Uncertainty, Reward, and Attention in the Bayesian Brain.

Louise Whiteley

Dissertation submitted for the degree of

doctor of philosophy

of the

university of london

Gatsby computational neuroscience unit
university college london

Declaration

I, Louise Emma Whiteley, 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.

17^{th} September 2008

Abstract

The 'Bayesian Coding Hypothesis' formalises the classic Helmholtzian picture of perception as inverse inference, stating that the brain uses Bayes' rule to compute posterior belief distributions over states of the world. There is much behavioural evidence that human observers can behave Bayes-optimally, and there is theoretical work that shows how populations of neurons might perform the underlying computations. There are, however, many remaining questions, three of which are addressed in this thesis. First, we investigate the limits of optimality, demonstrating that observers can correctly integrate an external loss function with their uncertainty about a very simple stimulus, but behave suboptimally with respect to highly complex stimuli. Second, we use the same paradigm in a collaborative fMRI study, asking where along the path from sensory to motor areas a loss function is integrated with sensory uncertainty. Our results suggest that value affects a fronto-striatal action selection network rather than directly impacting on sensory processing. Finally, we consider a major theoretical problem – the demonstrations of optimality that dominate the field have been obtained in tasks with a small number of objects in the focus of attention. When faced instead with a complex scene, the brain can't be Bayes-optimal everywhere. We suggest that a general limitation on the representation of complex posteriors causes the brain to make approximations, which are then locally refined by attention. This framework extends ideas of attention as Bayesian prior, and unifies apparently disparate attentional 'bottlenecks'. We present simulations of three key paradigms, and discuss how such modelling could be extended to more detailed, neurally inspired settings. Broadening the Bayesian picture of perception and strengthening its connection to neuroscientific and psychological literatures is critical to its future as a comprehensive theory of neural inference, and the thesis concludes with a brief discussion of future challenges in this direction.

Contents

List of Figures

List of Tables

Preface

Acknowledgements

It seems to be the inevitable lot of the new PhD student to scoff at tales of adversity imparted by battle-scarred final years, only to find themselves immersed in those very trials – the second year blues, the upgrade frustrations, the null result. I am extremely lucky to have gone through these rites of academic passage in a department, and with a supervisor, who have provided exceptional support. Being around so many talented scientists, with a range of approaches, interests, and backgrounds has greatly enriched my perspective on the brain and on the paradigmatic wranglings of the relatively novel neurosciences. My supervisor, Maneesh Sahani, has given me both freedom and guidance, and has been instrumental in the development of my thinking. His comments and suggestions never fail to be insightful, and his ability to combine technical expertise with philosophical thinking presents an inspiring example. To acknowledge his involvement in every aspect of this thesis I will use the first person plural throughout.

I have learnt from too many people at Gatsby to mention them all, but would like to thank Richard Turner and Misha Ahrens in particular – office mates, good friends, and much-valued sounding boards. I would also like to thank Stephen Fleming for his interest in our work on Bayesian decision making, and for exciting discussions and methodological guidance during the fMRI collaboration that followed. Throughout my PhD, the other students on the Wellcome Trust PhD program – Hanneke Den Ouden, Rosemary Milton, David Barker, Kieran Boyle, and Curtis Asante, and my fellow East-Enders – Lucy Neville and Sarah Brunell, have provided mutual encouragement and lots of fun. To my family, thanks for the ongoing support you have always shown, and for giving me from my earliest years a love of books and learning. Last but certainly not least, my partner Ollie Hulme. From our first meetings at 'Consciousness Club' to recent co-authorship, he has been a first port of call for matters academic and beyond, and has been an amazing source of inspiration and support in all my endeavours.

For reading part or all of this thesis in various forms, I would like to thank Peter Dayan, Josh Solomon, Richard Turner, Oliver Hulme and Stephen Fleming.

Collaborations and Contributions

The work reported in Chapters 4 and 5 was the result of a collaboration between myself and Dr Maneesh Sahani at the Gatsby Unit, Mr Stephen Fleming, Prof. Ray Dolan, and Prof. Chris Frith at the Wellcome Trust Centre for Neuroimaging, and Dr Oliver Hulme at the Institute for Ophthalmology. The project was led by Stephen Fleming as part of his PhD, and was based on the paradigm we present in Chapter 3. I contributed to adapting the experimental design for use in the scanner, and to its development based on the analysis of pilot behavioural data. Stephen Fleming collected the behavioural and fMRI data, and I coded psychometric and optimality analysis of the behavioural data. I then contributed to the design and interpretation of the fMRI analysis, and co-wrote a paper on the imaging results for submission to *Science*.

Publications and Other Work During the PhD

Chapter 3 has been published in *Journal of Vision* (Whiteley and Sahani, 2008). A paper based on Chapters 4 and 5 has been submitted to *Science*, and Chapter 6 is in preparation for submission to *Psychological Review*. During my PhD I also contributed to two fMRI studies investigating the subcortical basis of salience computation in the human brain, both in preparation. I co-wrote a commentary on a target article by Ned Block in *Behavioural and Brain Sciences* (Hulme and Whiteley, 2007), and a paper based on work done prior to my PhD *Acta Psychologica* (Whiteley et al., 2008).

1

Introduction

1.1 Introducing Bayesian Inference

Bayes' rule is a simple equation with a very complicated life. It was first presented in the 18$^{\text{th}}$ century by the Reverend Thomas Bayes as a solution to the 'inverse probability' problem central to the rather unholy pursuits of gambling and insurance[1] (Bayes, 1764). This problem occurs whenever we have to work backwards from an observation (or 'data') to a belief about the state of the world that generated it. For example, imagine a farmer has two varieties of tomato seed, one of which tends to produce much bigger fruits. If the labels on the bags of seeds were lost, he might decide to plant 20 seeds from each bag to work out which contained the larger variety. However, observing that the 20 tomatoes grown from the first bag were on average larger than the 20 grown from the second bag should not make him 100% confident that the first bag was the one he was after - it could have been a fluke. According to Bayes' rule, making an inference about a state of the world (here, seed type) from an observation (here, the size of tomatoes in two samples) requires a **likelihood** model of how observations are generated, and **prior** beliefs about the state of the world that generated them, which are used to compute a **posterior** belief distribution according to probability theory:

$$p\left(state\ of\ world \,|\, data\right) = \frac{p\left(data \,|\, state\ of\ world\right)\ p\left(state\ of\ world\right)}{p\left(data\right)} \tag{1.1}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \tag{1.2}$$

[1] When hurricane Barbara strikes Springfield, Marge Simpson reassures the long-suffering wife of evangelical Ned Flanders that insurance will cover their damaged house, and Maude replies "Uh, well, no. Neddy doesn't believe in insurance. He considers it a form of gambling."

The Bayesian farmer could use this equation to compute the posterior probability of each bag containing the larger seed variety. The likelihoods might embody knowledge such as the typical size of each tomato variety, and perhaps information about variability – for example, the larger variety might also yield a greater range of sizes. The prior might embody a bias, for example a suspicion based on where the bags were stored that the second bag contained the larger seed variety, and the stronger the prior, the more impact it has on the posterior. As can be seen in this simple example, despite its perhaps off-putting mathematical formulation, Bayesian reasoning embodies many intuitions about how we should combine information in arriving at a belief.

The denominator in Bayes' rule can be treated simply as a normalisation constant, functioning to ensure that the posterior probability distribution adds to one. It also measures the agreement between the likelihood and the prior, providing **evidence** for the choice of model structure independent of parameter settings (this quantity is also known as the 'marginal likelihood'). In order to compare competing models, the evidence term then plays the role of the likelihood for each model in a further application of Bayes' rule (Mackay, 2004). If our tomato farmer suspected that one bag in fact contained a mixture of the two seed varieties, he could compare the marginal likelihood of a model that predicts clusters of tomatoes around the average size of each variety, against the marginal likelihood for the single variety model. The posterior constitutes a full representation of the degree of belief about each possible state of the world, but in many scenarios a specific estimate is required. The state of the world that gives the highest value of the likelihood is known as the *maximum likelihood* or 'ML' estimate, and the state of the world that is accorded the highest probability by the posterior is known as the *maximum a posteriori* or 'MAP' estimate. With an uninformative prior, which is insensitive to reparameterisation, the two are equivalent (see e.g. Jaynes, 2003).

1.2 Application and Epistemology

The applications of Bayesian inference were considered by Bayes and Laplace in the 18^{th} – 19^{th} centuries (Dale, 1982), but its full potential was not realised until the late 20^{th} century, when machine learning techniques and computational power allowed priors and likelihoods of real world complexity to be used (see Fienberg, 2006). In the meantime, alternative approaches to statistical inference were developed, leading to philosophical arguments about the meaning of probability itself. The frequentist position prominent in the first half of the 20^{th} century constitutes a complementary, non-Bayesian framework for statistical inference and hypothesis-testing, differing from the Bayesian approach in the absence of priors and in

a focus on punctate statistics rather than degrees of belief (see Wonnacott and Wonnacott, 1990, for a description of both approaches). Behind these concrete differences lies a deeper disagreement about what a probability *is* – the frequentist treats probabilities as the relative frequency of each possible outcome in an infinite number of repetitions of a well-defined random experiment, whereas the Bayesian thinks of probability as a subjective degree of belief that can be assigned to any proposition[2]. To come back to our Bayesian farmer, if he wanted to repeat the tomato experiment, it would be hard to ensure that the conditions were identical. Scientists deal with such issues on a daily basis, designing well-controlled experiments according to accepted principles. However, the philosophical concept of an infinitely repeatable random experiment is rather strange, reflecting the difficulty with a human observer ever having access to this hypothetical scenario. If we configure the same scenario in terms of the beliefs of the farmer, this conceptual discomfort is ameliorated, though of course at the risk of sacrificing some degree of observer-independent 'objectivity'.

This epistemological debate continued somewhat tangentially to the booming development of statistical inference and hypothesis-testing methods in the latter half of the 20$^{\text{th}}$ century. Bayesian methods had huge success, and viable frequentist alternatives for many complex settings failed to materialise (Fienberg, 2006). However, the debate about subjectivity did not disappear – it migrated to questions about where the likelihood and prior in Bayesian inference *come from*. Bayes' rule is simply a theorem that tells you how to update your beliefs in the light of evidence – it does not tell you how to choose good priors, or how relevant data is to your hypothesis. When Bayes (1764) and later Laplace (1840) presented the theorem it was with flat or 'uninformative' priors, avoiding the issue of subjective prior beliefs. More recently, objective Bayesians have allowed priors, but insisted that they be derived from evidence, whilst subjective Bayesians are committed to the utility of probabilistic inference even in the face of subjectively biased priors – it is often pointed out that after enough incoming evidence, differences in prior beliefs will be largely ironed out. For most scientists, such questions are important for selecting and interpreting inference methods appropriate to particular settings, rather than for selecting an overarching epistemology from which all else will follow.

1.3 Bayes and the Brain

In this thesis we argue that the Bayesian approach is the right one for thinking about inverse inference in the brain. The brain is supplied with noisy sensory evidence, and must

[2]Although some of the knowledge embodied in Bayesian likelihoods and priors may of course have originated in the observation of relative frequencies.

work backwards to the underlying cause of that evidence, be it a physical principle of the universe, the intentions of another person, the tree in front of your eyes, or the firing of downstream neurons. Bayes' rule provides a way to formalise such inference, and does so in a way that is both deeply intuitive and mathematically sound. Indeed, it was shown by Cox (1961) that starting from simple common-sense desiderata such as consistency, the basic tenets of probability theory imply that Bayes' rule is the only correct way to reason about degrees of belief (see Jaynes, 2003).

Bayes rule can be applied to inverse inference in many different contexts – the details will differ, but the fundamental principles are the same. In this thesis we claim that it is the right way to think about both behavioural and neural inference, via two interlinked hypotheses. First, the **Bayesian Brain Hypothesis** (BBH) states that Bayes' rule provides an accurate and normative framework for understanding human behaviour. Second, the **Bayesian Coding Hypothesis** (BCH) states that the brain actually implements the computations implied by Bayesian descriptions of behaviour. In both settings, Bayes' rule forms the core of *computational models* of the brain – models that seek not just to describe a particular neural operation, but to understand its function (see Dayan, 1994). David Marr famously described computational modelling in terms of three levels of analysis: the **computational** level describes the goal of a particular system or operation, the **algorithmic** level describes strategies for achieving that goal, and the **implementational** level identifies the underlying substrate of the algorithm (Marr, 1982). In the remainder of the introduction we will unpack our two hypotheses, and consider how they fit into this framework. The evidence for Bayesian behaviour, and Bayesian neural coding, will be considered in detail in the literature review.

1.3.1 The Bayesian Brain Hypothesis

Viewing visual perception as noisy, ill-posed inverse inference (Green and Swets, 1966; Kersten et al., 2004; Knill and Richards, 1996) long precedes Bayes, perhaps in part because it can be investigated through introspection. Aristotle's shadowy images of the world presaged the realisation that neural processing is characterised by uncertainty, and Al Hazen's 11^{th} century treatise on optics made explicit the idea that the brain inverts the evidence of the sense organ to reveal a picture of the causative world, emphasising that vision involves judgement rather than "pure sensation". In the 19^{th} century Helmholtz (1925) and Mach (1980) developed a related theory of vision that informed modern experimental psychology and continues to dominate the field in various guises.

So what reason do we have for believing that the inverse inferences of perception are ill-posed and noisy? In order for an inverse inference to be *well*-posed, there must be a one-to-one mapping between data and its cause – i.e. every possible state of the world must evoke a distinct pattern of neural firing. This is highly unlikely. First, the world is three-dimensional (3D), and the retina forms a two-dimensional (2D) sheet of sensory transducers, with depth computations occurring later in the system. Second, the causes we infer are not unique states of the world – they generalise over these states by categorising patterns of photons into features and objects. This suggests that firing patterns cannot be mapped neatly and directly onto identifiable states of the world.

On a more biological level, the mapping from a state of the world to a neural firing pattern is one-to-many (see Glimcher, 2005; Faisal et al., 2008). This last contribution to ill-posedness is also a key source of variability (Tolhurst et al., 1983), exacerbated by the imperfect transmission of physical energy from the world to the sensory epithelia (Geisler, 1989), and amplified as neural signals are passed around the brain. Whether or not action potentials are truly random at the microscopic or even quantum level need not trouble us here – at the neuronal level there is abundant evidence for variability, despite controversy about whether this should be characterised as 'noise' (see e.g. Averbeck et al., 2006; Ma et al., 2006; Stein et al., 2005). And although perception might seem introspectively deterministic, illusions, mistakes, and a multitude of perceptual phenomena that shock the suddenly enlightened speak to a flexible and constructive process. The accuracy of our judgements is also often affected by context, such as in colour-constancy phenomena (Maloney, 1999) and the Müller-Lyer illusion (Dragoi and Lockhead, 1999). Not only are there multiple possible interpretations of sensory evidence, our knowledge and expectations about the world can change what we perceive (Hansen et al., 2006; Stocker and Simoncelli, 2006a) – our brain, if not our eyes, can indeed deceive us.

