}$ group ($t(15) = 0.30, P = 0.77$).

#### 7.3.2.3 Relationship between reaction times and metacognitive ability

The relationship between decision confidence and reaction time (RT) is complex (Baranski & Petrusic, 1998; Pleskac & Busemeyer, 2010), but the general finding is

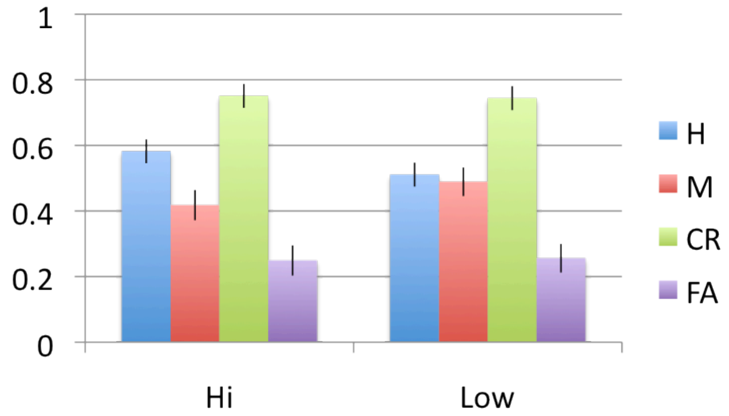

**Figure 7.7:** Relative proportions of responses categorised using Type 2 SDT, split according to whether individuals demonstrated high or low metacognitive ability.

that greater decision confidence is associated with faster Type 1 RTs (this relationship can reverse in situations where extended evidence accumulation is beneficial; see section 2.2.5). However, the relationship between individual differences in Type 2 sensitivity and RTs is largely unexplored. Here I examine RTs for both the Type 1 and Type 2 judgment as a function of confidence rating (figure 7.8), and link these relationships to changes in metacognitive ability.

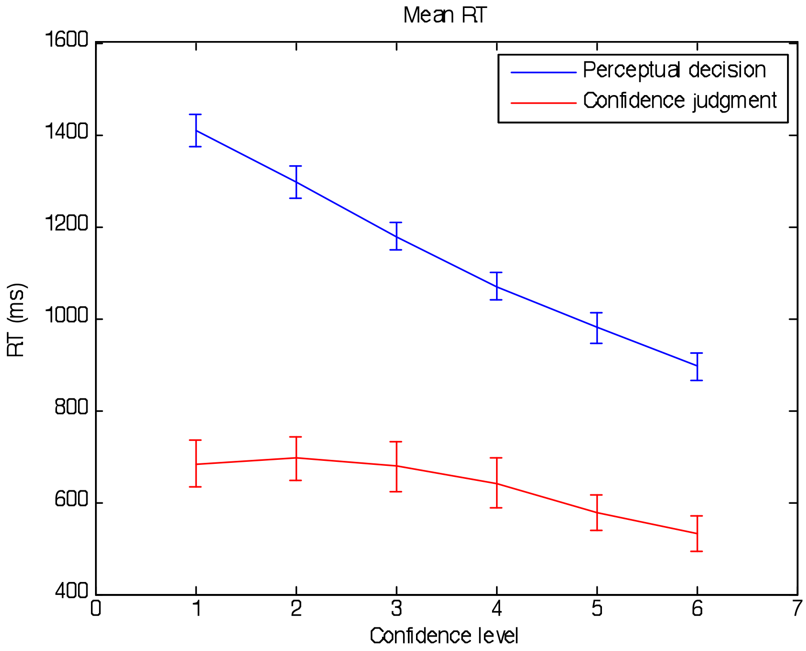

**Figure 7.8:** Mean RTs measured in milliseconds for both the perceptual decision (blue) and the confidence judgment (red) from blocks 2-6, plotted as a function of reported confidence level. Data are averaged across 32 participants and the error bars represent one standard error of the mean.

I entered each participant's RTs for both the initial task judgement into a multiple regression designed to predict reported confidence level. As expected based on previous work, faster RTs to the first perceptual judgement were significantly predictive of greater subjective confidence at a group level ($t(31) = -8.43, P < 0.001$; figure 7.9). The mean regression coefficient for RTs to the second metacognitive judgement was on average positive, but was not significant at the group level ($t(31) = 0.99, P = 0.33$). However, of interest was whether subjects RTs for the metacognitive judgment provided insight into the Type 2 decision process. I therefore examined the relationship between this Type 2 regression coefficient (beta) and $A_{roc}$. Subjects who showed an increasingly positive relationship between Type 2 RT and confidence also demonstrated greater $A_{roc}$, as indexed by a positive correlation between the relevant multiple regression beta and $A_{roc}$ ($r = 0.48, P = 0.006$; left panel of figure 7.9).

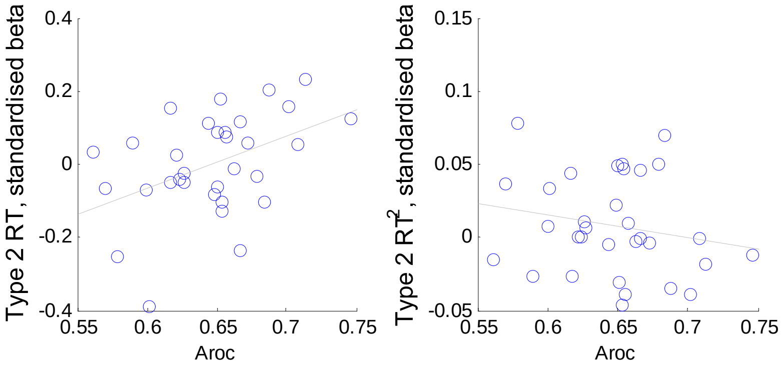

**Figure 7.9:** The left panel plots the significant positive relationship observed between the extent to which Type 2 RT predicted confidence (via multiple regression) and $A_{roc}$ across individuals (see text for further details). The right panel plots the same (non-significant) relationship for the quadratic term.

This result is somewhat difficult to interpret, as it reflects a significant linear effect at the group level on a linear effect (between RT and confidence) at the individual level. To gain a better understanding of what is driving this effect, I examined the relationship between Type 2 RTs and confidence separately for high and low $A_{roc}$ individuals (figure 7.10). A shift can be seen from an inverted-U pattern in high $A_{roc}$ individuals to a decreasing-linear pattern for low $A_{roc}$ individuals. In other words, our data suggest that subjects who have high $A_{roc}$ treat the metacognitive report as a second-order decision, with confidence ratings at the extremes of the scale (1 and 6) being made with greater speed than intermediate confidence ratings; in contrast, low $A_{roc}$ subjects show a pattern of RTs that is relatively indistinguishable from that seen for Type 1 RTs. Given this inverted-U seen for high $A_{roc}$ individuals, we might expect a quadratic term (Type 2 RT$^2$) to be a better predictor of $A_{roc}$ than a linear term; however this was not the case ($r = -0.20, P = 0.27$; right panel of figure 7.9). Further work is therefore required to disambiguate the complex relationship between RT and $A_{roc}$.

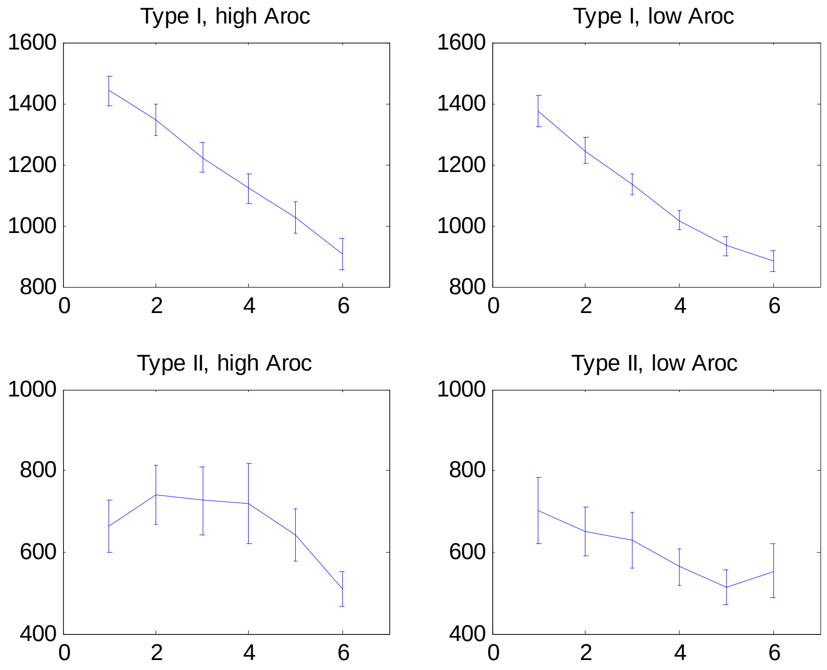

**Figure 7.10:** The upper row plots the relationship between Type 1 RT and reported confidence; the lower row plots the relationship between Type 2 RT and confidence. For both analyses participants are sorted into high $A_{roc}$ (left column) and low $A_{roc}$ (right column) based on a median split.

## 7.4 Discussion

In a Type 2 SDT model, probability distributions over the stimulus are assumed to directly give rise to Type 2 confidence, without invoking the notion of intermediate processing stages. Higham *et al.* (2009) identify this generation of confidence from assumed stimulus distributions as the 'direct translation hypothesis'. They empirically tested this hypothesis by varying Type 1 response criteria in a memory paradigm, noting that Type 2 sensitivity should be affected if direct translation holds, a conclusion supported by the data (Higham *et al.* 2009; Experiment 3). Building on this work, I show that the Type 2 model provides an accurate fit to post-decision confidence data in a perceptual task. In addition, the average ROC slope on normal-normal coordinates was found to be significantly less than 1, consistent with model predictions. However, a Type 2 SDT model predicts that perceptual performance should be yoked to metacognitive ability. Against this prediction, by constraining perceptual performance to be near-threshold (71%) using a staircase procedure, I show that there is considerable variation in the Type 2 ROC ($A_{roc}$) despite perceptual performance ($d'$) remaining constant across individuals. This result suggests that Type 2 performance may be partially dissociable from Type 1 performance.

Indeed, there are clear examples, such as blindsight and subliminal perception, where commentaries and performance dissociate (Lau *et al.*, 2006; Weiskrantz *et al.*, 1974; Wilimzig *et al.*, 2008). For example, Lau *et al.* (2006) showed that the subjective level of visibility of a stimulus could differ in two different metacontrast masking conditions, despite detection performance remaining identical. Similarly, recent human psychophysical data reveal dissociations between objective performance and subjective confidence, suggesting a more complex relationship between Type 1 and Type 2 processes (Wilimzig *et al.*, 2008). Metacognitive monitoring can also be manipulated by factors that are orthogonal to the task of interest: Bengtsson and colleagues found that priming subjects as clever or stupid altered the monitoring of errors, but not basic task performance, which was held constant through use of a staircase procedure (Bengtsson *et al.*, 2010); similarly, manipulating the ease of processing can affect metacognitive reports while leaving task performance relatively unaffected (Alter & Oppenheimer, 2009; Busey & Arici, 2009; Wenke *et al.*, 2010).

Reaction time measures suggest that decision confidence is at least partly determined by additional processing following the decision itself (Baranski & Petrusic, 1998, 2001). In support of this additional processing having functional relevance, I find that the extent to which post-decision reaction time is modulated by reported confidence is predictive of metacognitive ability. High $A_{roc}$ individuals show a qualitative pattern of slower RTs for intermediate confidence ratings. Similar to the pattern observed in Chapter 4, this result suggests that high $A_{roc}$ individuals treat the Type 2 rating as a second-order decision, with greatest uncertainty seen under equivocal evidence in favour of being correct or incorrect (equation 7.2). Similarly, Baranski & Petrusic (1998) found an inverted-U relationship between confidence and Type 2 RT when Type 1 decisions were made, as here, under time pressure. However, the addition of a quadratic term to the multiple regression analysis (which would be expected to capture the inverted-U effect) did not improve predictions of $A_{roc}$, indicating further work is needed to determine the relationship between RT and metacognitive ability.

Differences in the monitoring of correct, as opposed to incorrect, decisions appears to underlie individual differences in metacognitive ability in the present dataset. High $A_{roc}$ individuals tended to use a greater proportion of high than low confidence ratings when their decisions were correct; this sensitivity was largely absent in low $A_{roc}$ individuals. In contrast, metacognitive sensitivity to incorrect, error trials was approximately equal in the two groups. Similarly, transcranial magnetic stimulation (TMS) of dlPFC has been shown to decrease metacognitive sensitivity through a selective effect on correct, but not incorrect, trials (Rounis *et al.*, 2010). Together this data suggests the intriguing possibility that error monitoring (thought to depend on the ACC and insula; Ridderinkhof *et al.* 2004; Ullsperger *et al.* 2010; Yeung *et al.* 2004) is dissociable from metacognitive processes that track increasing confidence in being correct. In the next Chapter 8 I go on to examine how individual differences in metacognitive ability identified here map onto individual differences in brain structure, with a particular focus on the prefrontal cortex.

# Chapter 8

# Relating individual differences in metacognitive ability to human brain structure

> Don't you think if I were wrong I'd know it?
>
> Sheldon Cooper, *The Big Bang Theory*

## 8.1 Introduction

In the previous Chapter 7 I identified a component of metacognitive ability that is dissociable from perceptual task performance across individuals. In this chapter I assess whether this ability has a distinct neural correlate by examining individual differences in brain structure.

Little is known about the biological basis of metacognitive ability, here defined as how well an individual's confidence ratings discriminate correct from incorrect decisions over time. I hypothesised that individual differences in metacognitive ability would be reflected in the regional anatomy supporting this function, akin to similar associations between brain anatomy and performance noted in other cognitive domains (Carreiras *et al.*, 2009; Fuentemilla *et al.*, 2009; Scholz *et al.*, 2009; Tuch *et al.*, 2005).

Earlier patient studies describe candidate brain regions where damage is associated with poor introspective ability, in particular, a prefrontal-parietal network (Del Cul *et al.*, 2009; Fletcher & Henson, 2001; Simons *et al.*, 2010). Theories of prefrontal function have emphasised a role for anterior (rostrolateral) prefrontal cortex in carrying out second-order operations on internally generated information (Christoff & Gabrieli, 2000; Fletcher & Henson, 2001), a core feature of metacognition. Consistent with prefrontal cortex playing a causal role in metacognition, patients with lesions to anterior prefrontal cortex show deficits in subjective reports compared to controls, after factoring out differences in objective performance (Del Cul *et al.*, 2009). Furthermore, impairing dorsolateral prefrontal cortex function with theta-burst transcranial magnetic stimulation compromises the metacognitive sensitivity of subjective reports of awareness, while leaving underlying task performance intact (Rounis *et al.*, 2010). I hypothesise that the local structure of these regions (both grey matter volume and white matter integrity) might reflect an individual's metacognitive ability.

As described in Chapter 7, I quantified variability in metacognitive sensitivity (which is specific to an individual) that was independent of both objective task performance and subjective confidence (which vary on a trial-by-trial basis). Here, I now ask whether this variability in metacognitive ability was predicted by variability in brain structure using two distinct measures: grey matter (GM) volume measured using MRI, and the fractional anisotropy of white matter (WM) measured using diffusion tensor imaging (DTI).

## 8.2 Methods

### 8.2.1 Subjects

Details of participants are given in Chapter 7. One participant was excluded from the analysis of brain structure due to aberrant psychophysical task performance (Type 1 $d' > 3SD$ from the group mean; circled point in figure 7.6).

### 8.2.2 Voxel-based morphometry analysis

Voxel-based morphometry (VBM) provides a quantitative measure (at each voxel) of the tissue volume per unit volume of spatially normalised image (see Chapter 3). A 1.5T Sonata scanner (Siemens Medical Systems, Erlangen, Germany) was used to acquire all images for each participant. T1-weighted anatomical whole-brain scans were acquired for VBM analysis (176 slices, echo time = 3.56ms, TR = 12.24ms, voxel size 1mm isotropic). VBM preprocessing was carried out using SPM8 (`http://www.fil.ion.ucl.ac.uk/spm`). The images were first segmented into GM, WM and cerebral spinal fluid in native space (Ashburner & Friston, 2005). The GM segment images from this process were then rigidly aligned and subsequently warped to an iteratively improved template using nonlinear registration in DARTEL (Ashburner, 2007). DARTEL's 'Normalise to MNI' module was then used to produce smoothed normalised images. The DARTEL template was affine registered to MNI space, and the GM images were transformed using the DARTEL flow-fields and this affine transformation, in a way that preserved their local tissue volumes (equivalent to a Jacobian 'modulation' step). Smoothing used a Gaussian kernel of 8mm full width at half maximum.

The pre-processed GM images were entered into a multiple regression model in SPM8 to determine which brain regions showed significant covariation with the SDT-based measures of metacognitive ability. I included $A_{roc}$, $d'$, Type 2 criterion ($B_{roc}$; the overall tendency to use high confidence responses), the absolute (unsigned) value of the Type I criterion ($|c|$) and gender ($M = 1$; $F = 0$) in the model. Type 1 criterion ($c$) measures the extent of the bias towards interval 1 or 2 on the perceptual decision task, with greater bias reflecting suboptimal performance. Positive values indicate bias towards interval 1, and negative values bias towards interval 2. I thus entered the absolute value of $c$ as a covariate of no interest, with higher values indicating suboptimal performance bias towards either interval.