If the mapping from sensory firing to the state of the world that caused it is indeed ill-posed, we need to take the resulting uncertainty into account when making inverse inferences. Bayes' rule provides an optimal prescription for doing so (see Equations 1.1–1.2), both for single inferences and for the combination and ongoing assimilation of information in a distributed, dynamic system like the brain. Figure 1.1 illustrates the Bayesian view of perception: given data $\mathbf{s}$, the brain computes a posterior belief or 'recognition model' over the possible states of the world $\mathbf{m}$. The posterior is proportional to the product of the likelihood or 'generative model', which describes how states of the world evoke sensory firing, and the prior belief over those states of the world.

However, computing a posterior is not the end of the story – the brain is not a passive perceiver, and perceptual accuracy is not the only thing it is trying to optimise – perception drives decisions and actions in a world of costs and rewards. Bayes' rule is therefore part-

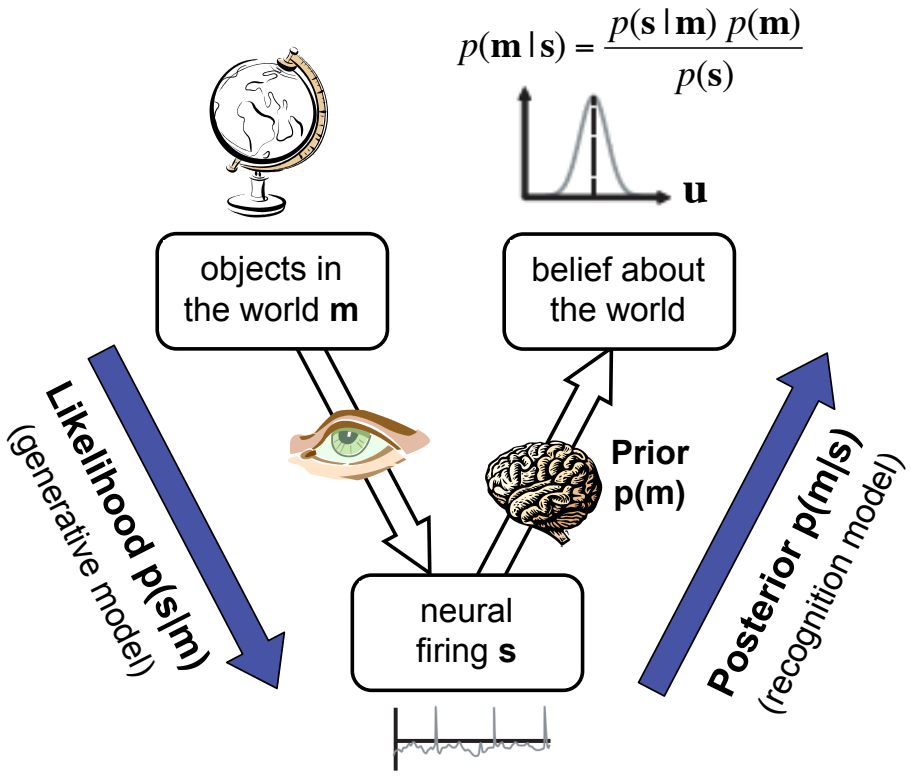

Figure 1.1: ***Schematic of Bayesian inverse inference.*** *This simple schematic illustrates the Bayesian view of perception: a generative model ('likelihood') maps the state of the world* $\mathbf{m}$ *to sensory firing* $\mathbf{s}$*, and the brain learns to use a recognition model ('posterior') which takes existing beliefs ('priors') into account when inverting the generative model to make inferences about the state of the world. In order for the acting brain to utilise posterior beliefs, they are combined with loss functions according to decision-theoretic principles, as formalised in Equation 1.3 and Figure 1.2 (see Kording, 2007).*

nered with decision theory, as formalised in Figure 1.2, in order to describe how posterior beliefs about the outcomes, o_j, that follow from a particular decision, d_i, can be combined with a 'loss function' that gives the distribution of costs and rewards for each outcome to yield an expected utility (EU) for each decision (Kording, 2007; Cox, 1961; Berger, 1985). This quantity is computing by summing over all possible outcomes the probability of that outcome multiplied by its utility;

$$\mathbf{E}\left[U(d_i)\right] = \sum_j p(o_j \,|\, d_i)\, U(o_j) \tag{1.3}$$

The decision with the maximum utility, $\hat{d}$, is then selected and executed via a motor action m, resulting in a particular outcome $\hat{o}$ with utility $U(\hat{o})$. To give an real world example,

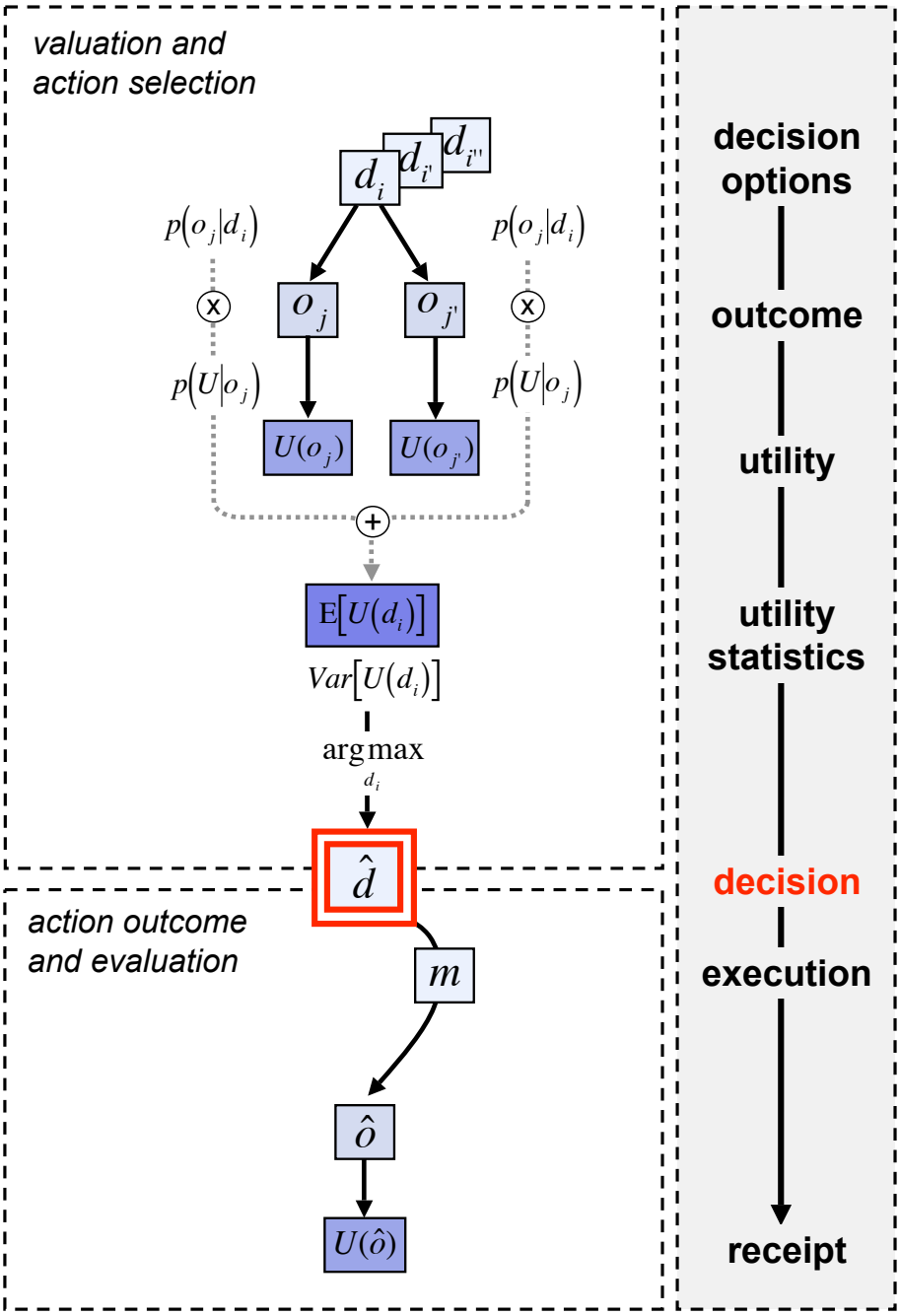

Figure 1.2: ***Formalising the Bayesian decision.*** *A decision $\hat{d}$ is selected by comparing the expected utility of different options; $\mathbf{E}[U(d_i)]$. Expected utility is computed by summing, over all possible outcomes, the product of the mapping from decision to outcome; $p(o_j|d_i)$, and from outcome to utility; $p(U|o_j)$. The decision $\hat{d}$ is then executed via a motor action m, and the animal receives an actual outcome $\hat{o}$ with utility $U(\hat{o})$. See text for details.*

imagine a referee trying to call a difficult penalty in the face of a vociferous home crowd - in making a decision, his posterior belief over where the ball hit the pitch might be unconsciously combined with the respective cost and reward for calling it either way. As illustrated by this example, when making a categorical decision on the basis of a posterior belief (here, deciding whether the ball was over the line or not), many states of the world will be subsumed under each possible decision and mapped onto each possible outcome. Thus $p(o_j \mid d_i)$ can be thought of as marginalising over states of the world $\mathbf{m}$;

$$p(o_j \mid d_i) = \sum_{\mathbf{m}_k} p(o_j \mid d_i, \mathbf{m}_k)\, p(\mathbf{m}_k) \tag{1.4}$$

As indicated in Figure 1.2, the loss function can also be stochastic, expressed in a distribution over possible utilities for each outcome; $p(U \mid o_j)$, which would also have to be marginalised in the EU computation.

What goes into the loss function ($U(o_j)$ in Equation 1.3) depends on what kind of decision is at stake – the decision could be a motor command (deciding how to move), a cognitive state (deciding what to think), or a perception ('deciding' what to see). There are semantic arguments about whether a perception should strictly be called a decision, but for our purposes an operational definition is sufficient, and a decision constitutes selection between a number of potential outcomes. The nature of the loss function and the relation of Bayesian decision making to classical neurobiological approaches will be considered in more detail in Section 2.3 of the literature review. Some critics of the Bayesian approach have argued that if we have to collapse sensory uncertainty into commands for a small number of motor effectors, surely there is no reason to carry around belief distributions rather than point estimates? But this only holds water if sensory processing consists of a single posterior over the motor command, or if all distributions are Gaussians of fixed variance. In a distributed, hierarchical system the optimal commands for even a single effector are obtained by integrating and updating information according to Bayes' rule as sensory input is transformed into a motor output (see e.g. Friston, 2005; Lee and Mumford, 2003; Rao and Ballard, 1999).

A Bayesian approach has the potential to unify apparently disparate models of perception and decision-making under a single framework (see e.g. Dayan and Daw, 2008). An important example, which will be relevant throughout the thesis, is the relationship of the BBH to Signal detection theory (SDT; Green and Swets, 1966). Signal detection approaches were initially developed during the Second World War, as a response to the problem of separating signal from noise in communication technology, but led to a revolution in the formalisation of Helmholtzian perceptual inference. In the SDT framework, a perceptual decision-maker asks which of a number of distributions, each generated by a

different cause, an internal datum came from – the canonical example is a detection task in which an observer must decide if a signal is present in noise. Proponents of SDT have emphasised its ability to separate sensitivity, which is dictated by the width and separation of the distributions, from bias, which is independently determined by the threshold for determining which distribution a datum came from. Modern Bayesian analyses are the probabilistic, inferential analogue of the SDT observer, extending the language of single distributions produced by particular stimulus classes to posterior probability distributions over continuously valued stimulus dimensions. When combined with decision theory (Kording, 2007; Cox, 1961; Berger, 1985; Jaynes, 2003), Bayes' rule provides a normative description of how to integrate uncertainty with loss functions to maximise the value of perceptual or motor 'decisions' – roughly analogous to setting an optimal SDT threshold.

The BBH states that Bayes' rule (see Equations 1.1–1.2), coupled with decision theory (see Equation 1.3), provides a normative and explanatory description of behaviour. Evidence for the BBH comes from showing that performance in perceptual and motor tasks follows Bayes rule, and that this performance can not be explained by equivalent non-Bayesian strategies (see Knill and Pouget, 2004, for a summary), and will be discussed in Section 2.1 of the literature review. In Marrian terms, we might say that the **computational** goal of behaviour is to maximise expected utility as laid out in Equation 1.3. The BBH states that the **algorithm** for doing so is to combine perceptual uncertainty with utility, rather than learning rigid decision-utility pairs over time[3]. In searching for behavioural evidence for the BBH, we therefore look for signatures of the flexible combination of uncertainty with loss functions, and for the flexible combination of different belief distributions that contribute to the posterior (see Section 2.1). The distinction between computation and algorithm is of course flexible, and implementational constraints can provide important bottom-up information – working in a purely top-down way is rarely a good strategy, and can place too much weight on the theorist's conception of computational problems and possible algorithmic formalisms.

1.3.2 The Bayesian Coding Hypothesis

The BCH extends the BBH by claiming that not only is a Bayesian algorithm expressed in behaviour, but that it is implemented by neural populations that explicitly represent and compute with probability distributions. What we mean by this is that we can point to particular properties of a neural population that correspond to elements of a probability

[3]This corresponds to the distinction between 'model-based' and 'model-free' learning strategies in studies of decision making – see Section 2.3.

distribution, via a neural coding model of the mapping between the two (see Knill and Pouget, 2004; Doya et al., 2007) – i.e. that $p(\mathbf{m} \mid \mathbf{s})$ is represented by identifiable neural populations[4]. An 'implicit' implementation produces the same, Bayesian output, but via a mechanism that does not have identifiable neural correlates of the components of $p(o_j \mid d_i)$.

According to deCharms and Zador (2000), "A neural code is a system of rules and mechanisms by which a signal carries information." Determining the neural code is one of the grand challenges of neuroscience, asking how trains of action potentials encode sensory stimulation, and how they can be decoded to reveal something about the state of the world that evoked them. Many theoretical studies have focused on early cortical areas, where input comes from the sensory epithelia, and neural responses are well studied, usually involving simple feature analysers (see Doya et al., 2007). For a hierarchical, distributed cortical system, the question of what a population encodes, and how its representations are interpreted by other populations and manifested in perception, is fraught with difficulties. Probabilistic methods have played an important part both in providing technical tools for investigating encoding and decoding models, and in providing a conceptual framework for thinking about what neurons represent (see Doya et al., 2007; Dayan and Abbott, 2001).

The BCH grew out of arguments for the benefits of probabilistic encoding – computing with distributions rather than point estimates improves flexibility and robustness, allowing the integration of different sources of information to be informed by the uncertainty inherent in each (see Barber et al., 2003; Zemel et al., 1998). Carrying around posteriors also prevents the brain from committing to one interpretation too early in processing. These arguments were supported by observations that existing coding models could be interpreted in Bayesian terms. First, it became clear that influential neural network models such as the Hopfield net (Mackay, 2004) could be reinterpreted as implementing approximate Bayesian inference on analog quantities (Anderson and Abrahams, 1987). Second, attempts to decode externally presented stimuli from trains of spikes in simple systems were reinterpreted in probabilistic terms, or extended to probabilistic settings (see Pouget et al., 2003)

Following these developments, Anderson (1994) and Zemel et al. (1998) proposed that biologically inspired neural networks should be thought of as representing probability density functions (pdfs) over variables, and performing statistical inference over these representations. This prompted a flurry of studies demonstrating how populations and networks of simple model neurons might implement Bayesian inference in simple settings (Ma et al.,

[4]There is of course a danger that "identifiable" is a property of our limited knowledge, and the line between a distributed neural correlate and a conjunctive list of all kinds of activity associated with a particular Bayesian quantity, can be slippery. These considerations should temper conclusions drawn from the success of particular implementational schemes, but do not prevent a useful distinction being made.

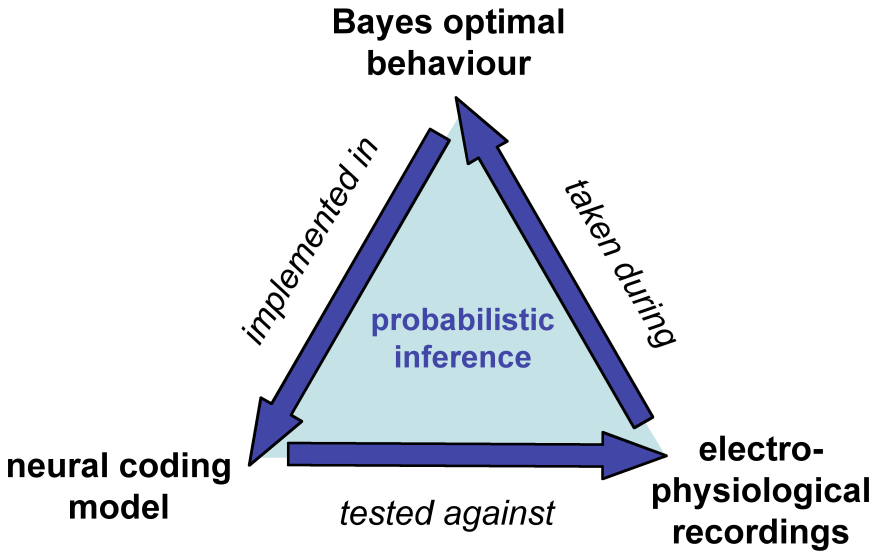

Figure 1.3: ***Triangulating evidence for the BCH.*** *Finding direct experimental evidence for the BCH is complex, and involves a triangulation of methodologies, centered on a Bayesian description of a particular inference. Ideally, electrophysiological recordings taken whilst an animal performs Bayes optimal inference can be compared to the predictions of neural coding models built to implement the inference implied by the behaviour.*

2006; Pouget et al., 2003, 2000; Eliasmith and Anderson, 2003; Deneve et al., 2001; Rao, 2004a), discussed in Section 2.2.1. This work is complemented by observations that certain aspects of cortical architecture and neural firing statistics are well suited to multilayer Bayesian inference (e.g. Friston, 2003, 2005) and inference in time (e.g. Deneve, 2008a,b; Huys et al., 2007), reviewed in Section 2.2.3.

To provide stronger evidence for the BCH, it is critical to link neural coding models to electrophysiological recordings from neurons thought to implement the relevant computations. The smoking gun would be a series of experiments which show Bayesian inference in behaviour, build neural coding models that implement the inference implied by the behaviour, and then test the predictions of this model against electrophysiological recordings from neurons thought to implement it. This integrative trinity is illustrated in Figure 1.3. However, this is an under-constrained problem for all but the earliest of cortical areas, and the evidence reviewed in Section 2.2.4 consists of piecemeal correspondences rather than watertight tests of competing Bayesian and non-Bayesian theories of neural computation (e.g. Deneve et al., 2001; Rao, 2004a; Huys et al., 2007; Rowland et al., 2007; Yu and Dayan, 2005; Anastasio et al., 2000; Gold and Shadlen, 2001).

We might want to characterise the BCH as saying that the brain **implements** a Bayesian **algorithm**. However, it can again be difficult to draw a clean separation between Marrian levels – physiological variables can be described in varying degrees of detail, and neural coding models that map such variables to parameters of probability distributions will be shaped by the language of the algorithm. When we bring all three levels together the picture gets even more complicated. If we are persuaded by the BBH that a Bayesian algorithm is the right one for solving the computational problems of perception, then the brain must implement that algorithm – all that is left is to say *how*. But the BCH says more than this – it does not argue for one particular implementation, it says that representation of, and computation with, probability distributions is the right way to think about many neural operations – this is both an implementational and to some degree an algorithmic statement. In gathering evidence for these hypotheses, the interdependence of different levels of analysis is thrown into even sharper relief. Although the BBH can be stated and investigated independently of the BCH, implementational constraints can usefully inform Bayesian characterisations of behaviour. And the BCH cannot be disentangled from the behaviour that follows from the neural operations it proposes – evidence for the BCH therefore consists in tightening all the links illustrated in Figure 1.3, with each recurrently informing and constraining the other [5].

Most behavioural evidence has been acquired in simple scenarios where performance can be Bayes optimal. The neural coding models built to map Bayesian variables to neural quantities correspondingly deal in simple belief distributions over single features, and animal behaviour during electrophysiological recordings also follows suit (see Knill and Pouget, 2004; Doya et al., 2007). This simplicity provides a good substrate for integrative methodology, and for mapping the scenarios in which behaviour can be Bayes optimal – below we use a simple behavioural task to consider whether performance can be optimal with regard to very simple and very complex unimodal visual categorisations.

For the BCH to be more widely applicable as a theory of neural inference, we need to consider more complex scenarios where performance and computation may not be optimal. This raises a number of tricky questions about what it means to have suboptimal, or 'approximate' Bayesian inference. In machine learning, approximate Bayesian techniques are commonly used to solve highly complex inferential problems (see e.g. Mackay, 2004; Minka, 2001). These are not thought of as changing the semantics of Bayesian inference – we still talk of likelihoods, priors, and posteriors – but rather as performing a full Bayesian computation as closely as possible, retaining the core properties of representing belief distributions and accumulating evidence via the machinery of Bayes' rule. However, deviations from op-

[5] Below we will thus sometimes refer to behavioural experiments that bolster the claims of the BBH as providing evidence *for the BCH*, even though the BCH is superficially a statement about neural coding.

timality can be very hard to interpret on the basis of behaviour alone – they could be due to limits on neural processing, due to interference from other operations, or due to a failure of the experimenter to properly characterise the ecological prior the participant brings to the experiment. Having a clear conception of what an approximate algorithm looks like, informed by implementational constraints, makes it far easier to distinguish behaviour that is approximately Bayesian from that which instead embodies a non-Bayesian algorithm.

As the Bayesian formalism is applied to more complex settings, the need to integrate it with relevant neuroscientific and psychological literatures will also become more pressing. Measuring behavioural performance in settings that invoke attention, reward, memory, and other cognitive functions will render the mapping of an isolated sensory posterior to neurons in the area thought to represent it inadequate. An important first step is to consider how Bayesian decision theory relates to research in decision-making more generally – below we consider how the neuroanatomical bases and theoretical concepts might overlap. Dealing with more complex perceptual decisions might also necessitate approximate representation of posterior distributions, and approximate algorithmic solutions to full Bayesian computations. Below we consider how this might relate to classic ideas of neural capacity limits, and how traditional ideas of attentional selection acting to allocate a limited capacity resource could be configured in a Bayesian framework.

1.4 Organisation of the Thesis

The thesis begins with a literature review, considering evidence for the BCH and relevant aspects of the psychology and neuroscience of decision-making and attentional selection (**Chapter 2**). Subsequent chapters report work on three unanswered questions:

1. First, we investigate the limits of Bayes-optimal behaviour, demonstrating psychophysically that uncertainty about a very simple visual quantity is integrated with reward information to yield optimal perceptual decisions (**Chapter 3**), but that uncertainty about complex stimulus categorisation is not (**Chapter 4**).

2. Second, in collaboration with Mr Stephen Fleming, Prof. Ray Dolan, Prof. Chris Frith, and Dr. Oliver Hulme, we use an fMRI version of the same paradigm to ask where along the path from sensory to motor areas changing the external value landscape impacts on the formation of a decision. We discuss how this adds to our understanding of the neurobiology of value-based vs. perceptual decision making, and the implications for the implementation of Bayesian decision making in a hierarchical, recurrent neural architecture (**Chapter 5**).

3. Third, we consider a major theoretical problem: when faced with a complex scene the brain cannot be Bayes-optimal everywhere. We characterise a general limitation in representing complex posteriors, and suggest how attentional selection might refine the impoverished representations that result. We discuss how this reciprocally informs and unifies some of the classic literature on attentional selection, and extends previous approaches to attention as a Bayesian prior, and as performing noise reduction in a signal detection framework (**Chapter 6**).

The final chapter summarises the contributions of this thesis, and identifies common themes in the three questions that are approached. Here we also consider future work, and briefly discuss the potential of the BBH as a comprehensive theory of perceptual inference, and the potential of the BCH as a theory of its neural substrate (**Chapter 7**).

2

Literature Review

In this thesis we address three outstanding questions for the Bayesian Coding Hypothesis. To situate this work, we will first review the experimental literature supporting the claim that the inverse inferences of perception can be Bayes optimal, and that decisions about perception and action can maximise value in the face of uncertainty and variably desirable outcomes. The limitations of this body of work will point the way to the psychophysics studies reported in Chapter 3 and 4. The BCH states that the probabilistic inferences that describe optimal behaviour are in fact explicitly implemented in the brain, and is supported by theoretical models of how neural populations might represent and compute with probability distributions, and by showing that these models are consistent with neurophysiological data. This work will also be reviewed below, and the limitations that parallel those in the behavioural literature will point the way to the theoretical work reported in Chapter 6. In Chapter 5 we ask about the anatomical implementation of components of a Bayesian decision, addressing unanswered questions about the interface between value-based and perceptual decision making, and in Chapter 6 we take a Bayesian approach to the psychological literature on attentional selection. We therefore also review the neural basis of decision making and models of attentional selection, highlighting issues not currently addressed by the BCH, and also identifying where a Bayesian approach might reciprocally enrich the literature.

2.1 Behavioural Evidence for Bayesian Inference

The most extensive source of evidence that Bayesian computation is the right way to think about what the brain is doing in perception and decision making comes from showing that Bayesian models explain behavioural data better than alternative accounts. This argues specifically for the BBH (see Section 1.3.1), and is an important component of the integrative evidence required for the BCH (see Section 1.3.2). Behavioural evidence comes in a number of forms (see Knill and Pouget, 2004, for a review). First, there are demonstrations that observers implicitly adjust cue weights in a Bayes optimal fashion given viewing or stimulus parameters (see Section 2.1.1), and open questions concerning how cue combination depends on judgements about a common cause (see Section 2.1.2). Second, tasks have been designed to show that perception and action take account of uncertainties in sensory and motor variables (see Section 2.1.3). Third, there is evidence that perceptual biases and illusions, which are often unexplained, can be accounted for by ecologically valid priors, even when the prior doesn't explicitly represent the bias to be explained (see Section 2.1.4). Below we will summarise the evidence in each of these domains, discussing issues of interpretation that arise from the flexibility of the Bayesian formalism and from failures of optimality.

2.1.1 Cue Combination

In many everyday scenarios we have multiple sources of information about an object or event, for example when watching someone's lips move whilst listening to them speak. In such scenarios, the various 'cues' to a particular perception should be combined with weights determined by their reliability. The most common example of this is visual capture – in general, vision is much more reliable than hearing for spatial localisation, and so when we have reason to believe that a visual and auditory stimulus were caused by the same object we will perceive the location of that object as heavily biased towards that suggested by the visual information, for example in the ventriloquist illusion (Pick et al., 1969; Welch and Warren, 1980; Bertelson and Radeau, 1981). Recently, these classic psychological effects have been considered from the perspective of optimal integration of probabilistic information, which provides a flexible and accurate description of cue combination within modalities (e.g. Jacobs, 1999; Knill and Saunders, 2003; Hillis et al., 2004; Knill, 1998), and across modalities (e.g. van Beers et al., 1999; Ernst and Banks, 2002; Battaglia et al., 2003; Helbig and Ernst, 2007b). Indeed, Alais and Burr (2004) showed that by degrading visual information, classic visual capture could be reversed in favour of the usually less accurate auditory signal.

The simplest model of Bayesian cue combination assumes that an underlying quantity x evokes a stochastic estimate ξ for each cue, which is distributed as a Gaussian likelihood centered on the true value and with a width, σ^2, that reflects the cue reliability;

$$p(\xi|x) \;=\; \mathcal{N}(\xi; x, \sigma^2) \tag{2.1}$$

If there is no prior information about x, the posterior is simply proportional to the likelihood, and is a Gaussian with the same variance, but centered on the cue estimate ξ;

$$p(x|\xi) \;=\; \mathcal{N}(x; \xi, \sigma^2) \tag{2.2}$$

For a single cue the optimal MAP estimate (see page 15) of x is thus given by the mean of the posterior; ξ. When there is more than one cue, the optimal estimate is a weighted combination of the estimates for each cue, where the weights depend straightforwardly on the relative posterior variances. For example, if ξ_V and ξ_H are the MAP estimates from visual and haptic posteriors over the size of an object, and σ_V and σ_H are the respective widths of the two posteriors, the optimal estimate $\hat{\xi}$ is given by;

$$\hat{\xi} = \frac{\frac{1}{\sigma_V^2}}{\frac{1}{\sigma_V^2} + \frac{1}{\sigma_H^2}}\,\xi_V \;+\; \frac{\frac{1}{\sigma_H^2}}{\frac{1}{\sigma_V^2} + \frac{1}{\sigma_H^2}}\,\xi_H \tag{2.3}$$

There is much evidence that people can conform to this model – Jacobs (1999) found that different cues to depth were optimally weighted by their reliabilities, and Ernst and Banks (2002) demonstrated a similar, crossmodal effect for haptic and visual cues to the size of a virtual object. However, there are some scenarios in which a simple linear combinations of weighted cues is inappropriate. For example, there are at least twenty cues to depth perception and some cannot support individual estimates. Here, a model in which non-linear additions to Equation 2.3 alter the effect of different cues has been used to model behavioural data (Fine and Jacobs, 1999) – in general, it is a challenge for the BBH to move away from simple Gaussian likelihoods that lead to simple, highly tractable optimal algorithms.

In moving from Bayesian descriptions of behaviour to arguments that the neural substrate explicitly represents and computes with the implied probability distributions (i.e. that ξ is a neural variable), it is crucial to consider whether *non*-Bayesian strategies could mimic the observed behaviour. Implementing Equation 2.3 involves just linear computations, and it has been argued that the brain could just learn the optimal weights for particular viewing parameters and stimulus properties that covary with cue uncertainty, and use these weights to combine point estimates (see Knill and Pouget, 2004; Jacobs, 2002). But this learning is prohibitive when you consider the variety of scenarios that might require different sets of weights, and the problem of labelling scenarios reliably. Using probability distributions is

more flexible and can accommodate unexpected changes – it has been shown that people can retain optimality in the face of changes in feedback, cue reliability or prior distributions (Atkins et al., 2001; Ernst, 2007; Knill, 2007). There is also evidence for behavioural optimality in non-linear systems, where it is necessary to compute with the full posterior (Knill, 2003). For example, Saunders and Knill (2001) presented a Bayesian model that accounts for tilt-related biases in the perception of 3D surface orientation by optimally integrating a highly non-Gaussian likelihood function over skew symmetry with stereoscopic information about orientation. However, the difficulty of deriving a non-Bayesian algorithm here does not rule out the existence of such a scheme, and conclusions about the plausibility of implementation must be made with care.