Adjustment for 'global' brain volume using proportional scaling was applied, resulting in voxel values that were proportions of total GM volume. A binary mask (SPM8 grey.nii template $> 0.3$) was used to restrict the search volume to changes in GM. $T$-statistic maps reflecting the correlation between each regressor and regional GM volume were created. Cluster-based statistics were used to locate significant regions based on both their peak value and spatial extent after applying an initial cluster-defining threshold of $P < 0.001$. Due to structural images displaying local variation in smoothness, standard applications of cluster-based random field theory are inappropriate (Hayasaka *et al.*, 2004). I thus applied non-stationary cluster extent correction when calculating family-wise error (FWE) corrected $P$ values using the NS toolbox (`http://www.fmri.wfubmc.edu/cms/NS-General`). Computational simulations (Hayasaka *et al.*, 2004) show that for designs with high degrees of freedom and sufficient smoothness, as here, using a cluster defining threshold of $P < 0.001$ with correction for non-stationarity provides adequate control over the family-wise false positive rate ($P < 0.05$).

### 8.2.3 Diffusion tensor imaging analysis

The DTI dataset comprised of 68 images with 60 slices and 2.3mm isotropic resolution. The first 7 images were collected with $b = 100$ s/mm$^2$. The diffusion encoding directions were isotropically distributed on the surface of the sphere (Jansons & Alexander, 2003) for the remaining 61 images and the $b$-value was 1000 s/mm$^2$. The echo time was 90ms, each 2D image slice took 150ms to collect, and the field of view was 220mm. DTI data sets are often collected using echo-planar imaging (EPI) methods which are affected by susceptibility-induced artefacts. To reduce the extent of these artefacts two datasets were collected for each participant, with the only difference being that the phase encoding direction was reversed for the second run. This method ensures the susceptibility-induced distortions are equal and opposite in the two datasets, providing the opportunity to correct their effect (Andersson *et al.*, 2003).

Diffusion-weighted images were first aligned using FSL's `eddycorrect` (`http:`

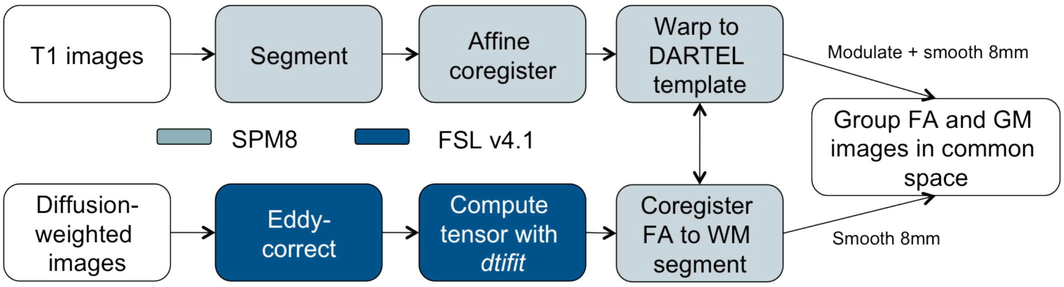
T1 images
Segment
Affine coregister
Warp to DARTEL template
Modulate + smooth 8mm
Group FA and GM images in common space
SPM8
FSL v4.1
Diffusion-weighted images
Eddy-correct
Compute tensor with *dtifit*
Coregister FA to WM segment
Smooth 8mm

**Figure 8.1:** Image preprocessing pipeline for VBM and DTI analysis

`//www.fmrib.ox.ac.uk/fsl/`), and then combined into a single dataset with reduced susceptibility-induced artefacts. The main diffusion tensor was then fitted at each voxel using FSL's `dtifit`. From the tensor a rotationally invariant measure of diffusion anisotropy can be calculated. One such measure is fractional anisotropy (FA) with values ranging from 0 (representing isotropic, or undirected, diffusion) to 1 (representing a single preferred direction of diffusion). This measure has been used extensively to investigate local WM integrity, as diffusion of water molecules is more restricted perpendicular to, rather than along, neuronal fibres (see section 3.5). The calculated FA map for each participant (in native space) was imported into SPM8 and coregistered to the WM segment image of the same participant created during VBM analysis. Coregistration was carried out by maximising normalised mutual information between the images. The DARTEL flowfields and affine (MNI) transformation were then applied to each participant's coregistered FA image, producing normalised FA images in MNI space. Unlike the VBM normalisation (which preserved the original local tissue volume), the FA images were normalised in a way that preserved their original voxel values (without 'modulation'). Normalised FA images were smoothed with the same 8mm full-width at half maximum Gaussian kernel prior to statistical analysis. For one participant, DTI scans were unavailable, leaving 30 subjects in the FA analysis.

Statistical analysis of FA proceeded in an identical fashion to that of GM volume (see section 8.2.2). A multiple regression model was constructed consisting of $A_{roc}$, $d'$, Type 2 criterion ($B_{roc}$; the overall tendency to use high confidence responses) and the absolute (unsigned) value of the Type I criterion ($|c|$). A binary mask (mean normalised FA $> 0.2$) was used to restrict the search volume to changes in WM. Statistical inference was conducted as for VBM. Probable tract labels were obtained using the JHU White-Matter Tractography Atlas within FSL.

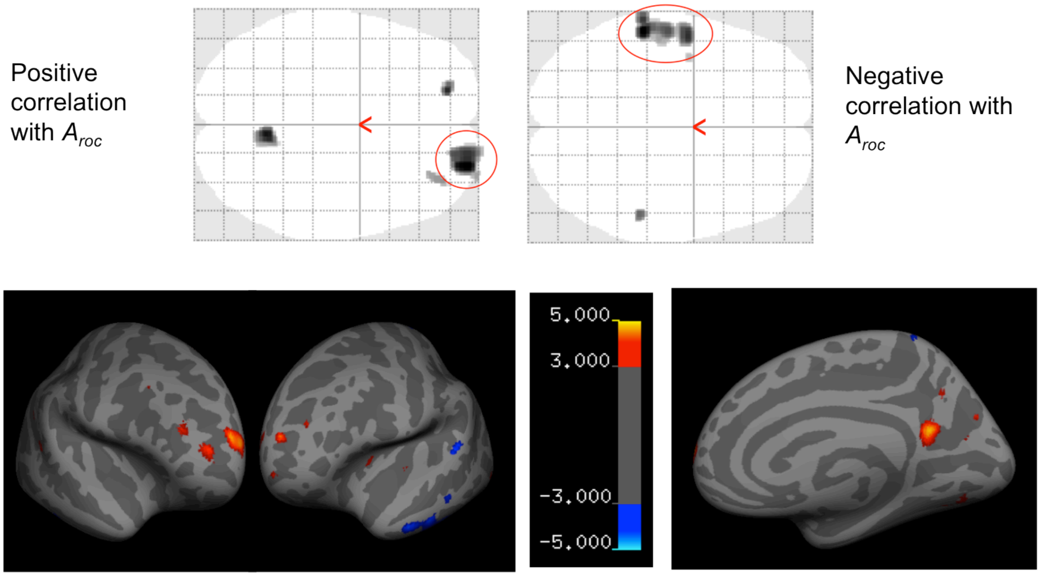
Positive correlation with $A_{roc}$
Negative correlation with $A_{roc}$
5.000
3.000
-3.000
-5.000

**Figure 8.2:** Grey matter volume correlated with metacognitive ability. Upper row: Axial 'glass brains' (viewed from above) showing areas where grey matter volume correlates positively (left) and negatively (right) with $A_{roc}$. Regions circled are cluster-level corrected for multiple comparisons across the whole brain volume. Lower row: Projection of statistical ($T$) maps for positive (hot colormap) and negative (cool colormap) correlations with $A_{roc}$ onto an inflated cortical surface T1-weighted template, thresholded at $T > 3$ for display purposes.

## 8.3 Results

### 8.3.1 Grey matter

I examined for a relationship between brain structure and four different psychological measures: the metacognitive ability ($A_{roc}$) of our participants, objective performance on the perceptual task ($d'$ and $c$), and the tendency to use high or low confidence responses on individual trials ($B_{roc}$). Having removed the potentially confounding factors (Smith *et al.*, 2007) of overall brain size and gender (by entering them as regressors of no interest), I found that an individual's metacognitive ability ($A_{roc}$) was significantly correlated with grey matter volume in right rostrolateral prefrontal cortex (figure 8.2) (Brodmann area (BA) 10, peak voxel coordinates: $[24, 65, 18], T_{max} = 4.8, P < 0.05$, corrected for multiple comparisons). Notably, gray matter volume in this region did not correlate with task performance as indexed by $d'$ ($r = 0.15, P = 0.42$) or overall confidence ($B_{roc}$; $r = -0.023, P = 0.90$). Gray matter volume in a homologous region in left rostrolateral prefrontal cortex was also correlated with $A_{roc}$, but did not survive correction for multiple comparisons across the entire brain volume. Details of this and other clusters that did not survive a whole brain correction are listed in table 8.1.

Thus, variability in introspective judgements of performance on a simple visual detection task was predicted by variability in the anatomical structure of anterior prefrontal cortex (BA 10) independently of both objective performance and level

| Regressor | Voxels | Z-score | cluster $P$-value | Peak voxel | Laterality | Label |
|---|---|---|---|---|---|---|
| $A_{roc}$ | **675** | **4.02** | **0.029** | **24,65,18** | **R** | **BA 10** |
| | 291 | 3.93 | 0.191 | 6,-57,18 | L/R | Precuneus |
| | 31 | 3.78 | 0.703 | -20,-53,12 | L | BA10 |
| | 25 | 3.45 | 0.829 | 36,39,21 | R | BA46 |
| | 29 | 3.44 | 0.497 | 35,50,9 | R | BA46 |
| $A_{roc}$ | **713** | **3.92** | **0.026** | **-56,-30,-26** | **L** | **ITG** |
| | 76 | 3.62 | 0.753 | -63,-30,10 | L | STG |
| | 80 | 3.54 | 0.457 | 51,-33,-21 | R | ITG |
| | 15 | 3.22 | 0.995 | -41,3,-48 | L | ITG |
| $B_{roc}$ | 28 | 3.93 | 0.313 | -33,-73,34 | L | BA19 |
| $-B_{roc}$ | 93 | 3.47 | 0.233 | -59,-27,-14 | L | MTG |
| | 20 | 3.35 | 0.826 | -66,-10,3 | L | STG |
| $d'$ | 82 | 3.77 | 0.175 | -3,-84,-21 | L/R | Cerebellum |
| | 16 | 3.68 | 0.939 | 53,-25,-15 | R | MTG |
| | 389 | 3.66 | 0.112 | 60,-39,51 | R | Sup. parietal |
| | 47 | 3.45 | 0.817 | 6,-61,4 | L/R | Lingual gy. |
| | 18 | 3.26 | 0.953 | -3,-9,66 | L | BA6 |
| $-d'$ | ... | ... | ... | ... | ... | ... |

**Table 8.1:** Grey matter volume associated with SDT parameters across subjects. Whole-brain corrected clusters ($P < 0.05$, corrected for multiple comparisons) are indicated in bold type. For completeness, correlations that survive a height threshold of $P < 0.001$, uncorrected, and an extent threshold of 10 voxels are also reported.

of confidence. While my primary question addressed positive dependence of gray matter on $A_{roc}$, I also found that left inferior temporal gyrus showed a negative correlation with metacognitive sensitivity (figure 8.2) (coordinates: $[-56-30-26], T_{max} = 4.66, P < 0.05$, corrected for multiple comparisons), accompanied by a similar region on the right that did not survive correction for multiple comparisons (see table 8.1 for full details and coordinates).

### 8.3.2 White matter

I next analysed white matter microstructure based on the following considerations. If the structure of anterior prefrontal cortex is functionally related to metacognitive performance, then one would expect that white matter tracts connected with this region would also show a similar microstructural correlation with expression of this behavioural trait. Statistical analysis of white matter fractional anisotropy (FA) proceeded in an identical fashion to that of GM volume. In a whole-brain analysis of white matter microstructure, I found that FA in the genu of the corpus callosum was positively dependent on $A_{roc}$ (figure 8.3) ($P < 0.05$, corrected for multiple comparisons). Neither objective performance (stimulus contrast or $d'$) nor overall confidence ($B_{roc}$) correlated with grey matter volume or white matter fractional

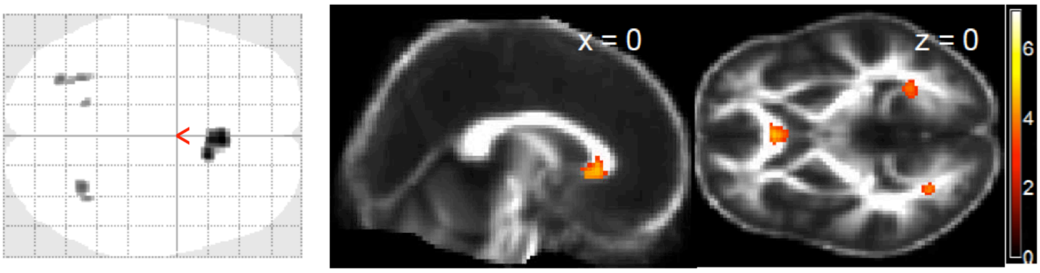

**Figure 8.3:** The leftmost panel shows an axial glass brain indicating areas where white matter fractional anisotropy (FA) correlates positively with $A_{roc}$. No suprathreshold FA clusters were found for negative correalations with $A_{roc}$ (see also tables 8.1 and 8.2). The right panel shows a statistical ($T$) map of voxel-wise correlations between fractional anisotropy (FA) and $A_{roc}$, thresholded at $T > 3$ for display purposes and overlaid on sagittal (left) and axial (right) slices of the average FA image across subjects, at the $x$ and $z$ co-ordinates indicated. A region within the genu of the anterior corpus callosum showed a correlation between FA and metacognitive accuracy that was statistically significant after correcting for multiple comparisons ($P < 0.05$).

anisotropy elsewhere in the brain ($P > 0.05$, corrected for multiple comparisons (see table 8.2 for uncorrected correlations). I note that an absence of structural correlations with these parameters may have been due to our design deliberately minimising variability in both $d'$ and $B_{roc}$ in order to isolate the neural correlates of introspective ability ($A_{roc}$).

### 8.3.3 Control analyses of grey and white matter correlations

I carried out additional analysis to rule out potential alternative interpretations of the findings. One concern is that variation in underlying perceptual acuity could confound the anatomical variance ascribed to metacognitive ability ($A_{roc}$). Good perceptual ability may be reflected in low mean stimulus contrast and/or low staircase variability (though I note that extraneous environmental or ocular factors also affect these variables). To rule out this interpretation, I computed the partial correlation between brain structure and $A_{roc}$ while controlling for both mean stimulus contrast and the variability (SD) in the staircase required to achieve a constant level of performance within each individual. Both the GM cluster in BA10 ($r = 0.39, P = 0.036$) and the FA cluster in anterior corpus callosum ($r = 0.74, P < 0.001$) remained significantly correlated with $A_{roc}$ after controlling for mean contrast and staircase variability.

This partial correlation analysis only examines the correlation of predefined regions. As a further test, I constructed a second design matrix in which mean stimulus contrast and staircase variability were directly entered as predictors of GM/FA, with gender again present as a covariate of no interest. Neither measure correlated with grey matter or FA at the statistical thresholds used in the main analysis ($P > 0.05$, corrected for multiple comparisons), even when applying a mask (8mm sphere) to

| Regressor | Voxels | Z-score | cluster $P$-value | Peak voxel | Laterality | Label |
|---|---|---|---|---|---|---|
| $A_{roc}$ | **308** | **3.93** | **0.033** | **2 26 -2** | **L/R** | **Corpus callosum** |
| | 66 | 3.58 | 0.492 | 29,-55,-2 | R | Post. corpus callosum |
| | 31 | 3.54 | 0.502 | -32,-67,0 | L | Inf. fronto-occipital fasciculus |
| | 26 | 3.48 | 0.680 | -32,-55,14 | L | Longitudinal fasciculus |
| | 13 | 3.44 | 0.824 | 35,-52,-15 | R | Inferior longitudinal fasciculus |
| | 11 | 3.39 | 0.631 | -18,-52,28 | L | Cingulum |
| $-A_{roc}$ | ... | ... | ... | ... | ... | ... |
| $B_{roc}$ | ... | ... | ... | ... | ... | ... |
| $-B_{roc}$ | 128 | 4.24 | 0.226 | -8,20,-9 | L | Corpus callosum |
| | 49 | 3.91 | 0.132 | 26,-51,-9 | R | Post. corona radiata |
| | 21 | 3.54 | 0.438 | -18,29,24 | L | Cingulum |
| $d'$ | 74 | 3.89 | 0.251 | -17,6,39 | L | Sup. corona radiata |
| | 18 | 3.33 | 0.721 | -18,-7,45 | L | Sup. corona radiata |
| $-d'$ | ... | ... | ... | ... | ... | ... |

**Table 8.2:** White matter fractional anisotropy associated with SDT parameters across subjects. Whole-brain corrected clusters ($P < 0.05$, corrected for multiple comparisons) are indicated in bold type. For completeness, correlations that survive a height threshold of $P < 0.001$, uncorrected, and an extent threshold of 10 voxels are also reported.

| Analysis | Regressor | Voxels | Z-score | cluster $P$-value | Peak voxel | Laterality | Label |
|---|---|---|---|---|---|---|---|
| GM | Negative mean contrast | 88 | 3.67 | 0.917 | 14,-10,24 | R | Caudate |
| | | 51 | 3.56 | 0.783 | -65,-57,4 | L | MTG |
| | | 78 | 3.44 | 0.913 | 5,-76,21 | L/R | Calcarine sulcus |
| | | 80 | 3.38 | 0.908 | 3,36,42 | L/R | dmPFC |
| | | 11 | 3.21 | 0.982 | -14,29,-20 | L | OFC |
| | Negative SD | 128 | 4.00 | 0.301 | 59,-42,1 | R | MTG |
| | | 29 | 3.70 | 0.577 | -51,33,36 | L | Inf. parietal |
| | | 34 | 3.36 | 0.938 | 47,-15,-48 | R | Postcentral gy. |
| | | 22 | 3.24 | 0.953 | -44,-21,46 | L | Postcentral gy. |
| FA | Negative mean contrast | ... | ... | ... | ... | ... | ... |
| | Negative SD | 30 | 3.86 | 0.942 | -29,-15,48 | L | Sup. corona radiata |

**Table 8.3:** GM and FA correlating with negative stimulus contrast and staircase variability (low-level measures of perceptual performance). After correcting for multiple comparisons, no significant clusters were observed, but correlations that survive a height threshold $P < 0.001$, uncorrected, and an extent threshold of 10 voxels are reported for completeness.

isolate voxels within the vicinity of the BA10 (GM) or the anterior corpus callosum (FA) peak voxels. While I am cautious about interpreting uncorrected findings, one result of potential interest is that GM volume in the medial calcarine sulcus, consistent with the location of early visual cortex, showed increased volume in subjects with greater perceptual acuity as defined by negative mean stimulus contrast ($P < 0.001$, uncorrected). Table 8.3 details uncorrected results from these models for completeness. Together these control analyses indicate that the correlations I observe between $A_{roc}$ and structure relate to differences in metacognitive ability rather than low-level differences in performance.