So far we have assumed that the brain can 'read off' the reliabilities of different cues from neural representations of the posterior distributions they invoke – i.e. that there is an explicit neural representation of $p(\mathbf{m} \,|\, \mathbf{s})$ for each cue. Whether this is the case for novel cues or stimulus dimensions, as well as for conventional cues to well-established sensory quantities, is an important question (see Chapter 4). Recently, Michel and Jacobs (2008) found that observers were sensitive to the reliability of arbitrary low-level features they had never seen before. The authors constructed stimuli out of linear combinations of 20 visual features or 'basis vectors', then trained observers to discriminate between two prototypes corrupted by Gaussian noise. A 'classification image' was derived that showed observers' weighting of the different features moving towards optimal during the experiment. This demonstrates the ubiquity of computations with uncertainty, strengthening evidence for the BBH. However, using this as evidence for the *neural* representation of posteriors, from which uncertainty signals can be read, is more difficult – the feedback provided to help observers learn the categorisation task could support an adaptive strategy that only implicitly represents information contained in the posterior. Showing that optimal processing of uncertainty can proceed *without* feedback-related learning is thus critical to arguing for posterior representations in the brain (see Chapter 3, and Whiteley and Sahani (2008)). However, this does not imply that any scenario in which learning about uncertainties takes place has a non-Bayesian substrate, and indeed learning paradigms offer important constraints on our understanding of the kinds of representations that support cue combination.

Understanding how cue combination develops during childhood is another important source of information. A recent study by Gori et al. (2008) found that the optimal integration of visual and haptic cues to size and orientation is first observed between the ages of eight and ten, but it is not clear why this should be so. Gori et al. (2008) suggest that plasticity during development makes it advantageous to preserve individual sensory signals for efficient recalibration as different systems change. However, adult studies show optimal cue combination in the face of long- and short-term plasticity or environmental changes

(e.g. Atkins et al., 2001; Ernst, 2007; Knill, 2007), and there is no reason why individual sensory signals might not be retained alongside a combined estimate. Indeed, Müller et al. (2007) showed that when cues to surface slant are combined, information about each can still be used. Ernst (2008) have suggested that perhaps the problem for children comes not from an inability to combine cues cross-modally, but rather from a difficulty with calibrating their perceptual decision-making to match task demands. In the task used by Gori et al. (2008) the visual and haptic cues were presented in different apparent locations, and whilst adults can tolerate this discrepancy if informed that both cues come from the same object (Helbig and Ernst, 2007a,b), perhaps young children cannot. Experiments by Atkins and colleagues (Atkins et al., 2001, 2003) have demonstrated that observers can adjust their assessment of cue reliability by comparing the correlation of a novel cue with one whose reliability is known. This is not necessarily incompatible with a Bayesian approach, perhaps being expressed in the prior, or in an external signal similar to a loss function.

2.1.2 Binding and the Common Cause

Cue combination takes place when there are multiple sources of information about the same object, and for it to be non-trivial, these sources must point in contradictory directions. But if they contradict each other too much, the observer might infer that they don't in fact have the same underlying cause, which changes how an estimate about some feature of the supposed common cause should be made. It has long been known that cue combination effects depend on the spatial and temporal separation between cues (Bertelson and Radeau, 1981; Gepshtein et al., 2005; Lewald et al., 2001; Lewald and Guski, 2003), and with large discrepancies cues often cease to influence each other. However, there is contradictory evidence about when a common cause is inferred, when inference is conditioned on such judgements, and how cognitively penetrable or automatic cue combination is in different scenarios.

Roach et al. (2006) asked people to judge either visual or auditory rate whilst ignoring conflicting information in the other modality, and observed a gradual transition between partial integration and total cue segregation as inter-modal discrepancy was increased. They modelled this result via a prior that expressed knowledge about the typical correspondence between auditory and visual rate signals, effectively overruling instructions that one signal was irrelevant when uncertainty was great enough. This suggests a certain degree of automaticity for crossmodal integration, but other studies have found greater flexibility. Hillis et al. (2002) showed that when visual and haptic cues to object size were provided, participants could access both combined and unimodal estimates, and Helbig and Ernst (2007a) showed that explicit instructions about object identity enhanced the integration of

crossmodal cues, implying that the observer's prior was cognitively penetrable. According to Bayes' rule, prior information about cue reliability or bias, for example due to development, tool use, or alterations in the environment, should be taken into account in deciding whether cues carry usefully redundant information about a common cause (see Ernst, 2008; Bresciani et al., 2006; Ernst, 2005). But the prior is unlikely to reflect just experimental manipulations – as the study by Roach et al. (2006) suggests, there are some prior biases that are hard to overturn. Delineating the various contributions to the Bayesian prior, and how they might compete for influence, is an important open question.

Another open question is whether categorical judgements *about* a common cause influence estimation according to Bayesian principles. Kording et al. (2007) and Sato et al. (2007) showed that a mixture model, in which observers optimally integrate over posteriors conditioned both on common and separate causes, provides a good model of location judgements for a visuo-auditory event. But it is unclear how widespread this kind of impartial consideration of both causal hypotheses might be. Stocker and Simoncelli (2008) modelled simple perceptual judgements reported by Jazayeri and Movshon (2007), in which observers were asked to categorise a direction-of-motion stimulus as lying either side of a boundary, before judging its exact direction. Judgements of motion direction showed a repulsion away from the boundary, and were well-described by a Bayesian model in which inference was conditioned on the observer's earlier categorical decision. Although this is not a cue combination experiment, it raises relevant questions – when is inference conditioned on a higher-level or previous judgement, as opposed to integrating over or different possible scenarios? Stocker and Simoncelli (2008) suggest that although conditioning on a previous categorical judgement leads to suboptimal estimation, the brain benefits from the attendant self-consistency, evoking psychological notions of cognitive dissonance (Festinger, 1957).

This discussion illustrates the problem with moving away from simple paradigms in which Bayes optimal performance is well-defined, and easily achieved. If performance does not match a simple Bayesian model, it can be hard to decide whether inference is optimal under a biased prior or constraint not included in the model, or whether it is suboptimal. And if inference is suboptimal, it can be even harder to distinguish approximate Bayesian inference from a non-Bayesian solution. The parameters of Bayes' rule are dangerously underconstrained, and it is important to avoid the mistake of thinking that all behaviour must be evolutionarily optimal, and therefore that we can work backwards from observed biases to define the underlying likelihoods and priors. To avoid this problem, experimenters usually encourage observers to leave behind any prior expectations or biases, hoping to reveal

Bayesian computation in a restricted model of a particular task.[1] When this doesn't work, constraints from multiple levels can be used to guide the development of Bayesian models – measuring parameters such as uncertainty or prior bias directly from behaviour can provide important constraints (e.g. Whiteley and Sahani, 2008; Stocker and Simoncelli, 2006a), and considering implementational constraints from neural properties or evolutionary arguments is also key. As discussed in Section 1.3.2, defining approximate Bayesian inference is difficult from behaviour alone – comparing machine learning models that perform approximate Bayesian computation to behaviour, and to relevant electrophysiological data via neural coding models, is an important future challenge (see Chapter 6 for further discussion).

Instructions about a common cause, or abstract knowledge of cue contingencies, constitute 'top-down' signals that might be expressed in a Bayesian prior. Another important source of top-down information is *attention*, which according to influential feature-integration (FIT) theories (Treisman and Gelade, 1980), is required for features represented in different cortical regions to be 'bound' together. Experimental work by Bertelson et al. (2000) and Vroomen et al. (2001) suggested that neither exogenous nor endogenous attention affect the binding of discrepant visual and auditory cues. However, their task involved a very simple perceptual judgement, and it may be that attention is required only when the binding at stake is difficult – not in order to perform relatively simple judgements about a common cause (see Wolfe et al., 1989, for an extension of FIT that allows limited pre-attentive binding). In Section 2.4 of the literature review we discuss the relationship between binding and attention in more detail, which helps to contextualise the Bayesian framework for attentional selection reported in Chapter 6, in which attention acts much like a Bayesian prior that can serve to improve judgements about co-located features.

2.1.3 Sensorimotor Integration

As described in Section 1.3.1, perception drives action – even in purely perceptual studies, observers usually report a discrete decision about a perceived object via motoric actions such as speech or button presses. The fullest version of the Bayesian paradigm therefore casts statistical problems in the framework of decision making, where posterior beliefs are combined with loss functions to compute expected utilities for competing decision options (see Kording, 2007). As well as motor actions being used to report and implement decisions, sensorimotor integration, in which sensory feedback is integrated with knowledge

[1] This problem is of course not unique to the Bayesian analysis of behaviour – studies of consciousness in particular suffer from the need to override an observer's expectations about what the experimenter wants them to do in what is often a very unnatural environment.

of the motor system in order to plan and guide movements, is a domain of noisy, ill-posed inference in its own right. Mirroring the noisy, ill-posed nature of perceptual inference, motor neurons and effectors are themselves stochastic, and deciding how to implement a particular motor goal is again an under-constrained problem (Ghahramani et al., 1995). Although this thesis focuses on Bayesian perception, evidence for Bayesian inference in the motor system is an important contributor to the BCH and raises issues of general relevance.

There are two main classes of evidence for sensorimotor Bayesian inference. The first involves showing that people take into account the uncertainties inherent in their motor system when planning movements – much like they take into account the uncertainties inherent in their perceptual system when deciding how to combine and update information from different sources (Bays and Wolpert, 2007). Behavioural studies have shown that the extent to which people use visual feedback to make online corrections to reaching movements depends on visual and motor uncertainty in a Bayesian way, and that optimal behaviour is preserved in the face of manipulations to feedback (Saunders and Knill, 2001, 2003; Kording and Wolpert, 2004). Estimating the position of the body is a sensory estimation task, and various sources of information about posture can be combined according to Bayes' rule (Clarke and Yuille, 1990; Ghahramani and Wolpert, 1997). In these studies, the loss function simply expresses the need for accuracy, but the output is not just an estimate of position – the posterior provides sensory feedback that is used to update motor commands. Models combining optimal perceptual estimators with control systems that compare predictive signals from an internal model to sensorily registered outcomes (see Wolpert et al., 1995) have provided a framework for thinking about movement optimality (see Todorov and Jordan, 2002; Todorov, 2005; van Beers et al., 2002), but there is ongoing debate about whether this is the right way to approach the motor system (see Guigon et al., 2008).

Another strategy for investigating these questions has been to manipulate external loss functions for the endpoint of a movement, and show that they can be combined optimally with information about the intrinsic uncertainty of the motor system to maximise reward (Tassinari et al., 2006; Trommershauser et al., 2006, 2005, 2003b). To give an intuitive example, if you are trying to throw a ball to a friend who is standing on the other side of a river, there is a higher penalty for undershooting than overshooting, so unless you're an amazing shot you should probably aim past her open hands. Now if I alter the loss function by filling the river with crocodiles, you'll probably risk a little more effort on her part and aim even further away. This kind of motor task provides more direct evidence for the BBH, as the complexities of online adjustment are subsumed in a simple estimate of endpoint variability that dictates an optimal mean endpoint. Linking these results to neural data, providing stronger evidence for the BCH, would require the identification of a neural population coding for movement endpoint, perhaps in regions known to reflect the

upcoming endpoint of a saccadic eye movement such as the frontal eye fields. The problem with subsuming motor computations into a single endpoint variable is that this neural population might not be doing the work of Bayes' rule – key representations of uncertainty might be elsewhere.

In real world perceptual judgements and actions, determining the loss function is rarely straightforward. In terms of behavioural evidence for the BCH, most perceptual tasks assume that accuracy is the only 'reward' that is optimised. However, for motor control the loss function is far more complex. First, it is not always obvious how to measure accuracy – for example, low variance in movement endpoints might be more important than the exact location of the mean (see Harris and Wolpert, 1998; Bays and Wolpert, 2007). Second, accuracy is not the only gain to be optimised – for a biological system, limits on quantities such as energy expended, speed, or flexibility of movements might constitute equally stringent constraints. Again we come up against the question about how to determine the constraints on components of a Bayesian decision-theoretic computation, and how to decide whether computations are optimal under constraints or simply suboptimal. Again, experimental approaches to measuring the relevant quantities directly are promising (see e.g. Kording et al., 2004), and implementational constraints from the motor system are important. It is clear that Bayesian approaches have shed some light on sensorimotor computations, and analyses of behaviour suggest that uncertainties in sensory and motor variables can be optimally taken into account in determining and updating motor plans. As potential contributors to integrative evidence for the BCH, sensorimotor paradigms are generally less easily linked to simple neural variables than their perceptual cousins, but the flip-side of this is that more complex correspondences are potentially more persuasive.

2.1.4 Priors and Illusions

Above, we touched on how information contained in the prior can bias decisions away from the ML estimate. Showing that priors affect judgements in a Bayes optimal way, and that their influence depends on the sensory uncertainty embodied in the likelihood, is another source of evidence for the BBH. Weiss et al. (2002) modelled motion perception with a likelihood function arising from the integration of noisy local cues, and a prior reflecting the predominance of slow speeds in the environment. The key idea is that as the width of the likelihood increases (for example via decreasing stimulus contrast), a biased prior will have a greater effect on the mode of the posterior. This model captures a remarkable number of motion illusions and phenomena that had previously resisted principled explanation, via the complex pattern of directional biases that arise in the Bayesian posterior as contrast, edge orientation, and other stimulus factors change. Recently, Welchman et al. (2008) reported

behavioural results suggesting that this same biased prior could explain surprising failures in the ability to estimate whether an approaching object would hit a target.

Another famous set of illusions is those involving tilt perception, a subset of which can be interpreted as a natural consequence of Bayesian inference in a model with a smoothness prior (Schwartz et al., 2006). Just as multiple sensory cues can be combined, multiple prior constraints can exist for the same quantity, as demonstrated for priors over light source and surface orientation for depth judgements by Mamassian and Landy (2001). Rather than explaining a static perceptual bias, van Ee et al. (2003) showed that introducing prior assumptions about the shape and orientation of objects in a scene can predict bistability in perceived slant.

A big challenge is to show that the priors used to explain biases and illusions are well constrained, rather than acting as free parameters. Even for cases such as the distribution of speeds, or the direction of light sources, where it is clear what the *shape* of an ecological prior should be, quantitative constraints from behaviour are important. Stocker and Simoncelli (2006a) modelled speed discrimination along the lines of the Weiss et al. (2002) but inferred the shape of the prior probability as well as the internal noise characteristics directly from psychophysical data, and used the resulting model to account for trial-by-trial variation. This issue does not only apply to determining priors – Stocker and Simoncelli (2006b) noted that repulsion effects following adaptation to a tilted grating are the opposite of what would be predicted by a prior biased towards the adapting stimulus. They instead modelled these effects via a local increase in the signal-to-noise ratio (SNR). This provided a neat explanation of the data, and appeals to neural data (Tranchina et al., 1984; Barlow, 1990), but illustrates the potential malleability of under-constrained Bayesian models - if allowed any manipulation to the prior or likelihood that fits the data, we can 'explain' any illusion as Bayes optimal.

2.2 Neural Coding Models for Bayesian Inference

In the introduction we argued for an integrative approach to gathering evidence for the neural implementation of Bayesian inference, illustrated in Figure 1.3. An important part of this integrative trinity is linking the kind of behavioual data discussed above to electrophysiological recordings taken from neurons thought to implement the implied computations. To do so, we need neural coding models that map between the two (e.g. Deneve et al., 2001; Rao, 2004a; Huys et al., 2007; Rowland et al., 2007; Yu and Dayan, 2005; Anastasio et al., 2000; Gold and Shadlen, 2001). Below, we review the various probabilistic population codes (PPCs) that have been proposed to play this role, and consider the kinds of distributions

they can encode, and the operations they can perform on the resulting representations (see Section 2.2.1). These codes represent posterior distributions over simple, static quantities, and cannot distinguish the presence of multiple objects from uncertainty over the parameters of a single object. We therefore consider extensions to population codes that can distinguish multiplicity from uncertainty (see Section 2.2.2), and also extensions to inference with dynamic stimuli and in hierarchical architectures (see Section 2.2.3), considering the issues these developments raise in terms of approximate inference.

2.2.1 Probabilistic Population Codes

The simplest probabilistic population code (PPC) is perhaps one that represents the likelihood ratio – the relative evidence for one of two possible explanations – for example, the ratio of the likelihoods that a random dot kinematogram (RDK) contains motion in an upward vs. a downward direction given the displacements of the constituent dots; $\mathbf{z}$:

$$\frac{p\left(\text{up} \mid \mathbf{z}\right)}{p\left(\text{down} \mid \mathbf{z}\right)} \tag{2.4}$$

This ratio can be represented with two populations of neurons, each responding preferentially to one of the two stimuli, or with one population whose firing represents the likelihood ratio between the two stimuli. Single cell recordings in the primate lateral intraparietal area (LIP), thought to play a role in integrating sensory evidence and in forming eye movement commands, have provided evidence for both strategies. When a monkey was trained to perform one of two possible saccades, some LIP neurons responded proportionally to the probability that the saccade would end in their receptive field (RF) (Platt and Glimcher, 1999). In a motion discrimination task like that described above, the activity of some LIP neurons reflects integration over time, consistent with the representation of a constantly-updated log likelihood ratio (Gold and Shadlen, 2001). Responses in the superior colliculus have also been interpreted in this way (Anastasio et al., 2000). Recording studies in which a single neuron reflects the log likelihood, or log likelihood ratio, raise the possibility that the 'populations' that code these simple probabilistic quantities could be very small indeed. However, population codes have important properties such as noise resistance and flexibility, and it is a mistake to interpret the firing of a single cell in too homuncular a way.

These studies provide suggestive evidence for the BCH, especially when considered in conjunction with theoretical work demonstrating a correspondence between unit responses in a recurrent neural network performing Bayesian inference and LIP recordings (Rao, 2004b). We will revisit these studies later, when considering work on the neural basis of decision-making (Section 2.3.3), but interpreting neural activities in terms of likelihood ra-

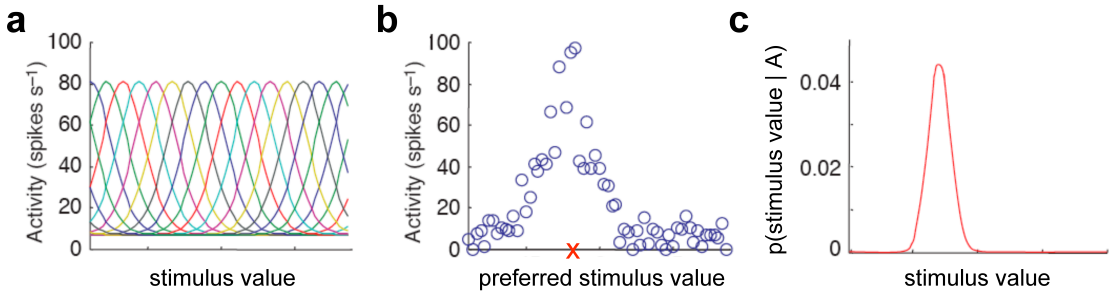

Figure 2.1: ***Probabilistic population codes.*** ***a,*** *Idealized Gaussian tuning curves showing a hypothetical group of neurons with different preferred values along the stimulus axis.* ***b,*** *Exemplar responses of 64 Gaussian-tuned neurons with preferred stimulus values indicated on the x-axis, in response to the stimulus value indicated with the red cross.* ***c,*** *the posterior distribution obtained by applying a Bayesian decoder to the noisy hill of population activity shown in* ***b****. In a gain encoding scheme, assuming independent Poisson noise means that the peak of the posterior distribution is given by the peak of the noisy hill shown in* ***b****, and the width of the distribution (i.e. the uncertainty) is inversely proportional its amplitude (adapted from Deneve et al. (2001)).*

tios does not solve the problem of how neurons can encode full pdfs and attendant measures of uncertainty. Two approaches to encoding full probability distributions over continuously valued variables are convolution codes and gain encoding schemes (see Knill and Pouget, 2004). Both are based on the observation that neurons have tuning curves – that they fire to a range of stimulus values but peak at a particular preferred value (see Figure 2.1a). When considered from the Bayesian perspective, it is tempting to interpret the neuron's firing as reflecting the probability that the stimulus takes the neuron's preferred value, and to suggest that the firing of a population of neurons with different preferred values (see Figure 2.1b) could represent a full pdf (see Figure 2.1c). Convolution and gain codes both formalise this intuition.

In a convolution code, each neuron's firing represents the dot product of its tuning curve with the pdf to be encoded. If the tuning curves of the population are simply displaced Gaussian functions with their means at a full range of 'preferred values' along the stimulus continuum (as in Figure 2.1a), the population activity represents a set of samples from the pdf smoothed with a Gaussian kernel (Zemel et al., 1998; Anderson, 1994). Computations with pdfs represented in this way, for example multiplying two likelihoods for motion and stereoscopic depth cues with a prior to obtain a posterior over depth, are relatively straightforward. For Dirac delta tuning functions (i.e. where each neuron responds *only* to its preferred value), point-by-point multiplications suffice (see Ernst and Banks, 2002;

Zemel and Dayan, 1997). These operations can be extended for more realistic tuning curves (Zemel and Dayan, 1997; Barber et al., 2003) and to dynamic variables (Anderson, 1994; Eliasmith and Anderson, 2003), but there is a general question mark over the biological plausibility of the product operation (though see Koch, 1994). Interpreting firing rates in terms of log probabilities instead converts this operation to a more readily implemented addition (Rao, 2004b), consistent with observations of how LIP neurons appear to integrate evidence over time (Gold and Shadlen, 2001).

Gain encoding schemes (Pouget et al., 2003; Deneve et al., 2001) are closely related to convolution codes, but take advantage of the near-Poisson nature of neural noise (Tolhurst et al., 1983) to encode the mean and variance (i.e. uncertainty) of a Gaussian distribution simultaneously. In such a code, each neuron i responds to a stimulus x with a firing rate r_i, whose mean value is given by its Gaussian tuning curve $f_i(x)$. The tuning curve peaks at the neuron's preferred value, x_i, and in a simple PPC all tuning curves have the same width σ^2 (as illustrated in Figure 2.1a);

$$f_i(x) \;=\; \mathbf{e}^{-\frac{(x-x_i)^2}{2\,\sigma^2}} \tag{2.5}$$

The tuning curve is then corrupted by noise, and if this follows Poisson statistics, the distribution of firing rates r_i for a single neuron is given by;

$$p\,(r_i \,|\, x) \;=\; \mathbf{e}^{-f_i(x)}\;\frac{(f_i(x))^{r_i}}{r_i\,!} \tag{2.6}$$

For a population of such neurons, whose tuning functions span the possible values of x, the firing rates in the population when a stimulus is presented will, when lined up according to each neurons' preferred value x_i, look like a noisy belief distribution (see Figure 2.1b for an example). A Bayesian decoder can be used to return the posterior distribution over stimulus value from these population activities (see Figure 2.1c and Sanger, 1996; Foldiak, 1993), which if the neurons are treated as independent is;

$$p\,(x \,|\, \mathbf{r}) \;\propto\; p(x)\;\prod_i\, p\,(r_i \,|\, x) \tag{2.7}$$

This posterior is a Gaussian whose mean and variance are defined very simply;

$$p\,(x \,|\, \mathbf{r}) \;=\; \mathcal{N}\left(x\,;\;\frac{\sum_i x_i\,r_i}{\sum_i r_i},\;\frac{1}{\sum_i r_i}\right) \tag{2.8}$$

The mean of the posterior is thus controlled mostly by the peak of the 'hill' of activities shown in Figure 2.1b, and the variance is inversely proportional to its gain or amplitude (Pouget et al., 2003). This result is due to fact that for Poisson noise, the variance of

the spike count is proportional to the gain, and thus to the SNR. A simple example of a neural computation where this might be implemented is orientation tuning in striate cortex (V1), as neurons are known to have bell-shaped tuning curves centered on all values for orientation (Hubel and Wiesel, 1962).

A recent paper by Ma et al. (2006) extended this observation about decoding a single posterior from a probabilistic population code with Poisson noise, and showed that a broad class of operations required for Bayesian inference reduce to simple linear combinations of population activities. They demonstrated further that these results hold for arbitrary probability distributions over the stimulus, for tuning curves of arbitrary shape and for realistic neuronal variability. This neurally plausible, comprehensive coding scheme for simple operations is an important starting point for understanding the neural implementation of perception as Bayesian inference. However, there are a number of constraints on the properties of the code and the noise involved that are controversial, as are interpretations of the gain-encoding scheme that reframe neural noise as useful rather than nuisance (see Ma et al., 2006). In order to rigorously assess whether these PPCs are a good description of what real neurons represent, it is critical not just to imbue them with biological plausibility, but to compare them against competing codes. At the moment, there are limited propositions for what neural firing represents in the Bayesian world – likelihood ratios, probabilities (or samples from probability distributions), sufficient statistics, and log probabilities being the main contenders – and these models do not make distinct enough predictions in complex enough domains to be able to distinguish between them experimentally.

2.2.2 Complexity in Population Codes

Even though recent coding models such as that of Ma et al. (2006) have shown wide applicability to different computations, there is a serious question about how they can be extended to complex, real world scenes. The pdfs considered in these codes tend to describe single objects, corresponding to the single objects usually considered in behavioural optimality tasks. As illustrated in Figure 2.2, the posterior over a simple lab task such as determining the orientation of a grating is really a distribution over the orientation *of a particular, labelled stimulus.* For a real world scene, the true posterior is a huge joint distribution over a multitude of features at a multitude of spatial locations and, critically, where these features do not belong to discretely labelled objects.

There are two issues here that probabilistic coding models need to address. First, how large joint posteriors over multiple correlated features could be represented. The failure of neural coding models to represent this kind of scenario mimics the conceptual failure

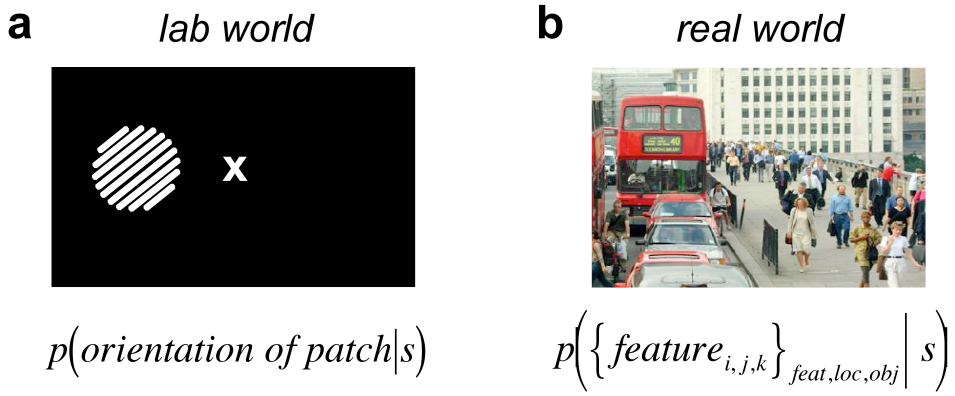

Figure 2.2: ***There's more than one object in the world.*** ***a,*** *A typical lab-based task use to demonstrate Bayesian optimal inference, where one or a small number of objects are present, and the relevant posterior is therefore over the value of a particular labelled feature given sensory firing.* ***b,*** *A real world London scene properly described with a joint posterior over a multitude of potentially correlated features, where multiple objects take values on a single feature dimension, and where a single object produces feature values on multiple dimensions. This represents problems of multiplicity and binding not addressed by existing probabilistic population codes (see Sahani and Dayan, 2003).*

of the BCH to consider the complexity limits of both representation and inference (see Section 2.4 below and Chapter 6 for further discussion). The second, related issue is how distributions over a single dimension can distinguish between multiplicity and uncertainty. For example, a bimodal posterior belief distribution over orientation which has two peaks – one centered on $+45°$ and one centered on $-45°$ – could correspond to two orientations being present at the same time (**multiplicity**), or uncertainty about which of the two was present (**uncertainty**) (see Figure 2.3). In order to extend the traditional PPC to this scenario, we would have to replace the multiplicity case with two probability distributions over orientation, one for each of the two objects. However, resolving which features belong to which 'objects' (i.e. solving the binding problem) is the *job* of the belief distribution, and should not be expressed within its notation.