## 8.4 Discussion

The central finding in this chapter is a delineation of a focal anatomical substrate that predicts inter-individual variability in metacognitive ability. As with any correlational method, I cannot establish whether the covariation I observed here between brain structure and metacognition reflects a causal role. However, given a wealth of evidence for changes in grey matter volume within and between individuals associated with a range of skills, these data indicate that underlying differences in metacognitive ability may be similarly dependent on local brain anatomy.

How might these regions contribute to metacognition? Anterior subdivisions of prefrontal cortex are implicated in high-level control of cognition (Boorman *et al.*, 2009; Burgess *et al.*, 2007; Christoff & Gabrieli, 2000; Daw *et al.*, 2006; Koechlin & Hyafil, 2007; Ramnani *et al.*, 2004), and are well placed to integrate supramodal perceptual information with decision output (Ramnani *et al.*, 2004), a process thought to be key for metacognition (Cleeremans *et al.*, 2007; Insabato *et al.*, 2010). The genu of the corpus callosum contains white matter fibres connected with the anterior and orbital prefrontal cortices in humans (Park *et al.*, 2008), consistent with metacognitive ability being dependent not only on anterior prefrontal grey matter but also on reciprocal projections to and from this area. I also found negative correlations with $A_{roc}$ in bilateral regions of anterior inferior temporal grey matter. While I am cautious about interpreting the relevance of a decrease in grey matter volume for increased metacognitive ability, I note these temporopolar regions are implicated in both self-related (Frith & Frith, 2003) and higher-order visual (Gross & Schonen, 1992) processing, and thus alterations in grey matter here might similarly place functional constraints on perceptual metacognition. Our present findings may reflect innate differences in anatomy, or alternatively reflect the effects of experience and learning, as has been found in the sensorimotor domain (Draganski *et al.*, 2004; Scholz *et al.*, 2009).

Dorsolateral prefrontal activity increases under conditions where subjective reports match objective perceptual performance (Lau *et al.*, 2006), suggesting a computational role in linking performance to confidence. Furthermore, convergent evidence from studies employing TMS (Rounis *et al.*, 2010), lesion (Del Cul *et al.*, 2009) and functional MRI (Yokoyama *et al.*, 2010) approaches provide support for the role of rostrolateral PFC as a mediator of metacognitive ability that is dissociable from primary task performance. Given that the variation in metacognitive ability in the present study was primarily driven by an effect on monitoring of correct trials, anterior PFC may be particularly important for monitoring subjective confidence, as opposed to errors (see also Rounis *et al.* 2010). However, based on these initial studies, the anatomical specificity with which metacognition can be related to the lateral PFC is still imprecise; for example, some studies have emphasised dorsolateral PFC (BA 9/46; Rounis *et al.* 2010; Lau *et al.* 2006), whereas the present study and others (Del Cul *et al.*, 2009; Yokoyama *et al.*, 2010) emphasise the importance of rostrolateral BA10. Further studies, ideally employing functional MRI, are required to precisely localise the activity relevant for metacognitive function.

In summary, these data provide an initial window onto the biological basis of the ability to link objective performance to subjective confidence. The demonstration that this ability may be dependent on a restricted area of prefrontal cortex, an area that is phylogenetically recent, is intriguingly consistent with a conjecture that metacognitive function has been selected for during evolution (Metcalfe, 2008), facilitating computations that allow us to introspect about self-performance, or, more prosaically, think about thinking.

# Chapter 9

# General discussion

## 9.1 Overview

The work contained in this thesis is concerned with analysing the neural basis of simple visual decisions at two levels: that of (first-order) links between perception and action, and second-order, or metacognitive, commentaries on these links. By comparing these levels of analysis, I set out to form bridges between mechanism on the one hand and subjectivity on the other. Below I discuss contributions, limitations, and future work relevant to each of the three sections of my thesis – a neuroimaging investigation of how the loss function is incorporated into the stages of perceptual decision making, experiments examining how the basal ganglia modulate the link between decision and action, and theoretical and experimental work on how metacognitive commentaries about decision-making are generated. In the final section I consider how metacognition may be related to particular higher-order aspects of consciousness, before concluding with a glance towards how I envisage the evolution of the future of the field.

## 9.2 Integration of the loss function in perceptual decision

### 9.2.1 Contributions

How the brain incorporates the loss function information into the perceptual decision process is unknown (see section 2.3.5). A Bayesian ideal observer should incorporate both a likelihood and a prior term when coming to a posterior belief about a particular state of the world (Kersten *et al.*, 2004); this belief is then incorporated with the knowledge of the expected value of each potential option to compute the optimal course of action (Kording, 2007). However, whether this sequence of computations is reflected in the brain is unknown; indeed, on one proposal the distinction between values and priors is lost (Friston, 2009). Thus while integration of values and priors into perceptual decision-making is well-studied at the behavioural level (e.g. Liston & Stone 2008; Johnstone & Alsop 2000; Proshansky & Murphy 1942; Whiteley & Sahani 2008), whether this integration affects neural mechanisms of perception, decision and/or action is unclear.

In Chapter 4, I devised a task that could address this question using brain imaging. Subjects were asked to categorise noisy images as either faces or houses, and were informed about the monetary losses for making an incorrect response on each category. On two-thirds of trials, these losses were asymmetric, such that faces were associated with higher losses than houses, or *vice versa*. This manipulation systematically biased subjects' decision-making towards the category associated with lower losses, consistent with the adoption of an asymmetric decision criterion. By using face and house stimuli, categories with dissociable perceptual representations in ventral visual cortex, I asked whether this behavioural bias was associated with shifts in activity at the level of category coding. I found no evidence for this; instead, ventral visual activity tracked the perceptual input, correlating with the amount of phase in the image informative of the face or house category. In contrast, during asymmetric loss trials, activation increased in left inferior frontal gyrus (L-IFG), posterior parietal cortex and the basal ganglia. This increase in activity was not associated with a change in performance ($d'$) or reaction time, suggesting that it reflects incorporation of the asymmetric loss function at a decision stage (see also Ferrera *et al.* 2009; Rorie *et al.* 2010). By design, response hand was orthogonal to the cost factor, allowing examination of effector-specific effects of cost, but none were found. Together these findings are consistent with asymmetries in value being incorporated at a decision stage rather than impacting on perception (*viz.* the computation of a posterior belief).

Since publication of results from this chapter (Fleming *et al.*, 2010c), Summerfield and colleagues have reported similar findings in a visual detection paradigm (Summerfield & Koechlin, 2010). Using an evidence accumulation model of subjects' reaction times, they found that asymmetric value shifted the baseline of accumulation, consistent with similar analyses of monkey behavioural data (Feng *et al.*, 2009; Rorie *et al.*, 2010). Shifts in decision criteria were tracked by the parietal cortex, but did not affect activity in regions encoding sensory evidence, including fusiform and middle occipital gyri. Signal in parietal cortex increased as decision criteria became more liberal in their yes-no task. In contrast, I observed signal increases in parietal and inferior frontal cortex for *any* shift in criterion away from the neutral point, a difference that may relate to the critical task features – yes/no and two-alternative forced-choice – used in the two studies. However, the broad pattern of results – effects on decision-making regions but not visual regions – is consistent with the findings reported in Chapter 4.

For the bias towards the house category, I additionally noted that subjects who showed greater shifts in decision criteria also had greater activity in ACC. This finding is echoed in a recent study demonstrating that event-related activity in the ACC predicted an individual's ability to use predictive (prior) information to adjust a decision bound (Domenech & Dreher, 2010), suggesting this activity may generalise to different causes of criterion shifts. The interactive role of ACC and more dorsomedial activity found to correlate with decision difficulty in both Chapter 4 and other studies (e.g. Grinband *et al.* 2006; Noppeney *et al.* 2010; Summerfield & Koechlin 2010; Thielscher & Pessoa 2007) remains to be determined. Activity correlating with decision difficulty was not affected by asymmetries in value in Chapter 4, suggesting one function of this region may be to recruit lateral frontal cortex to implement changes in decision criteria when uncertainty is high (Kouneiher *et al.*, 2009). However, here difficulty was correlated with reaction time, which was not included as a covariate in the model in this study (cf. Chapters 5 and 6). As Noppeney and colleagues point out, 'response times covary with many other cognitive processes that are unrelated to multisensory evidence accumulation such as stimulus processing times, working memory demands, etc.' (Noppeney *et al.*, 2010), and thus the existence of a putative difficulty signal in perceptual decision-making remains equivocal (Heekeren *et al.*, 2008).

### 9.2.2 Future directions

A recent theoretical suggestion holds that cost functions and priors perform a common role in Bayesian inference (Friston, 2009; Friston *et al.*, 2009). In other words, cost is implemented 'as if' it were a prior expectation (and vice versa). I additionally observed an increase in activity in the parahippocampal cortex to increases in value for the house category, outside the stimulus-sensitive region of interest. This activation may reflect top-down activation of category-specific cortex (Ekstrom *et al.*, 2008; Philiastides & Sajda, 2007) potentially consistent with changes in expectation (Summerfield & Koechlin, 2008; Egner *et al.*, 2010), but the lack of a similar effect for the face category is difficult to fit into this schema (although see section 4.5.2). In contrast, several studies have reported robust effects of prior expectations on activity in face- and house-specific ventral visual cortex (Summerfield *et al.* 2006a; Egner *et al.* 2010; Puri *et al.* 2009; but see Wenzlaff *et al.* 2011). Further work is required to explicitly examine the neural overlap between cost- and prior-induced biases in visual decision-making, ideally through a factorial design, and ask whether these biases may be incorporated at dissociable stages of the decision pathway.

The task in Chapter 4 required a forced choice between two categories, making it difficult to draw inferences as to whether subjects' perceptual experience was affected by the asymmetry in loss function. Liston & Stone (2008) tackled this question by asking observers to make a post-decision judgment during a spatial saccade task. Observers were asked to saccade to the brighter of two targets (the 'motor' response), which could appear to the left or right of fixation, and then make a 'perceptual' response using a keypress as to whether the chosen stimulus was brighter or dimmer than a subsequently presented test stimulus. They found that a decision bias induced by asymmetries in reward or probability associated with one side of the screen also affected the perceptual judgment, making the targets seem brighter and noisier. Other data suggest prior expectations alter the awareness of particular stimulus attributes such as brightness and motion direction (Carrasco *et al.*, 2004; Sterzer *et al.*, 2008). One key variable related to whether value affects perceptual processing may be the extent to which asymmetry in costs co-opt visual attention to induce biases towards one or other region of space (Maunsell, 2004; Serences, 2008; Weil *et al.*, 2010); in the present work and in Summerfield & Koechlin (2010), value was associated with a particular category, rather than a region of space, perhaps restricting the influence of the loss function to higher-level decision areas.

To the extent that the loss function is integrated at a post-perceptual stage, this raises intriguing questions as to subjects' knowledge of the antecedents of their decision process under biased and unbiased conditions. In other words, do subjects 'know' that they are biased on a particular trial? If we asked for a rating of the *stimulus*, rather than a reward-maximising response, would a different answer be given? If priors and values do act at perceptual and post-perceptual stages of the decision process respectively, one strong prediction from this work is that the influence of the former should be 'cognitively impenetrable' (Pylyshyn, 2003), whereas the latter should be amenable to self-report. A separate analysis of the influence of value and priors was not carried out by Liston & Stone (2008), but these conditions may plausibly have differential effects on any subsequent 'perceptual' estimation. In other words, subjects' confidence in their response to a trial biased by an asymmetric prior should be indistinguishable from an unbiased trial; in contrast, confidence on a trial biased through asymmetric value may show tell-tale signs of a mismatch between perception and action. This hypothesis remains to be tested, and could usefully draw upon the methods for measuring metacognitive confidence outlined in section 2.4.1 and Chapter 7.

## 9.3 Role of the basal ganglia in linking decision and action

### 9.3.1 Contributions

A hallmark of human decision-making is flexibility. Performing a predictable sequence of actions such as driving is usually carried out smoothly and swiftly, without the need for cognitive control. An unexpected change in the environment – a cat running into the road, say – wrenches us into immediate evasive action. This situation results in response competition – between the ongoing process of driving, and the need to hit the brakes to avoid the cat. The neural mechanisms underlying switches between automatic and controlled modes of action selection are unknown, but the subthalamic nucleus (STN) and pre-SMA have been proposed as key nodes in this process (Isoda & Hikosaka, 2007, 2008; Neubert *et al.*, 2010; Redgrave *et al.*, 2010). Through multiplexing of various cortical and intra-basal ganglia inputs (Shepherd, 2004), the STN is uniquely placed to integrate cognitive and motor information in the service of flexible action control (Frank, 2006; van den Wildenberg *et al.*, 2010). Further, deep-brain stimulation of the structure during treatment of Parkinsons disease can have the unwanted side effect of impairing cognitive control (Alberts *et al.*, 2008; Hershey *et al.*, 2004; Frank *et al.*, 2007; Ballanger *et al.*, 2009). However, direct evidence for the role of STN in action reprogramming in humans has remained elusive.

In Chapters 5 and 6 I provide evidence that the BOLD signal in human STN is increased during switches away from a default action. In Chapter 5 the default was signalled by the computer, and accepting the default was associated with inaction, making it difficult to interpret whether STN activity was specific to non-default 'go' responses, or non-default actions in general. In Chapter 6, I manipulated the default through changing response frequency for both 'go' and 'nogo' responses, permitting dissociation of the required action from changes in expectation. Here, STN activity was seen to be modulated by violation of expectations both for 'go' and 'nogo' responses, supporting a role for this structure in the control of unexpected actions. One influential account of these effects is that the need for cognitive control activates hyperdirect afferents to increase STN activity, allowing more time for the correct decision to be made, analogous to a raised decision threshold (Frank, 2006; van den Wildenberg *et al.*, 2010). A related set of studies has reported that striatal activation is linked to changes in the profile of reaction times, consistent with a *lowered* decision threshold (Forstmann *et al.*, 2008, 2010). In Chapter 6 the striatum was found to be marginally more active for low difficulty 'go' responses, potentially consistent with this view. However, here a change in threshold was transient (on surprising trials), whereas in the Forstmann *et al.* studies the speed-accuracy tradeoff was altered in a tonic fashion, from block to block. It would be useful to explicitly compare the role of STN and striatum under blockwise speed-accuracy tradeoff instructions using the anatomically-targeted techniques employed in Chapter 6.

The default bias is pervasive in several everyday scenarios (Thaler & Sunstein, 2009). For example, altering whether organ donation requires an opt-out or opt-in response (changing the default) produces a dramatic change in the proportion of a country's population agreeing to donate (Johnson & Goldstein, 2003). The studies in this thesis are considerably simplified in comparison to this category of higher-level default effect, which involves myriad factors including loss aversion and transaction costs (DeMartino *et al.*, 2009; Tversky & Kahneman, 1991). However, the study reported in Chapter 5 (Fleming *et al.*, 2010b), along with a related recent study (Yu *et al.*, 2010), represent initial attempts to isolate the neural correlates of accepting or rejecting a default option. Yu *et al.* (2010) manipulated the default in an gambling task, showing that insula and striatum were associated with switching away from, or sticking with, the default, respectively. In this experiment, participants were presented with two cards that when flipped could result in either a win or a loss, one of which was designated as the default. The discrepancies between the imaging results in these two studies might be ascribed to the focus on perceptual difficulty in the former case, and emotionally-laden processes in the latter. Consistent with this interpretation, when explicit feedback is added to the perceptual task used in Chapter 5, modulation of the default bias is associated with error-related insula activity (Nicolle *et al.*, 2011). Thus insula and striatal activity may modulate the strength of the default bias, with action control circuitry (including STN) being important in overcoming this bias. Further studies might usefully compare and contrast the role of emotional and visuomotor neural circuitry in modulating the default bias.

### 9.3.2 Future directions

One important distinction between studies that have studied perceptual 'prediction errors' or violations of expectation (Strange *et al.*, 2005; Näätänen *et al.*, 1987; Mars *et al.*, 2008; Egner *et al.*, 2010) and the study reported in Chapter 6 is that the former focussed on the sensory-mnemonic neural coding of expectation violation, whereas the present work examines how violations of expectation are linked to changes in action control. This link was also investigated by Bestmann *et al.* (2008), who found that surprising events were associated with slowing of RTs and a decrease in the excitability of the cortico-spinal motor tract using TMS, consistent with decreased drive to the motor system and a raised decision threshold (see section 2.2.5). In a similar analysis using paired-pulse TMS, Neubert *et al.* (2010) found that the right inferior frontal cortex and pre-SMA had inhibitory and excitatory effects on motor-evoked potentials during switch trials, respectively. This study also found that individual variability in the long-latency switch-related effect was associated with white-matter fractional anisotropy from right IFC to STN. In contrast, the short-latency effect size was associated with cortico-cortical white-matter fractional anisotropy, suggesting two potential routes by which premotor regions can influence motor output (see also Mars *et al.* 2009). Together, these studies suggest that cortical regions detecting violations in expectation may influence action control through a pathway that includes the STN. The inferior frontal cortex may be a key structure mediating this link, given its dual role in attention and action control (Dodds *et al.*, 2010; Hampshire *et al.*, 2007; Verbruggen *et al.*, 2010).