Sahani and Dayan (2003) present a notation that can represent both multiplicity and uncertainty, where instead of probability distributions being defined over features; $p(\theta)$, they are defined over *multiplicity functions* of features; $p(m(\theta))$. Multiplicity functions; $m(\theta)$, can be thought of as specific proposals about the state of the world, and consist of delta or step functions at the proposed feature values. For the example in Figure 2.3, a world in which both orientations are present would be represented by an $m(\theta)$ that had a delta function at both $+45°$ and $-45°$, and a world in which only one of those orientations

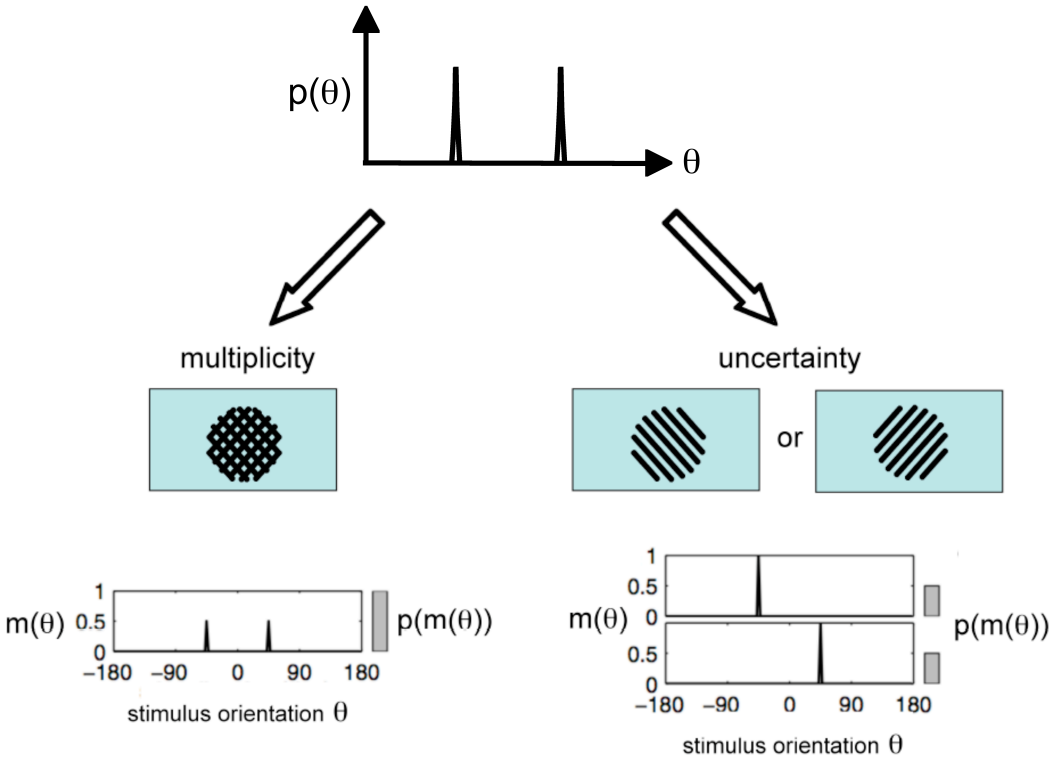

Figure 2.3: ***Distinguishing multiplicity and uncertainty.*** *A posterior belief with two peaks cannot distinguish between two objects being present with different orientations (**multiplicity**) and uncertainty about the orientation of a single object (**uncertainty**). Multiplicity functions represent proposals about the state of the world, and probability distributions over these functions can express multiplicity, uncertainty, or absence. In the context of a neural representation, multiplicity functions can be thought of as feature maps (adapted from Sahani and Dayan, 2003).*

is present would be represented by an $m(\theta)$ with a single delta function at the appropriate orientation. The probability distribution *over* the multiplicity functions, $p\left(m\left(\theta\right)\right)$, can then encompass both multiplicity (where all probability mass is given to the bimodal $m(\theta)$) and uncertainty (where the probability mass is divided between the two unimodal $m(\theta)$). These two alternatives are illustrated by the gray bars to the side of the multiplicity functions in the lower panel of Figure 2.3.

Moving beyond this simple example, there are three more key aspects of the multiplicity function notation that are crucial to its ability to represent complex scenes. First, whilst the locations of the delta peaks in $m(\theta)$ indicate the proposed values of the feature, the magnitude of $m(\theta)$ represents the 'strength' or amount of the proposed features. With an orientation patch, the location of the delta peak would correspond to the proposed

orientation, and the magnitude of $m(\theta)$ would correspond to something like contrast. For other features, contrast is replaced with an appropriate quantity such as motion coherence or colour saturation, and for features that are either present or absent $m(\theta)$ takes values of 0 or 1. Second, we need to be able to represent uncertainty in feature values as well as in the number of objects. In the multiplicity function notation illustrated in Figure 2.3 there is uncertainty about the number of objects present but no doubt that the relevant orientations were $+45^\circ$ and -45°. However, there is often additional uncertainty over feature values, which would be represented by a distribution over a *collection* of multiplicity functions with the location of their delta peaks displaced around the true values. Third, a multiplicity function can be empty, and $p(m(\theta))$ can therefore represent the probability of nothing being present, which is not possible in a probability distribution that must sum to 1.

A standard pdf cannot distinguish multiplicity and uncertainty, and a standard probabilistic population code likewise cannot tell them apart as it has only a single mechanism for representing multiple values (see Zemel and Dayan, 1999). Sahani and Dayan (2003) present a 'doubly distributional population code' (DDPC) which shows how distributions over multiplicity functions could be encoded in, and decoded from, populations of rate-coding neurons. The mechanism for encoding multiplicity is the same as that used in the standard PPC, and is distinct from the way in which uncertainty over multiplicity functions is encoded. The former intuitively corresponds to the addition of firing rates from multiple signals, and the latter intuitively corresponds to a scenario in which teacher signals would enable a system to learn to associate a single signal with multiple possible explanations. In general, this provides a framework for encoding distributions over *functions*, which greatly enhances the descriptive power of population coding approaches, but is yet to be applied to encoding and decoding posteriors of real world complexity.

2.2.3 Temporal, Hierarchical, and Network Extensions

The probabilistic (Zemel et al., 1998), and doubly distributional (Sahani and Dayan, 2003), population codes described above represent static encoding and decoding models in which, although information can be passed between neural populations (e.g. Zemel and Dayan, 1999; Yang and Zemel, 2000), these interactions are not dynamic. To improve the evidence these theoretical approaches provide for the implementational possibilities of the BCH it is important to show how populations of neurons in interacting, hierarchical networks can implement inference through the dynamic evolution of activity, and for information that itself evolves in time. To date, there has been some work on each of these elements, but this is ongoing and will evolve alongside relevant machine learning techniques.

Deneve et al. (2001) built an attractor network showing how the activity of individual populations representing unimodal likelihoods via a gain-encoding PPC scheme could be combined to return the MAP solution to the 'cue combination' problem (see Section 2.1.1). The initial activity of these populations reflects their uncertainty, due to the properties of Poisson noise discussed above (Section 2.2.1), which weights the drive each population provides as a separate pool of integrative neurons settles on a solution (Latham et al., 2003). The authors applied this biologically plausible network to the problem of object localisation given eye and head-centered signals, assumed to be solved in parietal cortex, but it can be extended to any cue combination scenario (e.g. Jacobs, 1999; Knill and Saunders, 2003; Hillis et al., 2004; van Beers et al., 1999; Ernst and Banks, 2002; Battaglia et al., 2003; Alais and Burr, 2004) and also to time-varying problems such as the sensorimotor estimation of the position of a moving arm.

This model provides an intuitive, biologically plausible mechanism for generic 'cue integration', but the attractor structure does not permit multimodal distributions of activity and thus cannot deal with multiplicity. Extending this work to the DDPC framework would present a powerful tool for inference in neural networks, but presents a serious challenge in terms of the representational language of the code. Deneve et al. (2001) used a set of joint basis functions over the individual population activities to map them into a common reference frame – extending this to joint basis functions over multiplicity functions would lead to a prohibitive explosion in the size of the basis set (see Sahani and Dayan, 2003). In another example, (Rao, 2004b) built a neural network model inspired by the architecture of the cerebral cortex that could implement Bayesian inference for a hidden Markov model. When he applied this to random-dot-kinematogram stimuli he found that responses of the model neurons mimicked those of evidence-accumulating neurons in primate LIP and frontal eye fields (FEF).

Dynamic network implementations of particular inferences acknowledge the constantly evolving nature of neural activity, but the world is constantly changing too. This is of course a central tenet of experimental work, where analyses of how individual spikes contribute to the representation of rapidly varying stimuli (Bialek et al., 1991; Reinagel and Reid, 2000; Johansson and Birznieks, 2004) and how fast-timescale spiking might contribute to population coding (Wilson and McNaughton, 1993; Schwartz, 1994; Zhang et al., 1998; Brown et al., 1998) raise important questions for population rate codes. Recently, complementary theoretical work has considered how Bayesian inference over continuously changing stimuli might be implemented in spiking neurons. Huys et al. (2007) and Natarajan et al. (2008) extended the PPC approach to perform maximum likelihood inference through time, but

found that for the simplest encoder, the correlations induced by smooth stimuli[2] mean that decoding requires information that is non-local in time and distributed across neurons. As decoding in time serves as a proxy for computation, this lack of tractability and biological plausibility is problematic. Huys et al. (2007) present an alternative encoder in which spikes contribute independent information and are independently decodable. This corresponds to treating each spike as an independent expert in a product of experts decoder (Hinton, 1999), and has comparable computational power.

In related work, Deneve (2008a) showed that the dynamics of spiking neurons can be interpreted as a form of Bayesian inference in time, where each neuron represents the probability of a binary variable, and where spikes represent new information not predicted from past activity. The model neurons in this scheme optimally integrate dynamic information, and exhibit biologically plausible properties – they are similar to leaky integrate-and-fire neurons, employ spike-dependent adaptation, and maximally respond to fluctuations of their input. In a companion paper, Deneve (2008b) showed how a network of such neurons could learn hierarchical causal models of the sensory input in a biologically plausible way, maximising information transfer whilst minimising energetically costly spikes.

The idea that the perceiving brain encodes 'prediction errors' – reflecting a difference between what was expected and what occurred – dates back to MacKay (1956), and can be applied to the brain on many different levels of analysis, from Deneve's single neurons to learning paradigms in which associations between stimuli and rewards are driven by errors in the prediction of reward (e.g. Montague et al., 1996; Hare et al., 2008, and Section 2.3). Bayesian inference through time can be interpreted as a Kalman filter (Kalman, 1960; Kalman and Bucy, 1961; Bryson and Ho, 1975), in which the current input is combined with an internal generative model to provide a prediction for the next input, whose error is then used to update the generative model ready for the next prediction (see Rao, 1999; Rao and Ballard, 1997). In hierarchical models, predictions in one layer constitute empirical priors for Bayesian inference in the next (Friston, 2003, 2005; Rao and Ballard, 1999), a algorithm that has also been described in terms of belief propagation and particle filtering (Lee and Mumford, 2003), and in terms of minimising the free energy induced by a stimulus (Friston, 2005). Predictive coding and preferred-value tuning may well turn out to be complementary perspectives rather than competing languages for neural coding – on a very simple, abstract level, a neuron that fires most to a preferred stimulus could perhaps be thought of as instead responding to a particular divergence from the predicted *absence* of stimulation.

[2]Think back to the priors over smooth motion and slow speeds discussed in Section 2.1.4.

It is important to note that PPC models, which have attempted to show how neural populations might represent diverse distributions (Sahani and Dayan, 2003), and perform diverse computations (Ma et al., 2006), are population rate codes. This comes down on one side of two big debates in neural coding – first, do populations of neurons carry more information than the sum of their parts, and second, is information contained in spikes averaged over time (rates) or in temporally correlated spiking patterns? (see deCharms and Zador, 2000; Averbeck et al., 2006). Huys et al. (2007) use population codes, and although they consider the evidence provided by single spikes, the spike trains are still treated independently and so this is essentially a rate code on a very short timescale – the distinction between rates and spikes is arguably moot. The work by Deneve (2008a) argues for individual neural codes, but it is possible that each neuron could be re-interpreted as a single expert in a product of experts model (Hinton, 1999; Huys et al., 2007). Population codes seem intrinsically suited for representing belief distributions – they exhibit noise resistance and the ability to encode uncertainty over continuous stimulus dimensions – and work on decoding through time can perhaps be configured within this framework.

A thornier question is how we can deal with correlations between individual neurons if, indeed, they are an important source of information. It has been argued that for certain simple systems, close to the sensory epithelia, codes in which the correlations in the pattern of spikes across time and between different neurons could carry information (see e.g. Gollisch and Meister, 2008; Pillow et al., 2008). There are serious challenges for the representation and 'read-out' of such information, but thus far neural coding models for Bayesian inference have largely avoided invoking them (though see Wills, 2004; MacKay and Wills, 2005; VanRullen and Thorpe, 2002; Lengyel et al., 2005). The dynamic population codes considered above (Deneve, 2008a; Huys et al., 2007) both treat neurons and temporal patterns as more independent than perhaps they really are – there are limitations on how much history codes can carry around. This provides an interesting complement to the limitations on exact representation of complex joint posteriors mentioned above in Section 2.2.2. The relationship between independence in population codes, and independence on the level of the probabilistic representation, will be considered further in Chapter 6.

2.2.4 Using Coding Models to Link Brain and Behaviour

In their 2004 review, Knill and Pouget called for proponents of the BCH to use the kind of theoretical neural coding models reviewed in the previous section to link evidence for Bayesian behaviour with electrophysiological data. Unfortunately, there has been relatively limited progress, primarily due to the indeterminate mappings between each component of the integrative methodology illustrated in Figure 1.3. A key issue is to what extent we can

record from neurons *in vivo* whilst optimal Bayesian inference is being carried out. Human electrophysiology is unlikely to be accessible, and demonstrating optimal performance in animals would involve extensive training regimes likely to lead to very low uncertainty as well as allowing feedback-related adaptive learning (see page 31). One possibility might be to take electrophysiological recordings whilst a minimally trained animal carries out a task similar to one in which human observers have demonstrated optimality, but this presents an obvious correspondence problem. Another issue with electrophysiological data is that the probability distributions of interest might be represented only transiently, and coding models therefore need to consider the temporal characteristics of the underlying inference.

The second serious issue is that it is hard to test *all Bayesian coding schemes* against some non-Bayesian alternative, and allegiance to a particular coding scheme is difficult given the relatively early stage of their development. After selecting a neural coding model, it is desirable to show that it better matches the electrophysiological data than competing models. In a brief feasibility study conducted at the start of my PhD, I found that for the simple kinds of computations for which optimal behaviour has been demonstrated, competing codes make quite similar qualitative predictions about neural firing properties. If the match between the behaviour and the circumstances of recording is not tight, quantitative differences are not a solid basis for comparison. But perfect realisations of the integrative trinity are not the only way to link Bayesian descriptions of behaviour to neural firing properties – above we encountered a number of more piecemeal sources of evidence that Bayesian neural coding models provide a good description of electrophysiological observations.

The first source of evidence comes from existence proofs that populations of spiking neurons can encode probability distributions (e.g. Pouget et al., 2003; Sahani and Dayan, 2003), combine those distributions according to Bayes' rule (e.g. Ma et al., 2006), and perform inference in dynamic networks and through time (e.g. Rao, 2004b; Deneve et al., 2001; Deneve, 2008a; Huys et al., 2007). Codes that present biologically plausible demands on neural tuning properties and variability, and for which computations consist in more biologically tractable operations clearly provide better evidence for the feasibility of explicit Bayesian encoding. The probabilistic population code scheme is the most fully developed in terms of biological plausibility, and in terms of the range of computations it can deal with (Ma et al., 2006). Extending this scheme to incorporate work on dynamic inference through time, across complex scenes, and in dynamic and hierarchical networks would add significant weight to the claim that we have shown how, in theory, the brain could explicitly encode and compute with probability distributions.

Some of the earliest evidence for the BCH consisted in showing that neural activity could be reinterpreted in terms of Bayesian inference (see Pouget et al., 2003; Anderson and Abrahams, 1987). Several of the theoretical studies above point to similarities between

properties of their model neurons and relevant electrophysiological observations (e.g. Deneve et al., 2001; Huys et al., 2007; Deneve, 2008a; Yu and Dayan, 2005; Anastasio et al., 2000; Gold and Shadlen, 2001; Rowland et al., 2007). The consistency of LIP firing rates with the accumulation of evidence to compute a likelihood ratio is perhaps the strongest example of this (Gold and Shadlen, 2001; Rao, 2004b; Churchland et al., 2008). In more abstract correspondences, Deneve (2008a) found that biologically plausible neural properties emerged from the demands of dynamic Bayesian inference on individual spikes, and Huys et al. (2007) note that their simple decoding scheme results in what looks like adaptation to the temporal stimulus statistics. Although this kind of approach rarely has the tight logic of the integrative trinity it is certainly suggestive – at least that a Bayesian scheme is consistent with neural properties, if not that it is a better match to those properties than non-Bayesian alternatives.

Other 'arguments from appropriateness' have focused on making more general observations about the suitability of the brain for implementing Bayesian inference. Hierarchical Bayesian schemes for inferring the causal structure of the visual world predict key aspects of cortical architecture such as massive recurrence (e.g. Mumford, 1992; Rao and Ballard, 1999), a functional asymmetry between forward and backward connections (e.g. Friston, 2005), simple and complex receptive field properties (e.g. Rao and Ballard, 1999; Friston, 2005; Lee and Mumford, 2003), and even predict perceptual phenomena (e.g. Rao, 1999; Friston, 2005; Kilner et al., 2007). Ma et al. (2006) argued that, from a Bayesian gain-encoding perspective, near-Poisson variability observed across the brain can actually be seen as advantageous. This plays into a controversial debate about the source and effects of apparent noise, evoking slippery questions about whether Bayesian inference evolved to deal with neural variability, or if neural variability evolved in order to support Bayesian inference (see Ermentrout et al., 2008; Averbeck et al., 2006; Stein et al., 2005). Finding generic 'signatures' of probability distributions and Bayesian inference in the brain is an attractive proposition, and Orbán et al. (2008) have recently proposed hallmarks of generative models (i.e. likelihoods) that might be observed in visual cortex.

Making explicit integrative links, arguing for identifiable mappings between neural properties and parameters of probability distributions, requires the identification of a neural population thought to represent the relevant quantities. This is difficult – it embodies assumptions about the kinds of 'features' that support the inference of interest, and then assumes that the chosen population not only responds to those features, but through its activity supports the judgement observed in behaviour. Even with the well-known mapping between log likelihood ratios and neural firing in LIP (see Gold and Shadlen, 2007), there is still debate about what LIP neurons really respond to (see Platt and Glimcher, 1999, and Section 2.3.3) It is no coincidence that most studies that have linked brain to behaviour

via neural coding models have considered inference over simple visual quantities linked to a well-studied visual cortical area, such as the processing of motion direction in middle temporal cortex (MT). A recent paper by Morgan et al. (2008) looked at the response properties of multisensory neurons in macaque area MSTd, which respond to both visual and vestibular self-motion cues. The authors found evidence that these neurons computed a weighted sum in which the weight of each cue depends on its reliability, and this kind of approach is important to providing reciprocal, implementational constraints on models of the Bayesian computations that support behaviour.

2.3 The Anatomical Basis of Bayesian Decision Making

In the introduction, we argued for expanding the remit of the BCH away from simple, optimal inference to more complex perceptual domains. In order for this to occur, Bayesian models must consider complex and perhaps approximate inference, and ask when, and to what degree, behaviour is Bayesian. As the behaviour that enters into the integrative loop of Figure 1.3 becomes more complex, identifying the neural substrates will require a better understanding of how elements of Bayesian decision making illustrated in Figure 2.5 (and summarised in Equation 1.3) map onto the functional anatomy identified in neuroscientific studies of decision-making. The potential of a Bayesian approach to the brain is far richer than suggested by the project of matching a single neural population to a single sensory posterior – uncertainty is almost always integrated with biases and utilities in the service of decision-making.

In this section of the literature review we consider the main findings from neuroscientific studies of decision making, and consider how they might be viewed in a Bayesian context. This situates the fMRI study reported in Chapter 5, in which we ask specifically where sensory uncertainty in $p(o_j \mid d_i)$ is integrated with externally imposed loss functions in $p(U \mid o_j)$. This invokes elements of both *perceptual* and *value-based* decision making – domains that have traditionally been investigated separately (see Heekeren et al., 2008; Rangel et al., 2008, respectively). Value-based decision making occurs whenever an animal chooses between several alternatives on the basis of their subjective value, and applies to behaviours from insect foraging to human stock-market trading. The components of value-based decision making have been investigated under several loose 'stages', that overlap with the formal notation introduced in Figure 1.2 and are schematised in Figure 2.4 (adapted from Rangel et al. (2008) and Heekeren et al. (2008)). In this picture, decision-making involves the **representation** of possible actions and relevant internal and external states, the

valuation of each action, **action selection** on the basis of the resulting values, and then **outcome evaluation** that drives **learning** in the representation and valuation processes.

The other component in Figure 2.4 is **risk and uncertainty**, which can modulate representation, valuation, and action selection. In the context of value-based decision making, especially in studies of economics, these uncertainties tend to involve probabilistic contingencies in the external world. However, as discussed in the introduction, *perceptual* decision making involves sensory uncertainty that can come from internal sources as well as from variability in the external world, and occurs whenever an animal decides between competing interpretations of sensory evidence – for example, deciding if a degraded image is a face or house, or whether an Gabor patch is oriented clockwise or anti-clockwise[3]. This kind of uncertainty is intrinsic to the **representation** of decision states, as explored in the evidence-accumulation models introduced in Section 2.2.1, where uncertainty also contributes to the **action selection** mechanism. And in the Bayesian decision-theory notation, sensory uncertainty also contributes to **valuation** – to the $p(o_j \mid d_i)$ component of expected utility. The components illustrated in Figure 2.4 can be a useful organising principle, expressing the rough temporal 'flow' from decision evaluation through selection, execution, evaluation, and learning. However, as is evident when we try to marry perceptual and value-based decision making, they do not constitute a clear computational taxonomy.

Reframing alternative approaches in a common probabilistic language has huge potential to unify different models, resolve apparent disagreements, and point the way to future studies for resolving remaining conflicts. A recent paper by Dayan and Daw (2008) takes this approach, expressing several existing decision-making models in a very general machine learning setting - the partially observable Markov decision process (POMDP). In Figure 2.5 we make a more local attempt to unify some of the concepts we will be considering below, expanding the notation of the Bayesian decision maker introduced in Figure 1.2 to incorporate evaluation and to indicate where models of perceptual decision-making fit in. We then survey the neuroscientific literature on each element of the decision-making flow-chart under this notation, and the pink boxes in Figure 2.5 indicate the main neural areas that have been associated with each component of the decision-theory equation. We pay particular attention to unanswered questions about how sensory uncertainty and external value might be integrated – there is some work in the SDT context using external loss functions to probe the integration of reward with perceptual accuracy (Maddox and Bohil, 2004; Davison and Tustin, 1978; Green and Swets, 1966), but little neurobiological data on this question.

[3]In value-based decision making, the decision maker chooses between actions. In perception, decisions are not always paired with actions, so both 'actions' and 'decisions' are used in the text to refer to the outcome of a decision-making process.

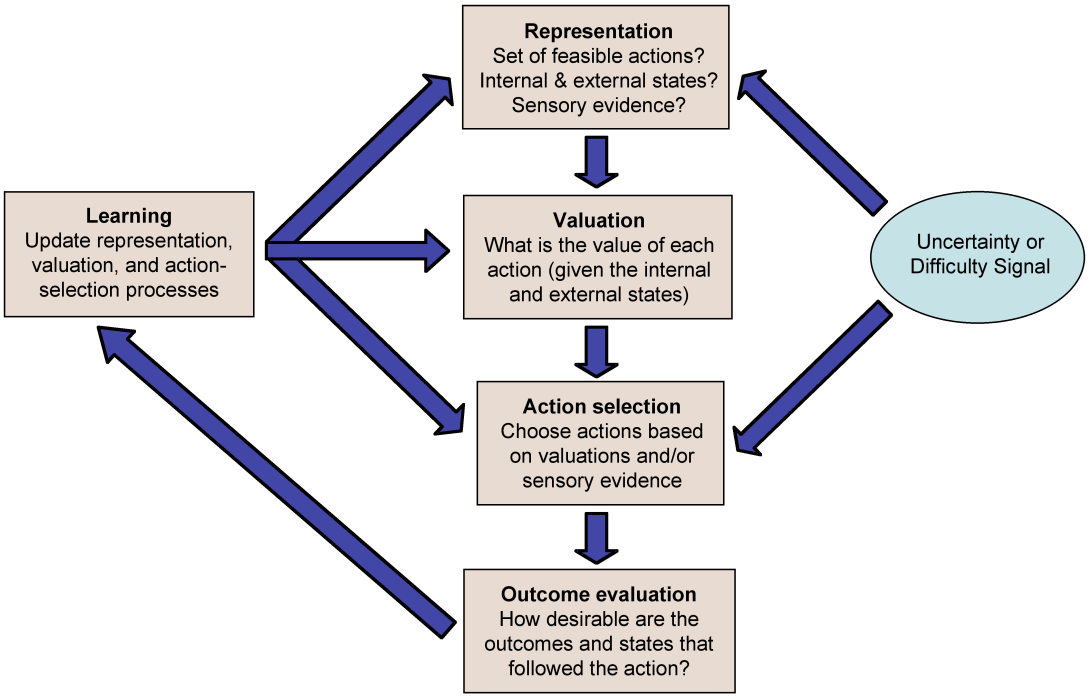

Figure 2.4: ***Components of decision making.*** *Possible decisions are represented, along with internal and external states of the world relevant to the assessment of value. The value of each decision is assessed and the one that optimises expected utility is taken. The outcome obtained is then evaluated and used to drive learning (Rangel et al., 2008). For perceptual decision making, posterior beliefs over the state of the world are used to compute decision variables that reflect the likelihood that the observer will take each possible decision. How value is integrated into this process is unknown, and uncertainty in the perceptual decision might directly affect both sensory representation and action-evaluation (Heekeren et al., 2008). There are many open questions about the theoretical and neural overlap of these variables in different kinds of decision making, and in Figure 2.5 we present a formal notation that we use to help pin down these relationships.*

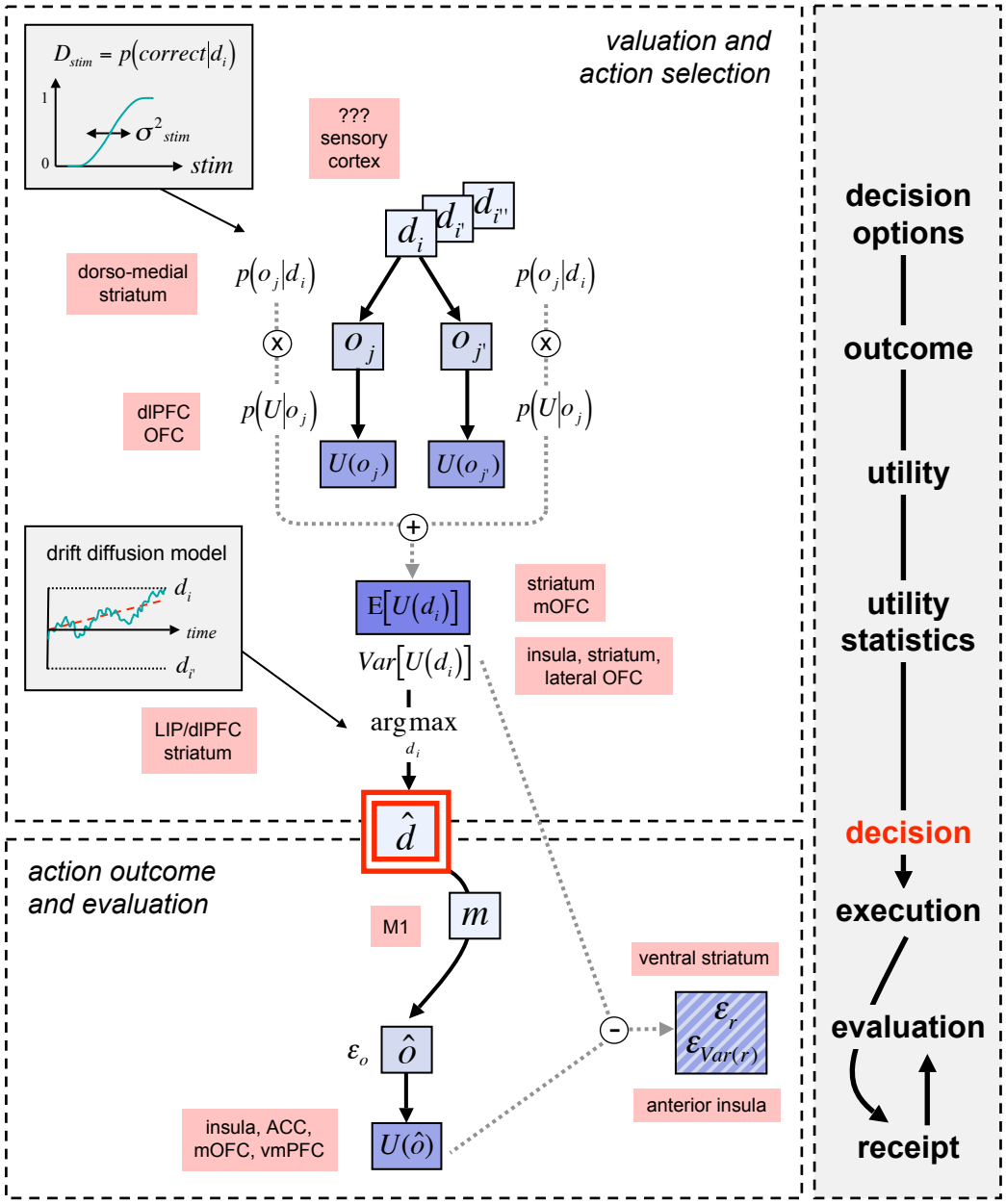

Figure 2.5: ***Formalising the components of decision making.*** *The components of decision-making illustrated in Figure 2.4 are formalised under Bayesian decision theory. A decision $\hat{d}$ is selected by comparing the expected utility of different options; $\mathbf{E}[U(d_i)]$. After the associated action m is taken, $\hat{d}$ is evaluated by comparing expected quantities to those obtained, yielding prediction errors ϵ. Perceptual decision-making can be incorporated within this framework – the upper grey box illustrates perceptual uncertainty in $p(o_j|d_i)$, and the lower grey box illustrates the drift diffusion model for the accumulation of sensory evidence. Pink boxes indicate neural areas associated with each component. This schematic extends the version presented as Figure 1.2 – see text for further details.*

The core of Figure 2.5 is the selected decision, $\hat{d}$. As explained in Section 1.3.1, $\hat{d}$ is chosen by comparing the expected utility (EU) of different decisions d_i; $\mathbf{E}[U(d_i)]$. EU is computed by summing over all possible outcomes the product of the mapping from decision to outcome; $p(o_j|d_i)$, and from each outcome to its utility; $p(U|o_j)$ (illustrated by the dashed grey arrows in the top panel of Figure 2.5);

$$\mathbf{E}[U(d_i)] = \sum_j p(o_j|d_i)\ p(U|o_j) \tag{2.9}$$

As illustrated in Figure 2.5, statistics other than the expectation of the utility can also be computed, for example the variance; $Var[U(d_i)]$.