I observed an interaction between default-related activity with stimulus difficulty in Chapter 5 but not 6. As discussed in section 6.4.2, differences in task design and the demands placed on cognitive control may potentially explain this discrepancy. More generally, the role of STN in responding to changes in task difficulty remains an open question. Deep-brain stimulation studies have reported that STN DBS leads to impairments on high-difficulty decision-making (Baunez *et al.*, 2001; Frank *et al.*, 2007), but results in this field are not always consistent (cf. van den Wildenberg *et al.* 2006). Indeed, detailed analysis of reaction times show a complex interactive effect of DBS, with a detrimental effect on the accuracy of fast responses, but an improvement for slower responses (Wylie *et al.*, 2009). Crucially, the relationship between BOLD signal, neural activity and the effects of deep-brain stimulation is likely to be complex (Gradinaru *et al.*, 2009); this relationship urgently needs to be established if the results from DBS studies and fMRI studies are to be combined to better understand STN function.

In Chapter 6, I assumed that participants' expectations of a high or low $p$(go) block were present on the first trial. An alternative approach is to use ideal observer models to predict the expectations encoded by the subject at any particular point in time (Strange *et al.*, 2005; Mars *et al.*, 2008; Bestmann *et al.*, 2008). In the present study such a model makes very similar predictions for subjective expectation to those assumed by our factorial design, but potentially provides a different perspective on cognitive control (Mars *et al.*, 2010). In Bayesian approaches to decision-making (Ma, 2010), our difficulty factor is neatly equated with the precision of the likelihood, whereas changes in expectation are reflected in the prior (equation 2.2). A further additional term that was not investigated here is entropy. Entropy reflects the uncertainty over the prior – how sure am I that my expectation is correct (Behrens *et al.*, 2007; Bestmann *et al.*, 2008; MacKay, 2003)? In future studies it would be useful to separate expectation from entropy to assess whether, in information theoretic terms, the STN response I observe here reflects a change in expectation (surprise), or change in the precision of one's expectation.

## 9.4 Metacognitive decision-making

### 9.4.1 Contributions

Since Nisbett & Wilson (1977) argued that we 'tell more than we can know' when asked for verbal explanations of our cognitive processes, several studies have shown that introspective access to the antecedents of our decision process is weak and inaccurate (Johansson *et al.*, 2005; Wegner, 2003). However, such studies contrive to artificially induce mismatches between intention and outcome that do not occur in the real world with any great frequency (Johansson *et al.*, 2005). In contrast, other work has shown that subjects make surprisingly accurate metacognitive judgments of their performance in simple decision tasks (Graziano & Sigman, 2009; Marti *et al.*, 2010). What mechanisms underlie these judgments, and why might they go awry in certain scenarios?

One approach to answering this question is to build upon what is known about the underlying mechanisms of simple decisions (section 2.3). There is usually a tied relationship between performance and metacognition in real-world scenarios: if I know that the answer is 'Ashgabat', then I will also be more likely to know I know the answer[1] (Kruger & Dunning, 1999). This tied relationship between performance and metacognition may confound several studies of higher-order awareness (Lau, 2008). By establishing how confidence tracks local fluctuations in whether a decision was correctly or incorrectly executed, while still adjusting for overall differences in performance between conditions using a staircase procedure, this relationship can be broken, enabling an effective isolation of metacognitive capacity (Lau & Passingham, 2006). The method outlined in Chapter 7 combined dynamic adjustment of performance with Type 2 signal detection theory to isolate individual differences in the awareness of decision performance. Individual differences in metacognitive ability were found that could not be explained by changes in decision performance. In a recent study using a variant of the paradigm outlined in Chapter 7, Song *et al.* (2011) found that while performance in two different perceptual tasks (contrast detection and orientation discrimination) was uncorrelated within individuals ($r = 0.05, P = 0.83$), metacognitive ability remained stable ($r = 0.71, P < 0.001$). Together these findings suggest that a 'direct translation' account of metacognitive confidence is incomplete (Galvin *et al.*, 2003; Higham *et al.*, 2009). Instead, recent theoretical models proposing a partially separable second-order stage of decision-making can more naturally accommodate this dissociation between performance and metacognition (Pasquali *et al.*, 2010; Insabato *et al.*, 2010; Pleskac & Busemeyer, 2010).

In Chapter 8 I report data linking metacognitive ability to increased grey matter volume in anterior prefrontal cortex, and increased fractional anisotropy in white matter tracts that may project to anterior prefrontal regions. In line with a role for anterior prefrontal cortex in metacognitive monitoring, Del Cul *et al.* (2009) found that patients with lesions to this region had a deficit in subjective reports of stimulus visibility, despite being performance-matched with controls. Since the publication of the results in Chapter 8 (Fleming *et al.*, 2010b), a functional imaging study has reported a link between metacognitive ability (the link between performance and confidence) and BOLD signal in right anterior prefrontal cortex, BA10 (Yokoyama *et al.*, 2010). However, the precise role of this activity remains to be determined: one hypothesis (see section 9.4.2) is that this region integrates various sources of uncertainty (perceptual, motor) in a higher-order frame of reference suitable for

[1] The question, of course, being 'What is the capital of Turkmenistan?'.

communication to others. Furthermore, the role of BA10 in comparison to more posterior BA46, a region also implicated in metacognitive decision-making (Lau & Passingham, 2006; Rounis *et al.*, 2010; Middlebrooks & Sommer, 2010), remains to be determined. In addition, measures of IQ or working memory ability were not available for the sample tested in Chapter 7. It would be useful to relate metacognitive ability to higher cognitive function in general, especially given that damage to anterior prefrontal regions leads to a loss of fluid intelligence (Woolgar *et al.*, 2010), and ask to what extent it represents a modular function.

The accurate fit of the Type 2 SDT model suggests that confidence-in-accuracy fluctuations are probabilistic, reflecting graded 'evidence' that a decision was correct or incorrect. A similar model of awareness of errors in a simple decision task was recently proposed based on graded fluctuations in the error-related ERP (Steinhauser & Yeung, 2010). The individual variability in metacognitive ability reported in Chapter 7 was primarily driven by differences in the efficacy of monitoring correct decisions, rather than signalling errors. Similarly, Rounis *et al.* (2010) found that TMS to dorsolateral PFC impaired monitoring of correct but not incorrect decisions. Error awareness has been linked to insula and ACC activity (Steinhauser & Yeung, 2010; Ullsperger *et al.*, 2010), whereas metacognitive confidence is associated with anterior and lateral prefrontal cortex (Chapter 8 and Chua *et al.* 2009; Rounis *et al.* 2010; Yokoyama *et al.* 2010). To the extent that this neural dissociation reflects separable systems, one might predict that decision confidence and error awareness are dissociable both within and between individuals.

### 9.4.2 Future directions

As indicated in the previous section, an important next step is to carry out functional studies that attempt to link 'first-order' decision-making mechanisms to activity, perhaps in rostrolateral prefrontal cortex, that carries out second-order computations on this information, and relate this process to individual differences. The studies of Middlebrooks & Sommer (2010); Rounis *et al.* (2010); Steinhauser & Yeung (2010); and Yokoyama *et al.* (2010) represent significant first steps towards achieving this goal. One important consideration is the method used to elicit subjective beliefs about decision-making (section 2.4.1). It was argued in Chapter 7 that given an arbitrary graded scale that can be mapped to changes in Type 2 criteria, metacognitive ability can be recovered from the relationship between hits and false alarms. However, this method assumes that subjects are motivated to use the scale appropriately. Thus, adopting a graded but naturally incentivised method such as post-decision wagering (Persaud *et al.*, 2007) or quadratic scoring (Holt & Smith, 2009) may be preferable. If this approach is taken, it will be important to carefully address concerns about loss aversion and the perhaps indirect mapping between confidence and wagering (Clifford *et al.*, 2008; Fleming & Dolan, 2010; Schurger &

Sher, 2008).

To the extent that decision confidence reflects uncertainty in either sensory or motor processing, techniques used to manipulate and separate these variables may prove useful in unpacking the components of metacognitive confidence (Faisal & Wolpert, 2009; Trommershauser *et al.*, 2003). In addition, Bayesian models that naturally specify how uncertainty is represented and used to adjust behaviour (Behrens *et al.*, 2007; Daunizeau *et al.*, 2010; Yoshida & Ishii, 2006) form a useful point of contact between work on decision-making on the one hand, and studies of metacognition and consciousness on the other. These models assume that uncertainty over particular states adjust how much one learns in a particular environment. If the environment is volatile, such that a subject is uncertain about his or her belief about any one point in time, recent information is given more weight (Kalman, 1960). How metacognitive commentaries may relate to such 'ideal' estimates of uncertainty is unknown. Moreover, metacognitive commentaries about *performance* require access to both belief and response uncertainty (figure 9.1). For example, just after hitting a shot in tennis, you might have high confidence (low uncertainty) that the spot you chose to aim at is out of reach of your opponent (your belief), but your confidence in correctly executing the shot (your response) might be low. This distinction is often overlooked in the literature, where the term 'decision confidence' is applied to confidence about beliefs, responses or both. I believe it is crucial to parse metacognitive confidence into its constituents to understand its neural basis, and how metacognitive mechanisms drive changes in behaviour.

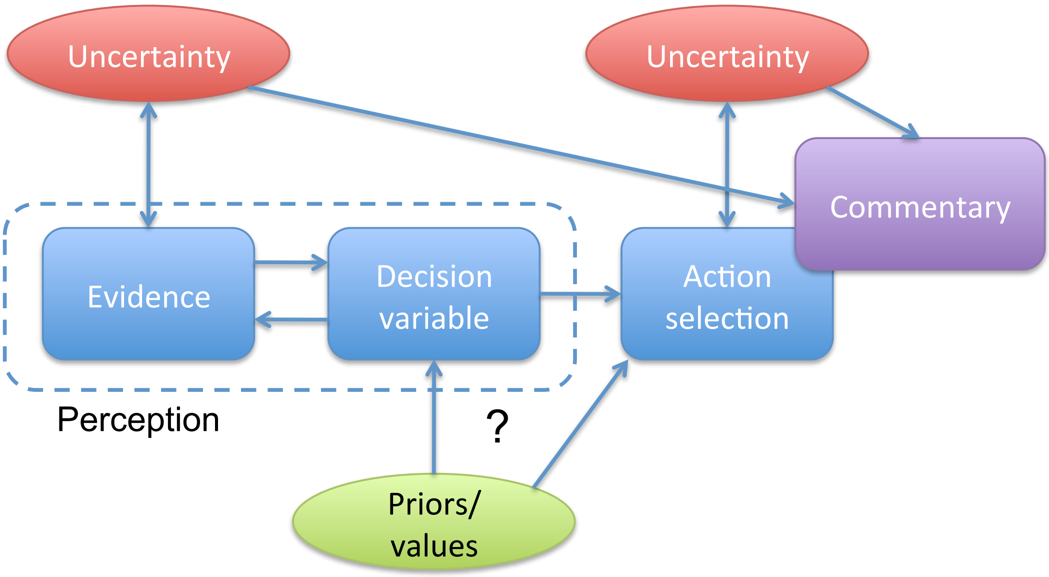

**Figure 9.1:** A conceptual model of the causal antecedents of metacognitive commentaries during decision-making, adapted from figure 2.3. Dissociable uncertainties about the different stages of the decision process (perception and action) are proposed to be integrated in a higher-order frame of reference, and be available for verbal report. That this commentary is itself a form of action selection is indicated by the overlap between these processes at the right of the figure.

For commentaries to integrate both belief and response uncertainty, the frame of reference in which decision variables are coded may be crucial. For instance, if information is maintained in segregated sensorimotor loops (Haber, 2003), a decision could be made and initiated without this information being more generally available for e.g. verbal report. It is currently an open question as to the extent to which perceptual decisions rely on 'embodied' or domain-general circuitry (Freedman & Assad, 2011). There are initial lines of evidence supporting a central role for the prefrontal cortex in representing sensory evidence in a more abstract frame of reference. Heekeren *et al.* (2006) found using fMRI that that a network of left posterior dlPFC, cingulate cortex, left IPS and left fusiform/parahippocampal gyrus responded to changes in sensory evidence independent of response modality. Similar results have been found for insular cortex (Ho *et al.*, 2009; Tosoni *et al.*, 2008), which is intriguing in relation to the possible role of this structure in awareness of errors. In value-based decision-making, a recent study has demonstrated a separation in the orbitofrontal cortex of regions showing response-independent and response-dependent value coding (Wunderlich *et al.*, 2009).

An important unanswered question is whether regions implicated in response-independent coding are also involved in metacognitive decision-making. In addition, if information is maintained in a response-independent frame of reference, the extent to which the loss function affects this response-independent belief state will affect whether first-order biases are expressed at the level of metacognitive reports. Usually, the information in first-order and metacognitive circuits will be highly correlated, due to the tied relationship between performance and metacognition highlighted above. However, in situations in which the low-level decision is affected by (motoric) biases induced by asymmetric priors or rewards (Stanford *et al.*, 2010), they may be decoupled. It is an empirical question as to whether putative response-dependent and response-independent processes will respond in a similar or dissimilar manner to biasing influences such as aymmetric priors, one which maps onto questions of whether bias alters subjective (post-decision) report, or only affects the first-order decision itself (see section 9.2.2).

### 9.4.3 Linking metacognition to consciousness

Several of the studies of metacognitive function that have been reviewed both in the present chapter and Chapter 2 have been couched in terms of visual awareness (Del Cul *et al.*, 2009; Lau & Passingham, 2006; Persaud *et al.*, 2007; Rounis *et al.*, 2010). This is perhaps not surprising; 'in an important sense, consciousness is knowing *that* you know *while* you know' (Metzinger, 2010). For a science of this higher-order aspect of consciousness, it is crucial to recognise that knowing that you know is usually confounded by changes in performance (Lau, 2008). By clamping performance at a constant level, either within or between subjects, one can in principle study higher-order aspects of consciousness in relative isolation.

Different constructs of consciousness map differently onto first- and second-order decision processes. For example, first-order confidence in the brightness of a stimulus might be associated with increased access (Block, 2005) to the phenomenology of brightness. This ability to make a first-order judgment might, somewhat counter-intuitively, relate to a higher-order thought in Rosenthal's theory of consciousness, in that the higher-order thought enables access to the quality of brightness (see Rosenthal 2005 for further discussion). In contrast, second-order confidence in a decision process is perhaps more akin to self-consciousness (Snodgrass *et al.*, 2009), or access to ones own mental states that arises due to monitoring of those states (Lycan, 1995). Thus metacognitive content may be quite separate from first-order access to consciousness (Seth, 2008). To what extent this theoretical separation occurs in the brain remains an empirical question.

I predict that the coding of beliefs in a response-independent framework is crucial for self-awareness. In other words, introspective ability requires a partial separation of perception and action. By maintaining information in a response-independent frame of reference, one breaks the boundaries of embodied sensorimotor loops, perhaps endowing the organism with a *perspective*. As the late Susan Hurley has written (Hurley, 2002):

> Having a perspective means in part what you experience and perceive depends systematically on what you do, as well as vice versa, and that you can keep track of some of the ways in which this is so, even if not in conceptual terms. In this sense having a perspective involves self-consciousness.

Such coding would permit multiple response systems to access the same information, including the motoric machinery underlying verbal reports. This proposal shares much in common with the idea that information becomes conscious by virtue of its presence in a 'global workspace' (Baars, 1993; Dehaene *et al.*, 2003), but grounds the workspace in circuits involved in first-order decision-making.

Finally, it will be important to ask whether metacognitive functionality confers an evolutionary advantage (Metcalfe, 2008). Why might response-independent circuitry have evolved? One hint can be drawn from the previous quotation. To keep track of errors in behaviour may require access both to one's belief about what should have happened, and a degree of certainty that the right response has been made (figure 9.1; although see Rosenthal 2008). Similarly, flexibility in being able to act upon information with any one of multiple effectors or cognitive processes may be increased by coding of information in a higher-order frame of reference. On this view, conscious information is exactly that information that is made available to multiple cognitive processes at a given point in time (Baars, 1993). This higher-order coding may in addition facilitate communication of mental states to others via verbal reports, allowing use of this sharing of confidence to mutually improve collective performance (Bahrami *et al.*, 2010). Whether this latter 'function' of consciousness is a by-product, or primary driver, of higher-order thought is an open question.

## 9.5 Conclusions

> She raised one hand and flexed its fingers and wondered, as she had sometimes before, how this thing, this machine for gripping, this fleshy spider on the end of her arm, came to be hers, entirely at her command. Or did it have some little life of its own? She bent her finger and straightened it. The mystery was in the instant before it moved, the dividing moment between not moving and moving, when her intention took effect. It was like a wave breaking. If only she could find herself at the crest, she thought, she might find the secret of herself, that part of her that was really in charge.
>
> *Atonement*, Ian McEwan

The idea that the human mind is composed of mechanisms that can be decomposed into stages, or interacting parts, has occupied generations of thinkers. In particular, the beguiling idea that a 'self' somehow sits between our perceptual apparatus and a motor system that acts as a wellspring of voluntary action, is both pervasive and intuitive (Cisek, 1999). Children are known to perceive the world and others in it in a naturally dualistic fashion (Bloom, 2004). Modern neuroscience and experimental psychology instead tell us that how things seem to us is often a poor guide to reality. Instead, ever-more complex and large-scale theories of the brain are emphasising the dynamic nature of competition for control of behaviour, and myriad interactions between expectations and perception.

However, what we, as social agents, know and are able to communicate about this process should not be ignored. By mapping self-reports onto the neural architectures that are being described for other types of decision, processes underlying self-awareness can be reframed in mechanistic terms. By maintaining a pivotal viewpoint both downwards, to a reductionist approach to perception and action, and upwards, to theories of higher-order thought and consciousness, we may be able to connect these two levels of analysis. This thesis has barely scratched the surface in this regard. Doing so will not only allow identification and characterisation of circuitry specific to higher-order aspects of human cognition, but may also refine our concept of the human mind and its mechanistic basis.

# List of Abbreviations

| Abbreviation | Details |
|---|---|
| ACC | Anterior cingulate cortex |
| ANOVA | Analysis of variance |
| BOLD | Blood-oxygen level-dependent |
| DBS | Deep-brain stimulation |
| dlPFC | Dorsolateral prefrontal cortex |
| dmPFC | Dorsomedial prefrontal cortex |
| DTI | Diffusion-tensor imaging |
| DV | Decision variable |
| EEG | Electroencephalography |
| EPI | Echo-planar imaging |
| FA | Fractional anisotropy |
| FEF | Frontal eye fields |
| FFA | Fusiform face area |
| FG | Fusiform gyrus |
| fMRI | Functional magnetic resonance imaging |
| FOK | Feeling of knowing |
| FWE | Family-wise error |
| GABA | Gamma-aminobutyric acid |
| GM | Grey matter |
| GPe | Globus pallidus externa |
| GPi | Globus pallidus interna |
| gy. | gyrus |
| HRF | Haemodynamic response function |
| IFC | Inferior frontal cortex |
| IFG | Inferior frontal gyrus |
| | |
| inf. | inferior |
| IPS | Intraparietal sulcus |
| ITG | Inferior temporal gyrus |
| LIP | Lateral intraparietal area |
| MFG | Middle frontal gyrus |
| MNI | Montreal Neurological Institute |
| MOG | Middle occipital Gyrus |
| MTG | Middle temporal gyrus |
| OFC | Orbitofrontal cortex |
| p. | pars |
| PDW | Post-decision wagering |
| Pe | Error-related positivity |
| PFC | Prefrontal cortex |

| **Abbreviation** | **Details** |
|---|---|
| PHG | Parahippocampal gyrus |
| post. | posterior |
| | |
| PPA | Parahippocampal place area |
| pre-SMA | pre-supplementary motor area |
| QSR | Quadratic scoring rule |
| RDK | Random dot kinematogram |
| ROC | Receiver operating characteristic |
| RT | Reaction time |
| SAT | Speed-accuracy tradeoff |
| SDT | Signal detection theory |
| SEM | Standard error of the mean |
| SMA | Supplementary motor area |
| SNc | Substantia nigra pars compacta |
| SNr | Substantia nigra pars reticulata |
| SOG | Superior occipital gyrus |
| SPL | Superior parietal lobule |
| SPM | Statistical parametric mapping |
| SSRT | Stop-signal reaction time |
| STG | Superior temporal gyrus |
| STN | Subthalamic nucleus |
| sup. | superior |
| SVC | Small volume correction |
| TE | Echo time |
| TMS | Transcranial magnetic stimulation |
| TR | Repetition time |
| VBM | Voxel-based morphometry |
| vlPFC | Ventrolateral prefrontal cortex |
| vmPFC | Ventromedial prefrontal cortex |
| VTA | Ventral tegmental area |
| VTF | Vibrotactile frequency |
| WM | White matter |

# Bibliography

Adler, R.J. (2009). *The Geometry of Random Fields*. SIAM.

Afraz, S., Kiani, R. & Esteky, H. (2006). Microstimulation of inferotemporal cortex influences face categorization. *Nature*, **442**, 692–695.

Alberts, J.L., Voelcker-Rehage, C., Hallahan, K., Vitek, M., Bamzai, R. & Vitek, J.L. (2008). Bilateral subthalamic stimulation impairs cognitive-motor performance in parkinson's disease patients. *Brain*, **131**, 3348–3360.

Alexander, A.L., Lee, J.E., Lazar, M. & Field, A.S. (2007). Diffusion tensor imaging of the brain. *Neurotherapeutics*, **4**, 316–329.

Alexander, G.E. & Crutcher, M.D. (1990). Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends in Neurosciences*, **13**, 266–271.

Alsop, B. & Davison, M. (1991). Effects of varying stimulus disparity and the reinforcer ratio in concurrent-schedule and signal-detection procedures. *Journal of the Experimental Analysis of Behavior*, **56**, 67–80.

Alsop, B. & Porritt, M. (2006). Discriminability and sensitivity to reinforcer magnitude in a detection task. *Journal of the Experimental Analysis of Behavior*, **85**, 41–56.

Alter, A.L. & Oppenheimer, D.M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, **13**, 219–235.

Amador, X.F. & David, A.S. (2004). *Insight and psychosis: awareness of illness in schizophrenia and related disorders*. Oxford University Press.

Amodio, D.M. & Frith, C.D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, **7**, 268–277.

Anderson, C.J. (2003). The psychology of doing nothing: forms of decision avoidance result from reason and emotion. *Psychological Bulletin*, **129**, 139–167.

Andersson, J., Hutton, C., Ashburner, J., Turner, R. & Friston, K. (2001). Modeling geometric deformations in EPI time series. *NeuroImage*, **13**, 903–919.

Andersson, J.L.R., Skare, S. & Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *NeuroImage*, **20**, 870–888.

Aron, A.R. (2010). From reactive to proactive and selective control: Developing a richer model for stopping inappropriate responses. *Biological Psychiatry*, Epub ahead of print.

Aron, A.R. & Poldrack, R.A. (2006). Cortical and subcortical contributions to stop signal response inhibition: role of the subthalamic nucleus. *Journal of Neuroscience*, **26**, 2424–2433.

Aron, A.R. & Verbruggen, F. (2008). Stop the presses: dissociating a selective from a global mechanism for stopping. *Psychological Science*, **19**, 1146–1153.

Aron, A.R., Schlaghecken, F., Fletcher, P.C., Bullmore, E.T., Eimer, M., Barker, R., Sahakian, B.J. & Robbins, T.W. (2003). Inhibition of subliminally primed responses is mediated by the caudate and thalamus: evidence from functional MRI and huntington's disease. *Brain*, **126**, 713–723.

Aron, A.R., Behrens, T.E., Smith, S., Frank, M.J. & Poldrack, R.A. (2007). Triangulating a cognitive control network using diffusion-weighted magnetic resonance imaging (MRI) and functional MRI. *Journal of Neuroscience*, **27**, 3743–3752.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, **38**, 95–113.

Ashburner, J. (2009). Computational anatomy with the SPM software. *Magnetic Resonance Imaging*, **27**, 1163–1174.

Ashburner, J. & Friston, K.J. (2000). Voxel-based morphometry–the methods. *NeuroImage*, **11**, 805–821.

Ashburner, J. & Friston, K.J. (2001). Why voxel-based morphometry should be used. *NeuroImage*, **14**, 1238–1243.

Ashburner, J. & Friston, K.J. (2005). Unified segmentation. *NeuroImage*, **26**, 839–851.

Ashkan, K., Blomstedt, P., Zrinzo, L., Tisch, S., Yousry, T., Limousin-Dowsey, P. & Hariz, M.I. (2007). Variability of the subthalamic nucleus: The case for direct MRI guided targeting. *British Journal of Neurosurgery*, **21**, 197–200.

Baars, B.J. (1993). *A Cognitive Theory of Consciousness*. Cambridge University Press.

Bahrami, B., Olsen, K., Latham, P.E., Roepstorff, A., Rees, G. & Frith, C.D. (2010). Optimally interacting minds. *Science*, **329**, 1081–1085.

Balcetis, E. & Dunning, D. (2006). See what you want to see: motivational influences on visual perception. *Journal of Personality and Social Psychology*, **91**, 612–625.

Balcetis, E. & Dunning, D. (2007). Cognitive dissonance and the perception of natural environments. *Psychological Science*, **18**, 917–921.

Ballanger, B., van Eimeren, T., Moro, E., Lozano, A.M., Hamani, C., Boulinguez, P., Pellecchia, G., Houle, S., Poon, Y.Y., Lang, A.E. & Strafella, A.P. (2009). Stimulation of the subthalamic nucleus and impulsivity: release your horses. *Annals of Neurology*, **66**, 817–824.

Baranski, J.V. & Petrusic, W.M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance*, **24**, 929–945.

Baranski, J.V. & Petrusic, W.M. (2001). Testing architectures of the decision-confidence relation. *Canadian Journal of Experimental Psychology*, **55**, 195–206.

Barnes, J., Ridgway, G.R., Bartlett, J., Henley, S., Lehmann, M., Hobbs, N., Clarkson, M.J., MacManus, D.G., Ourselin, S. & Fox, N.C. (2010). Head size, age and gender adjustment in MRI studies: A necessary nuisance? *NeuroImage*, Epub ahead of print.

Barthelmé, S. & Mamassian, P. (2009). Evaluation of objective uncertainty in the visual system. *PLoS Computational Biology*, **5**, 1124–1131.

Basser, P.J., Mattiello, J., LEBiHAN, D. *et al.* (1994). Estimation of the effective self-diffusion tensor from the NMR spin echo. *Journal of Magnetic Resonance-Series B*, **103**, 247–254.

Baunez, C., Humby, T., Eagle, D.M., Ryan, L.J., Dunnett, S.B. & Robbins, T.W. (2001). Effects of STN lesions on simple vs choice reaction time tasks in the rat: preserved motor readiness, but impaired response selection. *European Journal of Neuroscience*, **13**, 1609–1616.

Beck, J.M., Ma, W.J., Kiani, R., Hanks, T., Churchland, A.K., Roitman, J., Shadlen, M.N., Latham, P.E. & Pouget, A. (2008). Probabilistic population codes for bayesian decision making. *Neuron*, **60**, 1142–1152.

Becker, G.M., DeGroot, M.H. & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, **9**, 226–232.

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

Bengtsson, S.L., Nagy, Z., Skare, S., Forsman, L., Forssberg, H. & Ullén, F. (2005). Extensive piano practicing has regionally specific effects on white matter development. *Nature Neuroscience*, **8**, 1148–1150.

Bengtsson, S.L., Dolan, R.J. & Passingham, R.E. (2010). Priming for self-esteem influences the monitoring of one's own performance. *Social Cognitive and Affective Neuroscience*, Epub ahead of print.

Bennur, S. & Gold, J.I. (2011). Distinct representations of a perceptual decision and the associated oculomotor plan in the monkey lateral intraparietal area. *Journal of Neuroscience*, **31**, 913–921.

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

Bergman, H., Wichmann, T. & DeLong, M.R. (1990). Reversal of experimental parkinsonism by lesions of the subthalamic nucleus. *Science*, **249**, 1436–1438.

Bertelson, P. (1999). Ventriloquism: A case of crossmodal perceptual grouping. *Advances in Psychology*, **129**, 347–362.

Bestmann, S., Harrison, L.M., Blankenburg, F., Mars, R.B., Haggard, P., Friston, K.J. & Rothwell, J.C. (2008). Influence of uncertainty and surprise on human corticospinal excitability during preparation for action. *Current Biology*, **18**, 775–780.

Bihan, D.L. (2003). Looking into the functional architecture of the brain with diffusion MRI. *Nature Reviews Neuroscience*, **4**, 469–480.

Bihan, D.L., Mangin, J.F., Poupon, C., Clark, C.A., Pappata, S., Molko, N. & Chabriat, H. (2001). Diffusion tensor imaging: concepts and applications. *Journal of Magnetic Resonance Imaging*, **13**, 534–546.

Binder, J., Liebenthal, E., Possing, E., Medler, D. & Ward, B. (2004). Neural correlates of sensory and decision processes in auditory object identification. *Nature Neuroscience*, **7**, 295–301.

Bledowski, C., Prvulovic, D., Goebel, R., Zanella, F.E. & Linden, D.E.J. (2004). Attentional systems in target and distractor processing: a combined ERP and fMRI study. *NeuroImage*, **22**, 530–540.

Block, N. (2005). Two neural correlates of consciousness. *Trends in Cognitive Sciences*, **9**, 46–52.

Bloom, P. (2004). *Descartes' baby: How the science of child development explains what makes us human*. Basic Books , New York .

Bogacz, R. (2007). Optimal decision-making theories: linking neurobiology with behaviour. *Trends in Cognitive Sciences*, **11**, 118–125.

Bogacz, R. & Gurney, K. (2007). The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural Computation*, **19**, 442–477.

Bogacz, R., Wagenmakers, E., Forstmann, B.U. & Nieuwenhuis, S. (2010). The neural basis of the speed-accuracy tradeoff. *Trends in Neurosciences*, **33**, 10–16.

Bohil, C. & Maddox, W. (2001). Category discriminability, base-rate, and payoff effects in perceptual categorization. *Perception & Psychophysics*, **63**, 361–376.

Boorman, E.D., O'Shea, J., Sebastian, C., Rushworth, M.F.S. & Johansen-Berg, H. (2007). Individual differences in white-matter microstructure reflect variation in functional connectivity during choice. *Current Biology*, **17**, 1426–1431.

Boorman, E.D., Behrens, T.E.J., Woolrich, M.W. & Rushworth, M.F.S. (2009). How green is the grass on the other side? frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron*, **62**, 733–743.

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

Boyko, R., Boyko, A. & Boyko, M. (2007). Referee bias contributes to home advantage in english premiership football. *Journal of Sports Sciences*, **25**, 1185–1194.

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. *Cerebral Cortex*, **11**, 825 –836.

Brett, M., Anton, J., Valabregue, R. & Poline, J. (2002). Regions of interest analysis using an SPM toolbox. *Presented at the 8th International Conference on Functional Mapping of the Human Brain*.

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

Brodersen, K., Penny, W., Harrison, L., Daunizeau, J., Ruff, C., Duzel, E., Friston, K. & Stephan, K. (2008). Integrated bayesian models of learning and decision making for saccadic eye movements. *Neural Networks*, **21**, 1247–1260.

Brown, J.W. & Braver, T.S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science*, **307**, 1118–1121.

Bruner, J. (1957). On perceptual readiness. *Psychological Review*, **64** , 123–152.

Bruner, J. & Goodman, C. (1947). Value and need as organising factors in perception. *Journal of Abnormal and Social Psychology*, **64**, 33–44.

Buch, E.R., Mars, R.B., Boorman, E.D. & Rushworth, M.F.S. (2010). A network centered on ventral premotor cortex exerts both facilitatory and inhibitory control over primary motor cortex during action reprogramming. *Journal of Neuroscience*, **30**, 1395–1401.

Burgess, N., Barry, C. & O'Keefe, J. (2007). An oscillatory interference model of grid cell firing. *Hippocampus*, **17**, 801–812.

Burock, M., Buckner, R., Woldorff, M., Rosen, B. & Dale, A. (1998). Randomized event-related experimental designs allow for extremely rapid presentation rates using functional MRI. *Neuroreport.*, **9**, 3735–3739.

Busey, T.A. & Arici, A. (2009). On the role of individual items in recognition memory and metacognition: challenges for signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **35**, 1123–1136.

Busey, T.A., Tunnicliff, J., Loftus, G.R. & Loftus, E.F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review*, **7**, 26–48.

Carrasco, M., Ling, S. & Read, S. (2004). Attention alters appearance. *Nature Neuroscience*, **7**, 308–313.

Carreiras, M., Seghier, M.L., Baquero, S., Estévez, A., Lozano, A., Devlin, J.T. & Price, C.J. (2009). An anatomical signature for literacy. *Nature*, **461**, 983–986.

Carruthers, P. (2009). How we know our own minds: the relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, **32**, 121–138.

Changizi, M. & Hall, W. (2001). Thirst modulates a perception. *Perception*, **30**, 1489–1497.

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*, 168–186.

Chua, E.F., Schacter, D.L. & Sperling, R.A. (2009). Neural correlates of metamemory: a comparison of feeling-of-knowing and retrospective confidence judgments. *Journal of Cognitive Neuroscience*, **21**, 1751–1765.

Churchland, A.K., Kiani, R. & Shadlen, M.N. (2008). Decision-making with multiple alternatives. *Nature Neuroscience*, **11**, 693–702.

Cisek, P. (1999). Beyond the computer metaphor: behaviour as interaction. *Journa of Consciousness Studies*, **6**, 125–142.

Cisek, P. (2007). Cortical mechanisms of action selection: the affordance competition hypothesis. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, **362**, 1585–1599.

Clarke, F., Birdsall, T. & Tanner, W. (1959). Two types of ROC curves and definition of parameters. *Journal of the Acoustical Society of America*, **31** , 629–630.

Cleeremans, A., Timmermans, B. & Pasquali, A. (2007). Consciousness and metarepresentation: A computational sketch. *Neural Networks*, **20**, 1032–1039.

Clifford, C.W., Arabzadeh, E. & Harris, J.A. (2008). Getting technical about awareness. *Trends in Cognitive Sciences*, **12**, 54–58.

Corbetta, M. & Shulman, G.L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, **3**, 201–215.

Corbetta, M., Patel, G. & Shulman, G.L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron*, **58**, 306–324.

Coulthard, E.J., Nachev, P. & Husain, M. (2008). Control over conflict during movement preparation: Role of posterior parietal cortex. *Neuron*, **58**, 144–157.

Coxon, J.P., Goble, D.J., Impe, A.V., Vos, J.D., Wenderoth, N. & Swinnen, S.P. (2010). Reduced basal ganglia function when elderly switch between coordinated movement patterns. *Cerebral Cortex*, **20**, 2363–2379.

D'Ardenne, K., McClure, S.M., Nystrom, L.E. & Cohen, J.D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science*, **319**, 1264 –1267.

Daunizeau, J., den Ouden, H.E.M., Pessiglione, M., Kiebel, S.J., Stephan, K.E. & Friston, K.J. (2010). Observing the observer (I): meta-bayesian models of learning and decision-making. *PloS One*, **5**, e15554.

Davison, M. & Tustin, R. (1978). The relation between the generalized matching law and signal-detection theory. *Journal of the Experimental Analysis of Behavior*, **29**, 331–336.

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

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

Dayan, P. & Daw, N.D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective & Behavioral Neuroscience*, **8**, 429–53.

de Lafuente, V. & Romo, R. (2006). Neural correlate of subjective sensory experience gradually builds up across cortical areas. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 14266–14271.

Dehaene, S., Sergent, C. & Changeux, J. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 8520–8525.

Del Cul, A., Dehaene, S., Reyes, P., Bravo, E. & Slachevsky, A. (2009). Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain*, **132**, 2531.

DeMartino, B., Kumaran, D., Holt, B. & Dolan, R.J. (2009). The neurobiology of reference-dependent value computation. *Journal of Neuroscience*, **29**, 3833–3842.

Desimone, R. & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, **18**, 193–222.

Dienes, Z. & Seth, A. (2010). Gambling on the unconscious: a comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and Cognition*, **19**, 674–681.

Dodds, C.M., Morein-Zamir, S. & Robbins, T.W. (2010). Dissociating inhibition, attention, and response control in the frontoparietal network using functional magnetic resonance imaging. *Cerebral Cortex*, Epub ahead of print.

Domenech, P. & Dreher, J. (2010). Decision Threshold Modulation in the Human Brain. *Journal of Neuroscience*, **30**, 14305.

Dormont, D., Ricciardi, K.G., Tande, D., Parain, K., Menuel, C., Galanaud, D., Navarro, S., Cornu, P., Agid, Y. & Yelnik, J. (2004). Is the subthalamic nucleus hypointense on t2-weighted images? a correlation study using MR imaging and stereotactic atlas data. *American Journal of Neuroradiology*, **25**, 1516.

Douek, P., Turner, R., Pekar, J., Patronas, N. & Bihan, D.L. (1991). MR color mapping of myelin fiber orientation. *Journal of Computer Assisted Tomography*, **15**, 923.

Doya, K., Ishii, S., Pouget, A. & Rao, P. (2007). *Bayesian brain: probabilistic approaches to neural coding*. MIT Press, Cambridge, Mass.

Draganski, B., Gaser, C., Busch, V., Schuierer, G., Bogdahn, U. & May, A. (2004). Neuroplasticity: changes in grey matter induced by training. *Nature*, **427**, 311–312.

Duann, J., Ide, J.S., Luo, X. & shan Ray Li, C. (2009). Functional connectivity delineates distinct roles of the inferior frontal cortex and presupplementary motor area in stop signal inhibition. *Journal of Neuroscience*, **29**, 10171–10179.

Duvernoy, H.M. (1999). *The Human Brain: Surface, Three-Dimensional Sectional Anatomy with MRI, and Blood Supply*. Springer, 2nd edn.

Eagle, D.M., Baunez, C., Hutcheson, D.M., Lehmann, O., Shah, A.P. & Robbins, T.W. (2008). Stop-signal reaction-time task performance: role of prefrontal cortex and subthalamic nucleus. *Cerebral Cortex*, **18**, 178–188.

Egner, T., Monti, J.M. & Summerfield, C. (2010). Expectation and surprise determine neural population responses in the ventral visual stream. *Journal of Neuroscience*, **30**, 16601–16608.

Ekstrom, L.B., Roelfsema, P.R., Arsenault, J.T., Bonmassar, G. & Vanduffel, W. (2008). Bottom-up dependent gating of frontal signals in early visual cortex. *Science*, **321**, 414–417.

Epstein, R. & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature.*, **392**, 598–601.

Eriksen, B. & Eriksen, C. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, **16**, 143–149.

Eriksen, C.W. (1960). Discrimination and learning without awareness: a methodological survey and evaluation. *Psychological Review*, **67**, 279–300.

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

Esterman, M. & Yantis, S. (2010). Perceptual expectation evokes category-selective cortical activity. *Cerebral Cortex*, **20**, 1245–1253.

Evans, S. & Azzopardi, P. (2007). Evaluation of a 'bias-free' measure of awareness. *Spatial Vision*, **20**, 61–77.

Faisal, A.A. & Wolpert, D.M. (2009). Near optimal combination of sensory and motor uncertainty in time during a naturalistic perception-action task. *Journal of Neurophysiology*, **101**, 1901–1912.

Feng, S., Holmes, P., Rorie, A. & Newsome, W.T. (2009). Can monkeys choose optimally when faced with noisy stimuli and unequal rewards? *PLoS Computational Biology*, **5**, e1000284.

Ferrera, V.P., Yanike, M. & Cassanello, C. (2009). Frontal eye field neurons signal changes in decision criteria. *Nature Neuroscience*, **12**, 1458–1462.

Flandin, G. & Friston, K. (2008). Statistical parametric mapping. *Scholarpedia*, **3**, 6232.

Flavell, J. (1979). Metacognition and cognitive monitoring : A new area of cognitive - developmental inquiry. *American Psychologist*, **34**, 906–911.

Fleming, S.M. (2009). Shaping what we see: Pinning down the influence of value on perceptual judgements. *Frontiers in Human Neuroscience*, **3**, 9.

Fleming, S.M. & Dolan, R.J. (2010). Effects of loss aversion on post-decision wagering: Implications for measures of awareness. *Consciousness and Cognition*, **19**, 352–363.

Fleming, S.M., Mars, R.B., Gladwin, T.E. & Haggard, P. (2009). When the brain changes its mind: flexibility of action selection in instructed and free choices. *Cerebral Cortex*, **19**, 2352–2360.

Fleming, S.M., Thomas, C.L. & Dolan, R.J. (2010a). Overcoming status quo bias in the human brain. *Proceedings of the National Academy of Sciences*, **107**, 6005–6009.

Fleming, S.M., Weil, R.S., Nagy, Z., Dolan, R.J. & Rees, G. (2010b). Relating introspective accuracy to individual differences in brain structure. *Science*, **329**, 1541–1543.

Fleming, S.M., Whiteley, L., Hulme, O.J., Sahani, M. & Dolan, R.J. (2010c). Effects of category-specific costs on neural systems for perceptual decision-making. *Journal of Neurophysiology*, **103**, 3238–3247.

Fletcher, P.C. & Henson, R.N. (2001). Frontal lobes and human memory: insights from functional neuroimaging. *Brain*, **124**, 849–881.

Foote, A. & Crystal, J. (2007). Metacognition in the rat. *Current Biology*, **17**, 551–555.

Forstmann, B.U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D.Y., Ridderinkhof, K.R. & Wagenmakers, E.J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 17538.

Forstmann, B.U., Anwander, A., Schäfer, A., Neumann, J., Brown, S., Wagenmakers, E., Bogacz, R. & Turner, R. (2010). Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 15916–15920.

Frank, M.J. (2006). Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making. *Neural Networks*, **19**, 1120–1136.

Frank, M.J., Samanta, J., Moustafa, A.A. & Sherman, S.J. (2007). Hold your horses: impulsivity, deep brain stimulation, and medication in parkinsonism. *Science*, **318**, 1309–12.

Freedman, D.J. & Assad, J.A. (2011). A proposed common neural mechanism for categorization and perceptual decisions. *Nature Neuroscience*, **14**, 143–146.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, **13**, 293–301.

Friston, K., Jezzard, P. & Turner, R. (1994a). Analysis of functional MRI time-series. *Human Brain Mapping*, **1**, 153–171.

Friston, K., Worsley, K., Frackowiak, R., Mazziotta, J. & Evans, A. (1994b). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, **1** , 214–220.

Friston, K., Holmes, A., Worsley, K., Poline, J., Frith, C. & Frackowiak, R. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, **2**, 189–210.

Friston, K., Stephan, K., Lund, T., Morcom, A. & Kiebel, S. (2005). Mixed-effects and fMRI studies. *NeuroImage*, **24**, 244–252.

Friston, K., Kilner, J. & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, **100**, 70–87.

Friston, K.J., Harrison, L. & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, **19**, 1273–1302.

Friston, K.J., Daunizeau, J. & Kiebel, S.J. (2009). Reinforcement learning or active inference? *PloS One*, **4**, e6421.

Frith, U. & Frith, C.D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **358**, 459–473.

Fuentemilla, L., Càmara, E., Münte, T.F., Krämer, U.M., Cunillera, T., Marco-Pallarés, J., Tempelmann, C. & Rodriguez-Fornells, A. (2009). Individual differences in true and false memory retrieval are related to white matter brain microstructure. *Journal of Neuroscience*, **29**, 8698–8703.

Fumagalli, M., Giannicola, G., Rosa, M., Marceglia, S., Lucchiari, C., Mrakic-Sposta, S., Servello, D., Pacchetti, C., Porta, M., Sassi, M., Zangaglia, R., Franzini, A., Albanese, A., Romito, L., Piacentini, S., Zago, S., Pravettoni, G., Barbieri, S. & Priori, A. (2010). Conflict-dependent dynamic of subthalamic nucleus oscillations during moral decisions. *Social Neuroscience*, 1–14.

Galvin, S.J., Podd, J.V., Drga, V. & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, **10**, 843–876.

Garavan, H., Ross, T.J. & Stein, E.A. (1999). Right hemispheric dominance of inhibitory control: an event-related functional MRI study. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 8301–8306.

Gehring, W.J., Goss, B., Coles, M.G., Meyer, D.E. & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, **4**, 385.

Gigerenzer, G., Hoffrage, U. & Kleinbölting, H. (1991). Probabilistic mental models: a brunswikian theory of confidence. *Psychological Review*, **98**, 506–528.

Glaser, D. (2004). Variance components. In R. Frackowiak, ed., *Human Brain Function*, Academic Press.

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

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

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

Gold, J.I. & Shadlen, M.N. (2003). The influence of behavioral context on the representation of a perceptual decision in developing oculomotor commands. *Journal of Neuroscience*, **23**, 632–651.

Gottfried, J.A., O'Doherty, J. & Dolan, R.J. (2003). Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science*, **301**, 1104.

Gradinaru, V., Mogri, M., Thompson, K.R., Henderson, J.M. & Deisseroth, K. (2009). Optical deconstruction of parkinsonian neural circuitry. *Science*, **324**, 354–359.

Graziano, M. & Sigman, M. (2009). The spatial and temporal construction of confidence in the visual scene. *PLoS ONE*, **4**, e4909.

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

Grill-Spector, K., Kourtzi, Z. & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision research*, **41**, 1409–1422.

Grinband, J., Hirsch, J. & Ferrera, V. (2006). A neural representation of categorization uncertainty in the human brain. *Neuron*, **49**, 757–763.

Grinband, J., Wager, T.D., Lindquist, M., Ferrera, V.P. & Hirsch, J. (2008). Detection of time-varying signals in event-related fMRI designs. *NeuroImage*, **43**, 509–520.

Gross, C.G. & Schonen, S.D. (1992). Representation of visual stimuli in inferior temporal cortex [and discussion]. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **335**, 3–10.

Gurney, K., Prescott, T.J. & Redgrave, P. (2001). A computational model of action selection in the basal ganglia. i. a new functional anatomy. *Biological Cybernetics*, **84**, 401–410.

Haber, S.N. (2003). The primate basal ganglia: parallel and integrative networks. *Journal of Chemical Neuroanatomy*, **26**, 317–330.

Hampshire, A. & Owen, A.M. (2006). Fractionating attentional control using event-related fMRI. *Cerebral Cortex*, **16**, 1679–1689.

Hampshire, A., Duncan, J. & Owen, A.M. (2007). Selective tuning of the blood oxygenation level-dependent response during simple target detection dissociates human frontoparietal subregions. *Journal of Neuroscience*, **27**, 6219–6223.

Hampton, R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5359–5362.

Hariz, M.I., Krack, P., Melvill, R., Jorgensen, J.V., Hamel, W., Hirabayashi, H., Lenders, M., Wesslen, N., Tengvar, M. & Yousry, T.A. (2003). A quick and universal method for stereotactic visualization of the subthalamic nucleus before and after implantation of deep brain stimulation electrodes. *Stereotactic and Functional Neurosurgery*, **80**, 96–101.

Hayasaka, S., Phan, K.L., Liberzon, I., Worsley, K.J. & Nichols, T.E. (2004). Nonstationary cluster-size inference with random field and permutation methods. *NeuroImage*, **22**, 676–687.

Healy, A.F. & Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory*, **7**, 344–354.

Heeger, D.J. & Ress, D. (2002). What does fMRI tell us about neuronal activity? *Nature Reviews Neuroscience*, **3**, 142–151.

Heekeren, H., Marrett, S., Bandettini, P. & Ungerleider, L. (2004). A general mechanism for perceptual decision-making in the human brain. *Nature*, **431**, 859–862.

Heekeren, H., Marrett, S., Ruff, D., Bandettini, P. & Ungerleider, L. (2006). Involvement of human left dorsolateral prefrontal cortex in perceptual decision making is independent of response modality. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 10023–10028.

Heekeren, H., Marrett, S. & Ungerleider, L. (2008). The neural systems that mediate human perceptual decision making. *Nature Reviews Neuroscience*, **9**, 467–479.

Henderson, J.M. & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, **50**, 243–271.

Hernandez, A., Zainos, A. & Romo, R. (2002). Temporal evolution of a decision-making process in medial premotor cortex. *Neuron*, **33**, 959–972.

Hershey, T., Revilla, F.J., Wernle, A., Gibson, P.S., Dowling, J.L. & Perlmutter, J.S. (2004). Stimulation of STN impairs aspects of cognitive control in PD. *Neurology*, **62**, 1110–1114.

Hershey, T., Campbell, M.C., Videen, T.O., Lugar, H.M., Weaver, P.M., Hartlein, J., Karimi, M., Tabbal, S.D. & Perlmutter, J.S. (2010). Mapping Go-No-Go performance within the subthalamic nucleus region. *Brain: A Journal of Neurology*.

Higham, P.A. (2007). No special k! a signal detection framework for the strategic regulation of memory accuracy. *Journal of Experimental Psychology: General*, **136**, 1–22.

Higham, P.A., Perfect, T.J. & Bruno, D. (2009). Investigating strength and frequency effects in recognition memory using type-2 signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **35**, 57–80.

Ho, T.C., Brown, S. & Serences, J.T. (2009). Domain general mechanisms of perceptual decision making in human cortex. *Journal of Neuroscience*, **29**, 8675–8687.

Hodgson, T., Chamberlain, M., Parris, B., James, M., Gutowski, N., Husain, M. & Kennard, C. (2007). The role of the ventrolateral frontal cortex in inhibitory oculomotor control. *Brain*, **130**, 1525–1537.

Hollard, G., Massoni, S. & Vergnaud, J.C. (2010). Subjective belief formation and elicitation rules: experimental evidence. *unpublished manuscript*.

Holroyd, C.B. & Coles, M.G. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, **109**, 679–708.

Holt, C.A. & Smith, A.M. (2009). An update on bayesian updating. *Journal of Economic Behavior & Organization*, **69**, 125–134.

Howell, D.C. (2009). *Statistical methods for psychology*. Wadsworth Pub Co.

Hurley, S.L. (2002). *Consciousness in action*. Harvard University Press.

Insabato, A., Pannunzi, M., Rolls, E.T. & Deco, G. (2010). Confidence-related decision making. *Journal of Neurophysiology*, **104**, 539–547.

Isoda, M. & Hikosaka, O. (2007). Switching from automatic to controlled action by monkey medial frontal cortex. *Nature Neuroscience*, **10**, 240–248.

Isoda, M. & Hikosaka, O. (2008). Role for subthalamic nucleus neurons in switching from automatic to controlled eye movement. *Journal of Neuroscience*, **28**, 7209.

Iyengar, S.S. & Lepper, M.R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, **79**, 995–1006.

Jack, A.I. & Roepstorff, A. (2002). Introspection and cognitive brain mapping: from stimulus–response to script–report. *Trends in Cognitive Sciences*, **6**, 333–339.

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

Jahanshahi, M., Dirnberger, G., Fuller, R. & Frith, C. (2000). The role of the dorsolateral prefrontal cortex in random number generation: a study with positron emission tomography. *NeuroImage*, **12**, 713–725.

Janowsky, J.S., Shimamura, A.P., Kritchevsky, M. & Squire, L.R. (1989). Cognitive impairment following frontal lobe damage and its relevance to human amnesia. *Behavioral Neuroscience*, **103**, 548.

Jansons, K.M. & Alexander, D.C. (2003). Persistent angular structure: new insights from diffusion MRI data. dummy version. *Information Processing in Medical Imaging*, **18**, 672–683.

Johansson, P., Hall, L., Sikstrom, S. & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, **310**, 116–119.

Johnson, E. & Goldstein, D. (2003). Medicine: Do defaults save lives? *Science*, **302**, 1338–1339.

Johnstone, V. & Alsop, B. (1996). Human signal-detection performance: Effects of signal presentation probabilities and reinforcer distributions. *J.Exp.Anal.Behav.*, **66**, 243–263.

Johnstone, V. & Alsop, B. (2000). Reinforcer control and human signal-detection performance. *Journal of the Experimental Analysis of Behavior*, **73**, 275–290.

Kahneman, D. & Tversky, A. (1979). Prospect theory - analysis of decision under risk. *Econometrica*, **47**, 263–291.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, **82**, 35–45.

Kanwisher, N., McDermott, J. & Chun, M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, **17**, 4302–4311.

Kao, Y., Davis, E.S. & Gabrieli, J.D.E. (2005). Neural correlates of actual and predicted memory formation. *Nat Neurosci*, **8**, 1776–1783.

Kenner, N.M., Mumford, J.A., Hommer, R.E., Skup, M., Leibenluft, E. & Poldrack, R.A. (2010). Inhibitory motor control in response stopping and response switching. *Journal of Neuroscience*, **30**, 8512–8518.

Kepecs, A., Uchida, N., Zariwala, H. & Mainen, Z. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, **455**, 227–231.

Kersten, D., Mamassian, P. & Yuille, A. (2004). Object perception as bayesian inference. *Annual Review of Psychology*, **55**, 271–304.

Kiani, R. & Shadlen, M.N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, **324**, 759–764.

Kim, H. & Cabeza, R. (2007). Trusting our memories: dissociating the neural correlates of confidence in veridical versus illusory memories. *Journal of Neuroscience*, **27**, 12190.

Kim, J. & Shadlen, M. (1999). Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nature Neuroscience*, **2**, 176–185.

Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., Song, J.H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R.P., Mann, J.J. & Parsey, R.V. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*, **46**, 786–802.

Knill, D.C. & Saunders, J.A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research*, **43**, 2539–2558.

Koechlin, E. & Hyafil, A. (2007). Anterior prefrontal function and the limits of human decision-making. *Science*, **318**, 594–598.

Kording, K. (2007). Decision theory: What 'should' the nervous system do? *Science*, **318**, 606–610.

Koriat, A. & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, **103**, 490–517.

Kornbrot, D.E. (2006). Signal detection theory, the approach of choice: model-based and distribution-free measures and evaluation. *Perception & Psychophysics*, **68**, 393–414.

Koski, L. & Paus, T. (2000). Functional connectivity of the anterior cingulate cortex within the human frontal lobe: a brain-mapping meta-analysis. *Experimental Brain Research*, **133**, 55–65.

Kouneiher, F., Charron, S. & Koechlin, E. (2009). Motivation and cognitive control in the human prefrontal cortex. *Nature Neuroscience*, **12**, 939–945.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F. & Baker, C.I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, **12**, 535–540.

Kruger, J. & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, **77**, 1121–1134.

Kubicki, M., Shenton, M.E., Salisbury, D.F., Hirayasu, Y., Kasai, K., Kikinis, R., Jolesz, F.A. & McCarley, R.W. (2002). Voxel-based morphometric analysis of gray matter in first episode schizophrenia. *Neuroimage*, **17**, 1711–1719.

Kunimoto, C., Miller, J. & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition*, **10**, 294–340.

Landy, M., Goutcher, R., Trommershauser, J. & Mamassian, P. (2007). Visual estimation under risk. *Journal of Vision*, **7**, 4.

Lau, H. (2010). Are we studying consciousness yet? In L. Weiskrantz & M. Davies, eds., *Frontiers of Consciousness: Chichele Lectures*, Oxford University Press, Oxford, UK.

Lau, H. & Passingham, R. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proc.Natl.Acad.Sci.U.S.A*, **103**, 18763–18768.

Lau, H., Rogers, R. & Passingham, R. (2006). Dissociating response selection and conflict in the medial frontal surface. *NeuroImage.*, **29**, 446–451.

Lau, H.C. (2008). A higher order bayesian decision theory of consciousness. *Progress in Brain Research*, **168**, 35–48.

Lauwereyns, J., Watanabe, K., Coe, B. & Hikosaka, O. (2002). A neural correlate of response bias in monkey caudate nucleus. *Nature*, **418**, 413–417.

Lemus, L., Hernandez, A., Luna, R., Zainos, A., Nacher, V. & Romo, R. (2007). Neural correlates of a postponed decision report. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 17174–17179.

Leon, M. & Shadlen, M. (1999). Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. *Neuron*, **24**, 415–425.

Leung, H. & Cai, W. (2007). Common and differential ventrolateral prefrontal activity during inhibition of hand and eye movements. *Journal of Neuroscience*, **27**, 9893–9900.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, **49**, Suppl.

Li, C.R., Yan, P., Sinha, R. & Lee, T. (2008). Subcortical processes of motor response inhibition during a stop signal task. *NeuroImage*, **41**, 1352–63.

Li, S., Mayhew, S.D. & Kourtzi, Z. (2009). Learning shapes the representation of behavioral choice in the human brain. *Neuron*, **62**, 441–452.

Lie, C. & Alsop, B. (2010). Stimulus disparity and punisher control of human signal-detection performance. *Journal of the Experimental Analysis of Behavior*, **93**, 185–201.

Limousin, P. & Martinez-Torres, I. (2008). Deep brain stimulation for parkinson's disease. *Neurotherapeutics*, **5**, 309–319.

Lin, P., Hasson, U., Jovicich, J. & Robinson, S. (2010). A neuronal basis for Task-Negative responses in the human brain. *Cerebral Cortex*, Epub ahead of print.

Link, S.W. (1975). The relative judgment theory of two choice response time. *Journal of Mathematical Psychology*, **12**, 114–135.

Liston, D. & Stone, L. (2008). Effects of prior information and reward on oculomotor and perceptual choices. *Journal of Neuroscience*, **28**, 13866–13875.

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

Lobel, E., Kahane, P., Leonards, U., Grosbras, M., Lehéricy, S., Bihan, D.L. & Berthoz, A. (2001). Localization of human frontal eye fields: anatomical and functional findings of functional magnetic resonance imaging and intracerebral electrical stimulation. *Journal of Neurosurgery*, **95**, 804–815.

Logothetis, N.K. (2008). What we can do and what we cannot do with fMRI. *Nature*, **453**, 869–878.

Lu, Z. & Dosher, B.A. (2008). Characterizing observers using external noise and observer models: assessing internal representations with external noise. *Psychological Review*, **115**, 44–82.

Luce, R.D. (1991). *Response Times: Their Role in Inferring Elementary Mental Organization*. Oxford University Press US.

Luders, E., Narr, K.L., Thompson, P.M., Rex, D.E., Woods, R.P., DeLuca, H., Jancke, L. & Toga, A.W. (2006). Gender effects on cortical thickness and the influence of scaling. *Human Brain Mapping*, **27**, 314–324.

Lycan, W.G. (1995). Consciousness as internal monitoring, i: The third philosophical perspectives lecture. *Philosophical Perspectives*, **9**, 1–14.

Ma, W.J. (2010). Signal detection theory, uncertainty, and poisson-like population codes. *Vision Research*, **50**, 2308–19.

MacDonald, A.W., Cohen, J.D., Stenger, V.A. & Carter, C.S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science*, **288**, 1835 –1838.

MacKay, D.J.C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

Macmillan, N. & Creelman, C. (2005). *Detection theory: a user's guide*. Lawrence Erlbaum, New York.

Maddox, W. (2002). Toward a unified theory of decision criterion learning in perceptual categorization. *Journal of the Experimental Analysis of Behavior*, **78**, 567–595.

Maddox, W. & Bohil, C. (2003). A theoretical framework for understanding the effects of simultaneous base-rate and payoff manipulations on decision criterion learning in perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **29**, 307–320.

Maddox, W. & Bohil, C. (2004). Probability matching, accuracy maximization, and a test of the optimal classifier's independence assumption in perceptual categorization. *Perception & Psychophysics*, **66**, 104–118.

Magno, E., Foxe, J.J., Molholm, S., Robertson, I.H. & Garavan, H. (2006). The anterior cingulate and error avoidance. *Journal of Neuroscience*, **26**, 4769.

Maldjian, J.A., Laurienti, P.J., Kraft, R.A. & Burdette, J.H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *NeuroImage*, **19**, 1233–1239.

Mallet, L., Schüpbach, M., N'Diaye, K., Remy, P., Bardinet, E., Czernecki, V., Welter, M., Pelissolo, A., Ruberg, M., Agid, Y. & Yelnik, J. (2007). Stimulation of subterritories of the subthalamic nucleus reveals its role in the integration of the emotional and motor aspects of behavior. *Proceedings of the National Academy of Sciences*, **104**, 10661 –10666.

Maloney, L. (2002). Statistical decision theory and biological vision. In D. Heyer & R. Mausfeld, eds., *Perception and the Physical World: Psychological and Philisophical Issues in Perception*, Wiley, New York.

Maloney, L.T. & Thomas, E.A. (1991). Distributional assumptions and observed conservatism in the theory of signal detectability. *Journal of Mathematical Psychology*, **35**, 443–470.

Marani, E., Heida, T., Lakke, E. & Usunoff, K.G. (2008). *The subthalamic nucleus, Part I*. Springer.

Mars, R., Shea, N., Kolling, N. & Rushworth, M. (2010). Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. *The Quarterly Journal of Experimental Psychology*, Epub ahead of print.

Mars, R.B., Piekema, C., Coles, M.G., Hulstijn, W. & Toni, I. (2007). On the programming and reprogramming of actions. *Cerebral Cortex*, **17**, 2972–2979.

Mars, R.B., Debener, S., Gladwin, T.E., Harrison, L.M., Haggard, P., Rothwell, J.C. & Bestmann, S. (2008). Trial-by-Trial fluctuations in the Event-Related electroencephalogram reflect dynamic changes in the degree of surprise. *Journal of Neuroscience*, **28**, 12539.

Mars, R.B., Klein, M.C., Neubert, F., Olivier, E., Buch, E.R., Boorman, E.D. & Rushworth, M.F.S. (2009). Short-latency influence of medial frontal cortex on primary motor cortex during action selection under conflict. *Journal of Neuroscience*, **29**, 6926–6931.

Marti, S., Sackur, J., Sigman, M. & Dehaene, S. (2010). Mapping introspection's blind spot: reconstruction of dual-task phenomenology using quantified introspection. *Cognition*, **115**, 303–313.

Mather, G. (2008). Perceptual uncertainty and line-call challenges in professional tennis. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **275**, 1645–1651.

Maunsell, J.H. (2004). Neuronal representations of cognitive state: reward or attention? *Trends in Cognitive Sciences*, **8**, 261–265.

McKeeff, T. & Tong, F. (2007). The timing of perceptual decisions for ambiguous face stimuli in the human ventral visual cortex. *Cerebral Cortex*, **17**, 669–678.

Menon, V., Adleman, N.E., White, C.D., Glover, G.H. & Reiss, A.L. (2001). Error-related brain activation during a Go/NoGo response inhibition task. *Human Brain Mapping*, **12**, 131–143.

Metcalfe, J. (1996). *Metacognition: Knowing About Knowing*. MIT Press.

Metcalfe, J. (2008). Evolution of metacognition. In J. Dunlosky & R. Bjork, eds., *Handbook of Metamemory and Memory*, 27–46, Psychology Press.

Metzinger, T. (2010). *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. Basic Books.

Middlebrooks, P. & Sommer, M.A. (2010). Frontal eye field, lateral prefrontal cortex, and supplementary eye field single neuron activity during a visual-saccadic metacognition task. *Society for Neurocience Abstracts*, **280.13**.

Miller, E.K. & Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, **24**, 167–202.

Monakow, K., Akert, K. & Künzle, H. (1978). Projections of the precentral motor cortex and other cortical areas of the frontal lobe to the subthalamic nucleus in the monkey. *Experimental Brain Research*, **33**, 395–403.

Moreno-Bote, R. (2010). Decision confidence and uncertainty in diffusion models with partially correlated neuronal integrators. *Neural Computation*, Epub ahead of print.

Moritz, S., Gläscher, J., Sommer, T., Büchel, C. & Braus, D.F. (2006). Neural correlates of memory confidence. *NeuroImage*, **33**, 1188–1193.

Mostofsky, S.H. & Simmonds, D.J. (2008). Response inhibition and response selection: two sides of the same coin. *Journal of Cognitive Neuroscience*, **20**, 751–761.

Näätänen, R., Paavilainen, P., Alho, K., Reinikainen, K. & Sams, M. (1987). The mismatch negativity to intensity changes in an auditory stimulus sequence. *Electroencephalography and Clinical Neurophysiology. Supplement*, **40**, 125–131.

Nachev, P., Wydell, H., O'Neill, K., Husain, M. & Kennard, C. (2007). The role of the pre-supplementary motor area in the control of action. *NeuroImage*, **36 Suppl 2**, T155–163.

Nambu, A., Tokuno, H., Hamada, I., Kita, H., Imanishi, M., Akazawa, T., Ikeuchi, Y. & Hasegawa, N. (2000). Excitatory cortical inputs to pallidal neurons via the subthalamic nucleus in the monkey. *Journal of Neurophysiology*, **84**, 289–300.

Nambu, A., Tokuno, H. & Takada, M. (2002). Functional significance of the cortico-subthalamo-pallidal 'hyperdirect' pathway. *Neuroscience Research*, **43**, 111–117.

Nelson, T.O. & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The psychology of learning and motivation: Advances in research and theory*, **26**, 125–173.

Neubert, F., Mars, R.B., Buch, E.R., Olivier, E. & Rushworth, M.F.S. (2010). Cortical and subcortical interactions during action reprogramming and their related white matter pathways. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 13240–13245.

Newsome, W., Britten, K. & Movshon, J. (1989). Neuronal correlates of a perceptual decision. *Nature*, **341**, 52–54.

Neyman, J. & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A*, **231**, 289–337.

Nicolle, A., Fleming, S., Bach, D., Driver, J. & Dolan, R. (2011). A regret-induced default bias. *Journal of Neuroscience*, in press.

Nieuwenhuis, S., Ridderinkhof, K.R., Blom, J., Band, G.P. & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. *Psychophysiology*, **38**, 752–760.

Nieuwenhuis, S., Yeung, N., Wildenberg, W.V.D. & Ridderinkhof, K.R. (2003). Electrophysiological correlates of anterior cingulate function in a go/no-go task: Effects of response conflict and trial type frequency. *Cognitive, Affective, & Behavioral Neuroscience*, **3**, 17.

Nisbett, R.E. & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, **84**, 231.

Noppeney, U., Ostwald, D. & Werner, S. (2010). Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex. *Journal of Neuroscience*, **30**, 7434–7446.

Northoff, G. & Bermpohl, F. (2004). Cortical midline structures and the self. *Trends in Cognitive Sciences*, **8**, 102–107.

Ogawa, S., Lee, T.M., Kay, A.R. & Tank, D.W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 9868.

Ogawa, S., Tank, D.W., Menon, R., Ellermann, J.M., Kim, S.G., Merkle, H. & Ugurbil, K. (1992). Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 5951.

Palmer, J., Huk, A.C. & Shadlen, M.N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, **5**, 376–404.

Park, H.J., Kim, J.J., Lee, S.K., Seok, J.H., Chun, J., Kim, D.I. & Lee, J.D. (2008). Corpus callosal connection mapping using cortical gray matter parcellation and DT-MRI. *Human Brain Mapping*, **29**, 503–516.

Pasquali, A., Timmermans, B. & Cleeremans, A. (2010). Know thyself: metacognitive networks and measures of consciousness. *Cognition*, **117**, 182–190.

Pasquereau, B., Nadjar, A., Arkadir, D., Bezard, E., Goillandeau, M., Bioulac, B., Gross, C.E. & Boraud, T. (2007). Shaping of motor responses by incentive values through the basal ganglia. *Journal of Neuroscience*, **27**, 1176–1183.

Passingham, R.E., Bengtsson, S.L. & Lau, H.C. (2010). Medial frontal cortex: from self-generated action to reflection on one's own performance. *Trends in Cognitive Sciences*, **14**, 16–21.

Peirce, C.S. & Jastrow, J. (1885). On small differences in sensation. *Memoirs of the National Acadmey of Sciences*, **3**, 73–83.

Persaud, N., McLeod, P. & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience*, **10**, 257–261.

Philiastides, M.G. & Sajda, P. (2007). EEG-informed fMRI reveals spatiotemporal characteristics of perceptual decision making. *Journal of Neuroscience*, **27**, 13082–13091.

Philiastides, M.G., Ratcliff, R. & Sajda, P. (2006). Neural representation of task difficulty and decision making during perceptual categorization: a timing diagram. *Journal of Neuroscience*, **26**, 8965–8975.

Pick, H.L., Warren, D.H. & Hay, J.C. (1969). Sensory conflict in judgments of spatial direction. *Perception & Psychophysics*, **6**, 203–205.

Pleger, B., Ruff, C.C., Blankenburg, F., Bestmann, S., Wiech, K., Stephan, K.E., Capilla, A., Friston, K.J. & Dolan, R.J. (2006). Neural coding of tactile decisions in the human prefrontal cortex. *Journal of Neuroscience*, **26**, 12596–12601.

Pleger, B., Blankenburg, F., Ruff, C., Driver, J. & Dolan, R. (2008). Reward facilitates tactile judgments and modulates hemodynamic responses in human primary somatosensory cortex. *Journal of Neuroscience*, **28**, 8161–8168.

Pleger, B., Ruff, C.C., Blankenburg, F., Klöppel, S., Driver, J. & Dolan, R.J. (2009). Influence of dopaminergically mediated reward on somatosensory decision-making. *PLoS Biology*, **7**, e1000164.

Pleskac, T.J. & Busemeyer, J.R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, **117**, 864–901.

Ploran, E.J., Nelson, S.M., Velanova, K., Donaldson, D.I., Petersen, S.E. & Wheeler, M.E. (2007). Evidence accumulation and the moment of recognition: dissociating perceptual recognition processes using fMRI. *Journal of Neuroscience*, **27**, 11912–11924.

Poldrack, R.A., Halchenko, Y.O. & Hanson, S.J. (2009). Decoding the large-scale structure of brain function by classifying mental states across individuals. *Psychological Science*, **20**, 1364–1372.

Poline, J., Worsley, K., Evans, A. & Friston, K. (1997). Combining spatial extent and peak intensity to test for activations in functional imaging. *NeuroImage*, **5**, 83–96.

Posner, M.I., Snyder, C.R. & Davidson, B.J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, **109**, 160–174.

Preuschoff, K., Quartz, S.R. & Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience*, **28**, 2745–2752.

Proshansky, H. & Murphy, G. (1942). The effects of reward and punishment on perception. *Journal of Psychology*, **13** , 295–305.

Puri, A.M., Wojciulik, E. & Ranganath, C. (2009). Category expectation modulates baseline and stimulus-evoked activity in human inferotemporal cortex. *Brain Research*, **1301**, 89–99.

Pylyshyn, Z. (2003). *Seeing and visualizing*. MIT Press, Cambridge, MA.

Ramnani, N., Behrens, T.E.J., Penny, W. & Matthews, P.M. (2004). New approaches for exploring anatomical and functional connectivity in the human brain. *Biological Psychiatry*, **56**, 613–619.

Rangel, A., Camerer, C. & Montague, P. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, **9**, 545–556.

Rao, R.P. & Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, **2**, 79–87.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, **85**, 59–108.

Ray, N.J., Jenkinson, N., Brittain, J., Holland, P., Joint, C., Nandi, D., Bain, P.G., Yousif, N., Green, A., Stein, J.S. & Aziz, T.Z. (2009). The role of the subthalamic nucleus in response inhibition: evidence from deep brain stimulation for parkinson's disease. *Neuropsychologia*, **47**, 2828–2834.

Redgrave, P., Rodriguez, M., Smith, Y., Rodriguez-Oroz, M.C., Lehericy, S., Bergman, H., Agid, Y., DeLong, M.R. & Obeso, J.A. (2010). Goal-directed and habitual control in the basal ganglia: implications for parkinson's disease. *Nature Reviews Neuroscience*, **11**, 760–772.

Resulaj, A., Kiani, R., Wolpert, D.M. & Shadlen, M.N. (2009). Changes of mind in decision-making. *Nature*, **461**, 263–266.

Ridderinkhof, K.R., Ullsperger, M., Crone, E.A. & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science*, **306**, 443–447.

Roitman, J. & Shadlen, M. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience*, **22**, 9475–9489.

Romo, R., Hernandez, A., Zainos, A., Brody, C. & Salinas, E. (2002a). Exploring the cortical evidence of a sensory-discrimination process. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, **357**, 1039–1051.

Romo, R., Hernandez, A., Zainos, A., Lemus, L. & Brody, C. (2002b). Neuronal correlates of decision-making in secondary somatosensory cortex. *Nature Neuroscience*, **5**, 1217–1225.

Romo, R., Hernandez, A. & Zainos, A. (2004). Neuronal correlates of a perceptual decision in ventral premotor cortex. *Neuron*, **41**, 165–173.

Rorie, A.E., Gao, J., McClelland, J.L. & Newsome, W.T. (2010). Integration of sensory and reward information during perceptual decision-making in lateral intraparietal cortex (LIP) of the macaque monkey. *PLoS ONE*, **5**, e9308.

Rosa, M., Bestmann, S., Harrison, L. & Penny, W. (2010). Bayesian model selection maps for group studies. *NeuroImage*, **49**, 217–224.

Rosenthal, D.M. (2005). *Consciousness and mind*. Oxford University Press.

Rosenthal, D.M. (2008). Consciousness and its function. *Neuropsychologia*, **46**, 829–840.

Rounis, E., Maniscalco, B., Rothwell, J.C., Passingham, R.E. & Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience*, **1**, 165–175.

Rubia, K., Russell, T., Overmeyer, S., Brammer, M.J., Bullmore, E.T., Sharma, T., Simmons, A., Williams, S.C., Giampietro, V., Andrew, C.M. & Taylor, E. (2001). Mapping motor inhibition: conjunctive brain activations across different versions of go/no-go and stop tasks. *NeuroImage*, **13**, 250–261.

Salzman, C. & Newsome, W. (1994). Neural mechanisms for forming a perceptual decision. *Science.*, **264**, 231–237.

Samuelson, W. & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, **1**, 7–59.

Schoenbaum, G., Saddoris, M.P. & Stalnaker, T.A. (2007). Reconciling the roles of orbitofrontal cortex in reversal learning and the encoding of outcome expectancies. *Annals of the New York Academy of Sciences*, **1121**, 320–335.

Scholz, J., Klein, M.C., Behrens, T.E.J. & Johansen-Berg, H. (2009). Training induces changes in white-matter architecture. *Nature Neuroscience*, **12**, 1370–1371.

Schurger, A. & Sher, S. (2008). Awareness, loss aversion, and post-decision wagering. *Trends in Cognitive Sciences*, **12**, 209–210.

Serences, J. (2008). Value-based modulations in human visual cortex. *Neuron*, **60**, 1169–1181.

Serences, J.T., Schwarzbach, J., Courtney, S.M., Golay, X. & Yantis, S. (2004). Control of object-based attention in human cortex. *Cerebral Cortex*, **14**, 1346–1357.

Seth, A. (2008). Post-decision wagering measures metacognitive content, not sensory consciousness. *Consciousness and Cognition*, **17**, 981–983.

Shadlen, M. & Newsome, W. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, **86**, 1916–1936.

Sharp, D.J., Bonnelle, V., Boissezon, X.D., Beckmann, C.F., James, S.G., Patel, M.C. & Mehta, M.A. (2010). Distinct frontal systems for response inhibition, attentional capture, and error processing. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 6106–6111.

Shepherd, G.M. (2004). *The synaptic organization of the brain*. Oxford University Press.

Shimamura, A.P. (2000). Toward a cognitive neuroscience of metacognition. *Consciousness and Cognition*, **9**, 313–323.

Shimamura, A.P. & Squire, L.R. (1986). Memory and metamemory: a study of the feeling-of-knowing phenomenon in amnesic patients. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, **12**, 452–460.

Simen, P., Cohen, J.D. & Holmes, P. (2006). Rapid decision threshold modulation by reward rate in a neural network. *Neural Networks*, **19**, 1013–1026.

Simen, P., Contreras, D., Buck, C., Hu, P., Holmes, P. & Cohen, J.D. (2009). Reward rate optimization in two-alternative decision making: empirical tests of theoretical predictions. *Journal of Experimental Psychology. Human Perception and Performance*, **35**, 1865–1897.

Simons, J.S., Peers, P.V., Mazuz, Y.S., Berryhill, M.E. & Olson, I.R. (2010). Dissociation between memory accuracy and memory confidence following bilateral parietal lesions. *Cerebral Cortex*, **20**, 479–485.

Smith, C.D., Chebrolu, H., Wekstein, D.R., Schmitt, F.A. & Markesbery, W.R. (2007). Age and gender effects on human brain anatomy: A voxel-based morphometric study in healthy elderly. *Neurobiology of Aging*, **28**, 1075–1087.

Smith, J.D. (2009). The study of animal metacognition. *Trends in Cognitive Sciences*, **13**, 389–396.

Smith, J.D., Shields, W.E. & Washburn, D.A. (2004). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences*, **26**, 317–339.

Smith, J.M. (1982). *Evolution and the theory of games*. Cambridge University Press, Cambridge.

Snodgrass, M., Kalaida, N. & Winer, E.S. (2009). Access is mainly a second-order process: SDT models whether phenomenally (first-order) conscious states are accessed by reflectively (second-order) conscious processes. *Consciousness and Cognition*, **18**, 561–564; discussion 565–567.

Song, C., Kanai, R., Fleming, S., Weil, R., Schwarzkopf, D. & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and Cognition*, in press.

Spiridon, M., Fischl, B. & Kanwisher, N. (2006). Location and spatial profile of category-specific regions in human extrastriate cortex. *Human Brain Mapping*, **27**, 77–89.

Stanford, T.R., Shankar, S., Massoglia, D.P., Costello, M.G. & Salinas, E. (2010). Perceptual decision making in less than 30 milliseconds. *Nature Neuroscience*, **13**, 379–385.

Stehling, M.K., Turner, R. & Mansfield, P. (1991). Echo-planar imaging: magnetic resonance imaging in a fraction of a second. *Science*, **254**, 43.

Steinhauser, M. & Yeung, N. (2010). Decision processes in human performance monitoring. *Journal of Neuroscience*, **30**, 15643–15653.

Sterzer, P., Frith, C. & Petrovic, P. (2008). Believing is seeing: expectations alter visual awareness. *Current Biology*, **18**, R697–698.

Strange, B.A., Duggins, A., Penny, W., Dolan, R.J. & Friston, K.J. (2005). Information theory, novelty and hippocampal responses: unpredicted or unpredictable? *Neural Networks*, **18**, 225–230.

Summerfield, C. & Koechlin, E. (2008). A neural representation of prior information during perceptual inference. *Neuron*, **59**, 336–347.

Summerfield, C. & Koechlin, E. (2010). Economic value biases uncertain perceptual choices in the parietal and prefrontal cortices. *Frontiers in Human Neuroscience*, **4**, 208.

Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J. & Hirsch, J. (2006a). Predictive codes for forthcoming perception in the frontal cortex. *Science*, **314**, 1311–1314.

Summerfield, C., Egner, T., Mangels, J. & Hirsch, J. (2006b). Mistaking a house for a face: neural correlates of misperception in healthy humans. *Cerebral Cortex*, **16**, 500–508.

Sumner, P., Nachev, P., Morris, P., Peters, A.M., Jackson, S.R., Kennard, C. & Husain, M. (2007). Human medial frontal cortex mediates unconscious inhibition of voluntary action. *Neuron*, **54**, 697–711.

Swann, N., Tandon, N., Canolty, R., Ellmore, T.M., McEvoy, L.K., Dreyer, S., DiSano, M. & Aron, A.R. (2009). Intracranial EEG reveals a time- and frequency-specific role for the right inferior frontal gyrus and primary motor cortex in stopping initiated responses. *Journal of Neuroscience*, **29**, 12675–12685.

Swick, D., Ashley, V. & Turken, U. (2008). Left inferior frontal gyrus is critical for response inhibition. *BMC Neuroscience*, **9**, 102.

Thaler, R.H. & Sunstein, C.R. (2009). *Nudge: Improving Decisions About Health, Wealth and Happiness*. Penguin.

Thielscher, A. & Pessoa, L. (2007). Neural correlates of perceptual choice and decision making during fear-disgust discrimination. *Journal of Neuroscience*, **27**, 2908–2917.

Thompson, K.G. & Schall, J.D. (2000). Antecedents and correlates of visual detection and awareness in macaque prefrontal cortex. *Vision Research*, **40**, 1523–1538.

Tosoni, A., Galati, G., Romani, G. & Corbetta, M. (2008). Sensory-motor mechanisms in human parietal cortex underlie arbitrary visual decisions. *Nature Neuroscience*, **11**, 1446–1453.

Trommershauser, J., Maloney, L.T. & Landy, M.S. (2003). Statistical decision theory and the selection of rapid, goal-directed movements. *Journal of the Optical Society of America A: Optics, Image science and Vision*, **20**, 1419–1433.

Tsujimoto, S., Genovesio, A. & Wise, S.P. (2010). Evaluating self-generated decisions in frontal pole cortex of monkeys. *Nature Neuroscience*, **13**, 120–126.

Tuch, D.S., Salat, D.H., Wisco, J.J., Zaleta, A.K., Hevelone, N.D. & Rosas, H.D. (2005). Choice reaction time performance correlates with diffusion anisotropy in white matter pathways supporting visuospatial attention. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 12212–12217.

Tversky, A. & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Economics*, **106**, 1039–1061.

Ullsperger, M., Harsay, H.A., Wessel, J.R. & Ridderinkhof, K.R. (2010). Conscious perception of errors and its relation to the anterior insula. *Brain Structure & Function*, **214**, 629–643.

van den Wildenberg, W., Wylie, S., Forstmann, B., Burle, B., Hasbroucq, T. & Ridderinkhof, K. (2010). To Head or to Heed? Beyond the Surface of Selective Action Inhibition: A Review. *Frontiers in Human Neuroscience*, **4**.

van den Wildenberg, W.P.M., van Boxtel, G.J.M., van der Molen, M.W., Bosch, D.A., Speelman, J.D. & Brunia, C.H.M. (2006). Stimulation of the subthalamic region facilitates the selection and inhibition of motor responses in parkinson's disease. *Journal of Cognitive Neuroscience*, **18**, 626–636.

Verbruggen, F., Aron, A.R., Stevens, M.A. & Chambers, C.D. (2010). Theta burst stimulation dissociates attention and action updating in human inferior frontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 13966–13971.

Vickers, D. (1979). *Decision processes in visual perception*. Academic Press, New York.

Vuilleumier, P. (2004). Anosognosia: The neurology of beliefs and uncertainties. *Cortex*, **40**, 9–17.

Vul, E. & Kanwisher, N. (2010). Begging the question: The non-independence error in fmri data analysis. In S. Hanson & M. Bunzi, eds., *Foundational issues for human brain mapping*, MIT Press.

Wagner, A.D., Poldrack, R.A., Eldridge, L.L., Desmond, J.E. & Glover, G.H. (1998). Material-specific lateralization of prefrontal activation during episodic encoding and retrieval. *NeuroReport*, **9**, 3711.

Wald, A. & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, **19**, 326–339.

Watanabe, M. & Sakagami, M. (2007). Integration of cognitive and motivational context information in the primate prefrontal cortex. *Cerebral Cortex*, **17 Suppl 1**, i101–i109.

Wegner, D.M. (2003). *The illusion of conscious will*. MIT Press.

Weil, R.S., Furl, N., Ruff, C.C., Symmonds, M., Flandin, G., Dolan, R.J., Driver, J. & Rees, G. (2010). Rewarding feedback after correct visual discriminations has both general and specific influences on visual cortex. *Journal of Neurophysiology*, **104**, 1746–1757.

Weiskopf, N. & Helms, G. (2008). Multi-parameter mapping of the human brain at 1mm resolution in less than 20 minutes. In *Proceedings of 16th ISMRM, Toronto, Canada*, 2241.

Weiskopf, N., Hutton, C., Josephs, O. & Deichmann, R. (2006). Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: a whole-brain analysis at 3 t and 1.5 t. *NeuroImage*, **33**, 493–504.

Weiskrantz, L., Warrington, E., Sanders, M. & Marshall, J. (1974). Visual capacity in the hemianopic field following a restricted occipital ablation. *Brain*, **97**, 709–728.

Welch, R.B. & Warren, D.H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, **88**, 638–667.

Wendelken, C., Ditterich, J., Bunge, S.A. & Carter, C.S. (2009). Stimulus and response conflict processing during perceptual decision making. *Cognitive, Affective, & Behavioral Neuroscience*, **9**, 434–447.

Wenke, D., Fleming, S.M. & Haggard, P. (2010). Subliminal priming of actions influences sense of control over effects of action. *Cognition*, **115**, 26–38.

Wenzlaff, H., Bauer, M., Maess, B. & Heekeren, H.R. (2011). Neural characterization of the speed-accuracy tradeoff in a perceptual decision-making task. *Journal of Neuroscience*, **31**, 1254–1266.

Whiteley, L. (2009). *Uncertainty, reward and attention in the Bayesian brain*. Ph.D. thesis, University College London.

Whiteley, L. & Sahani, M. (2008). Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. *Journal of Vision*, **8**, 2–15.

Wichmann, F. & Hill, N. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & Psychophysics.*, **63**, 1293–1313.

Wichmann, T., Bergman, H. & DeLong, M.R. (1994). The primate subthalamic nucleus. i. functional properties in intact animals. *Journal of Neurophysiology*, **72**, 494.

Wilimzig, C., Tsuchiya, N., Fahle, M., Einhäuser, W. & Koch, C. (2008). Spatial attention increases performance but not subjective confidence in a discrimination task. *Journal of Vision*, **8**, 7.1–10.

Woolgar, A., Parr, A., Cusack, R., Thompson, R., Nimmo-Smith, I., Torralva, T., Roca, M., Antoun, N., Manes, F. & Duncan, J. (2010). Fluid intelligence loss linked to restricted regions of damage within frontal and parietal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 14899–14902.

Wunderlich, K., Rangel, A. & O'Doherty, J.P. (2009). Neural computations underlying action-based decision making in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 17199–17204.

Wylie, S.A., van den Wildenberg, W.P.M., Ridderinkhof, K.R., Bashore, T.R., Powell, V.D., Manning, C.A. & Wooten, G.F. (2009). The effect of parkinson's disease on interference control during action selection. *Neuropsychologia*, **47**, 145–157.

Yantis, S., Schwarzbach, J., Serences, J.T., Carlson, R.L., Steinmetz, M.A., Pekar, J.J. & Courtney, S.M. (2002). Transient neural activity in human parietal cortex during spatial attention shifts. *Nature Neuroscience*, **5**, 995–1002.

Yarkoni, T., Barch, D.M., Gray, J.R., Conturo, T.E. & Braver, T.S. (2009). BOLD correlates of trial-by-trial reaction time variability in gray and white matter: a multi-study fMRI analysis. *PloS one*, **4**, 4257.

Yarkoni, T., Poldrack, R.A., Essen, D.C.V. & Wager, T.D. (2010). Cognitive neuroscience 2.0: building a cumulative science of human brain function. *Trends in Cognitive Sciences*, **14**, 489–496.

Yeung, N., Cohen, J.D. & Botvinick, M.M. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological Review*, **111**, 931–959.

Yokoyama, O., Miura, N., Watanabe, J., Takemoto, A., Uchida, S., Sugiura, M., Horie, K., Sato, S., Kawashima, R. & Nakamura, K. (2010). Right frontopolar cortex activity correlates with reliability of retrospective rating of confidence in short-term recognition memory performance. *Neuroscience Research*, in press.

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

Yu, R., Mobbs, D., Seymour, B. & Calder, A.J. (2010). Insula and striatum mediate the default bias. *Journal of Neuroscience*, **30**, 14702–14707.

Zeki, S. & Bartels, A. (1998). The autonomy of the visual systems and the modularity of conscious vision. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **353**, 1911–1914.