In our notation, an outcome is defined as the state of the world that results from taking a particular decision or action at a particular time – for example, the delivery of food after a rat chooses to press one of a number of levers. The utility or 'loss' function, $p(U \mid o_j)$ then describes the mapping between that state of the world and its value to the decision maker, which can be stochastic or deterministic. For example, the food delivered to the rat might have a fixed subjective value, be modulated by satiation, or have an unpredictable taste. Once a decision is selected, it is expressed through a motor output m, leading to an actual outcome $\hat{o}$ with utility $U(\hat{o})$, both of which can be compared to their expected values to generate prediction errors ϵ_o and ϵ_r respectively. Other statistics, such as the predicted variance, can also be compared to the utility of a particular decision over time to generate a 'risk prediction error' (see Preuschoff et al., 2008).

Perceptual decision making has traditionally been viewed outside this framework, but can be incorporated within its notation. In a perceptual task, posterior distributions over states of the world are used to make decisions – for example, making a categorical judgement where the options are indexed by d_i, and where outcomes o_j may also be categorical states such as being correct vs. incorrect. In this setting, $p(o_j \mid d_i)$ can be thought of as marginalising out possible states of the world, as indicated in Equation 1.4. In a classic categorisation task such as that used in Chapters 3 - 5, observers classify stimuli drawn from a continuous axis as falling either side of a boundary – for example, deciding whether an offset is left vs. right, or whether an RDK is moving up vs. down. Data on such a task can be expressed as a psychometric function, which plots the stimulus axis against the relative proportion of the two answers (see the upper grey box in Figure 2.5). The lower grey box in Figure 2.5 illustrates the drift diffusion model proposed as a mechanism for accumulating sensory evidence, and for making a decision when the weight of evidence in favour of one option exceeds a threshold (see Gold and Shadlen, 2007), which can also be reframed in terms of Bayesian decision theory (see Dayan and Daw, 2008, and page 59).

2.3.1 Representation and Valuation in Value-Based Decision Making

The least well understood component of value-based decision making is perhaps **representation** – it is not clear how the brain decides to which actions to assign value to in a real-world context, how many actions can be represented at once, and how internal and external states are computed and transmitted to the valuation system. The **valuation** of actions has however been extensively studied in the context of animal learning models[4], and it is generally agreed that there are three different types of valuation system that compete for control of action (see Balleine et al., 2008; Dayan, 2008b); Pavlovian, habitual, and goal-directed, with recent proposals of a fourth 'episodic' controller (Lengyel and Dayan, 2008). There are still unanswered questions about whether this taxonomy is neurally consistent and species general, and how the different valuation systems compete on common ground for the control of action (see Rangel et al., 2008; Dayan et al., 2006; Daw et al., 2005). Here, we are specifically interested in model-based, goal-directed control, in which all components of expected utility are used flexibly to evaluate statistics of the utility of different decisions, rather than the observer forming rigid $d_i - U$ mappings as in Pavlovian or habitual learning. Model-based control is flexible to changes in the environment, and is clearly the scenario most relevant to the BCH (see also page 22).

Electrophysiological animal studies and human neuroimaging have begun to trace out the neural circuits involved in valuation – in very general terms, it seems to involve cortico-subcortical loops, with subcortical systems integrating different sources of value information with emotional and motivational contexts, in order to assist the cortical evaluation and control of behaviour. The cortical regions implicated are mainly in regions of prefrontal (PFC) and parietal cortex associated with the generation of decisions and control of action (Wood and Grafman, 2003). These include the orbitofrontal cortex (OFC), thought to reflect integrated sources of reward information (Wallis, 2007), the insula, thought to be involved in integrating emotional context with sensory experience (Augustine, 1996), and the anterior cingulate cortex (ACC), thought to be involved in reward anticipation, error detection, and conflict monitoring (Botvinick, 2007). The subcortical regions implicated include the striatum, thought to play a role in action selection and reward encoding as well as in regulating motor performance (Balleine et al., 2007; O'Doherty et al., 2003), and the amygdala, thought to help coordinate physiological responses and learning associated with emotional states (Phelps and LeDoux, 2005), and to guide attention to emotional salient stimuli (Vuilleumier, 2005). Subcortical dopaminergic nuclei such as the substantia nigra

[4] 'Value' is often used to refer to the utility of a particular outcome $U(o_j)$, but here 'valuation' refers to the computation of expected utility for different decisions, which might include uncertainty both in $p(o_j \mid d_i)$ and $p(U \mid o_j)$

have also been implicated in outcome evaluation and learning (Montague et al., 1996), which will be discussed below in Section 2.3.5.

Below we will focus in more detail on the neural basis of **goal-directed** valuation. Existing anatomical studies in rodents suggest that $p(o_j \mid d_i)$ might be encoded in the dorsomedial striatum (Yin et al., 2005), and $p(U \mid o_j)$ in the dorso-lateral prefrontal cortex (dlPFC) and OFC (Wallis and Miller, 2003; Padoa-Schioppa and Assad, 2006; Wallis, 2007; Barraclough et al., 2004; Schoenbaum and Roesch, 2005), with the basolateral amygdala and mediodorsal thalamus also implicated (Balleine, 2005). Human fMRI studies concur with the involvement of dlPFC (Plassmann et al., 2007) and OFC (Tom et al., 2007; Hare et al., 2008; Paulus and Frank, 2003; Plassmann et al., 2007) in encoding elements of $p(U \mid o_j)$. Studies have found evidence consistent with an expected utility signal in striatum and medial OFC (Rolls et al., 2008), and activity consistent with the variance of utility in the striatum (Preuschoff et al., 2006; Dreher et al., 2006), insula (Rolls et al., 2008; Preuschoff et al., 2008), and lateral OFC (Tobler et al., 2007). This picture is consistent with the separate representation of various components of expected utility, which are then integrated in the basal ganglia (BG) and communicated to decision-making regions of the cortex, but this is still very much an open question.

In this thesis we are particularly concerned with uncertainty or variability, in both $p(o_j \mid d_i)$ and $p(U \mid o_j)$, but there are other contributors to EU, including motivational, emotional, and social modulators of value. Prospect theory (PT) is closely related to Bayesian expected utility, but defines value relative to a reference point and passes the objective probabilities through a non-linear function that can express various heuristics and biases intrinsic to human economic and probabilistic reasoning (see Kahneman and Tversky, 1979). Debate about whether these biases represent suboptimal reasoning, or represent optimality under constraints, ecologically sound loss functions, and prior knowledge is ongoing (Gigerenzer, 2002). For our purposes EU and PT calculations of expected utility are likely to be equivalent so we stick to the more straightforward EU formalism. Human fMRI studies have found evidence for a PT-like signal in a wide network of regions including ventral and dorsal striatum, ventromedial and ventrolateral PFC, ACC and midbrain dopaminergic regions (Tom et al., 2007), overlapping with those associated with expected utility.

2.3.2 Representation and Valuation in Perceptual Decision Making

Unlike the representation of actions and internal/external states relevant to value-based decision making, the anatomical basis of *stimulus* representation has received a good deal of attention. There is a paradigmatic consensus that despite massive recurrency, visual cortex

is organised in a loose hierarchy that represents scenes in terms of increasingly complex features, with functionally specialised regions corresponding to features such as motion, colour, orientation and even semantic categories such as faces, tools, and building-like structures (Van Essen et al., 1992). For example, physiological studies of middle temporal visual area (MT) have shown a close correspondence between neural activity and behavioural reports of visual motion (e.g. Newsome et al., 1989), and microstimulation studies that evoke analogous perceptual experiences support the causal role of such areas (Ditterich et al., 2003). Regions of sensory cortex that selectively respond to faces (fusiform face area; FFA) and houses (parahippocampal place area; PPA) have been extensively used in human neuroimaging studies as clearly definable substrates of stimulus representation (e.g. Haxby et al., 1994). Evidence for their role in representing these stimulus categories includes the increasing probability that monkeys will report seeing a face following microstimulation of face-responsive neurons in FFA (Afraz et al., 2006), and the appropriate correlation of BOLD (Blood-oxygen level dependent) signal in the FFA and PPA with the difficulty of a face-house discrimination (see Heekeren et al., 2004), and with switches between face and house percepts in binocular rivalry (Tong et al., 1998).

The inverse inferences of perception, deriving a posterior belief about a state of the world given noisy sensory evidence, are expressed in perceptual decisions according to Equation 1.4. As described above, the classic categorisation task we use in Chapters 3–5 can be characterised by a psychometric function that plots the continuous stimulus axis against the relative proportion of two categorical answers – for example, whether an offset is left or right of centre (see the upper grey box in Figure 2.5). Two measures of sensory uncertainty, both of which contribute to $p(o_j \mid d_i)$, can be derived from the psychometric curve. The uncertainty with regard to a particular stimulus (or stimulus 'difficulty'; D_{stim}) increases away from the categorisation boundary in both directions. We can also characterise the uncertainty across the whole axis, measured by the slope of the psychometric function, σ^2_{stim}, and corresponding to signal detection 'sensitivity' (see Equation 5.2). In Chapter 3 we show that under a simple Gaussian model of uncertainty in the stimulus representation, this quantity can also serve as a proxy for the width of the posterior.

The proponent of the BCH argues that sensory uncertainty is explicitly represented in neural populations – i.e. that we can point to some property of neural firing that corresponds to the uncertainty in a posterior belief distribution. Under this view, uncertainty is embedded in the 'representation', but according to the notation of Figure 2.5, it contributes to $p(o_j \mid d_i)$ and thus to the valuation of decisions. We thus re-evaluate the notion of a functionally isolated uncertainty signal illustrated in Figure 2.4 – sensory uncertainty is an intrinsic part of the decision-making machinery. However, there may be specific anatomical correlates of these sensory contributions to $p(o_j \mid d_i)$. Human fMRI studies have found D_{stim}

correlating with the BOLD signal in medial frontal gyrus, inferior frontal gyrus/anterior insula, and ACC (see Thielscher and Pessoa, 2007; Binder et al., 2004), and reflected across the distributed decision-making circuit we have been reviewing (Philiastides and Sajda, 2007; Grinband et al., 2006), and a recent study in rats found a difficulty or 'confidence' signal for odor categorisation task reflected in the OFC (Kepecs et al., 2008).

It is interesting to note that difficulty-encoding regions overlap with those responding to risk (see Preuschoff et al., 2008). This highlights the potential interconnectedness of valuation components - for a task with different outcomes for correct and incorrect responses, an increase in difficulty will lead to an increase in the variance of rewards obtained over time. Heekeren and colleagues have suggested that a difficulty signal could also serve to recruit attentional resources in demanding situations (see Heekeren et al., 2008), consistent with the proposed role of the ACC in detecting error and monitoring conflict in order to instruct a cognitive control system in dlPFC (Botvinick, 2007). There are many remaining questions about how the correlates of perceptual contributions to $p(o_j \mid d_i)$ overlap with externally determined outcome contingencies in animal learning tasks, and are integrated in the computation of value statistics such as expected utility. For example, whether there is a discrete neural correlate of sensitivity or posterior uncertainty, which could be utilised in combining different estimates and loss functions, is unclear.

2.3.3 Models of Perceptual Decision Making

The contributions of Bayesian perceptual inference to $p(o_j \mid d_i)$ have traditionally been considered in terms of accumulating sensory evidence, where accuracy is the only reward and $p(U \mid o_j)$ can therefore be neglected. The lower grey box in Figure 2.5 illustrates a model of sensory evidence accumulation that has gained great currency in recent years, and presents the most widely-known example of a mapping from a Bayesian inference to a neural variable. In order to make a perceptual categorisation, the likelihood each stimulus category accords to the data can be compared – if the ratio of two likelihoods (see Equation 2.4) exceeds one, the category represented in the numerator should be selected over that represented in the denominator. But as we discussed with regard to extending simple probabilistic population codes (see Section 2.2.3), the brain receives information over time, even during the presentation of a single stimulus. The optimal Bayesian decision maker should therefore compute the likelihood ratio sequentially (more generally, a 'sequential probability ratio test'; SPRT), accumulating the evidence for each category until the weight of evidence in favour of one exceeds a threshold (see Wald, 1947; Smith and Ratcliff, 2004; Ratcliff and Rouder, 1998). In terms of the notation in Figure 2.5, rather than computing $p(o_j \mid d_i)$ where o_j is correct vs. incorrect, and d_i is left vs. right, a decision variable is com-

puted that corresponds to the *likelihood* of the stimulus being left vs. right given sensory evidence; something like $p(d_i \mid \mathbf{s})$. This evidence-accumulation process can be implemented by a diffusion-to-bound mechanism, which has been shown to give a good fit to behavioural accuracy and reaction time (RT) data (Ratcliff and Smith, 2004; Luce, 1986; Usher and McClelland, 2001; Reddi et al., 2003).

Following seminal work in the late 1990s (e.g. Kim and Shadlen, 1999), there has been an explosion of interest in how areas in the prefrontal and parietal cortex might encode perceptual decision variables such as the likelihood ratio (see Gold and Shadlen, 2007, for review). As mentioned above (page 57), motion-selective neurons in MT appear to reflect the probability of their preferred direction of motion being present in their receptive field. During a motion discrimination task in which saccades were used to report the direction of motion, activity in primate LIP, FEF, and dlPFC was found to predict the monkey's gaze shift, and to reflect a combination of sensory evidence and motor signals (Kim and Shadlen, 1999). Later analyses suggested that some LIP responses reflect the log likelihood ratio as sensory evidence is accumulated (Gold and Shadlen, 2001; Roitman and Shadlen, 2002; Shadlen and Newsome, 2001), with modelling work supporting the idea that LIP firing rates correspond to a diffusion-to-bound or race model that implements an SPRT (see Smith and Ratcliff, 2004; Bogacz, 2007; Ratcliff and McKoon, 2008).

Studies looking for perceptual decision-variables in humans broadly support the findings in primates, but without the spatial and temporal resolution of electrophysiological data that allows comparison to parameters of an evidence-accumulation model. In an fMRI study looking at face-house discrimination, Heekeren et al. (2004) observed activation in the dlPFC and both superior and inferior frontal sulci that was anti-correlated with the level of noise added to the stimuli, and was positively correlated with the difference in FFA and PPA activity, properties suggestive of a decision variable. Similarly, an MEG study of an auditory categorisation task found gamma band activity over the dlPFC that was anti-correlated with task difficulty, and which correlated with the difference in activity over sensors thought to reflect the two kinds of pattern change listeners were asked to discriminate (Kaiser et al., 2007).

Despite these successes, the SPRT and the diffusion-to-bound model apply to a very restricted range of settings – most obviously, this approach is limited to computing likelihood ratios for deciding between small numbers of options. Recently, a multiple-SPRT has been used to model electrophysiological responses for four-choice (rather than the typical two-choice) tasks (Churchland et al., 2008), and theoretical work has extended the diffusion-to-bound model to multiple decisions (Thornton and Gilden, 2007). However, this still involves choosing between predefined decisions on the basis of a discrete statistic, rather than forming a full probabilistic representation on the basis of which a multitude of decisions

could be made. The SPRT approach also lacks a formal way of dealing with alterations to components of valuation such as external utility or biased priors. For example, if we manipulate the utility of different outcomes by changing external value (expressed in $U(o_j)$), it is unclear how to optimally adjust the model parameters. Some studies have found that LIP neurons, rather than reflecting an SPRT, can indeed be modulated by the reward magnitude associated with a saccadic response into their receptive field (Platt and Glimcher, 1999), and by the probability of reward (Yang and Shadlen, 2007). But the interpretation of these results is a matter of contention (see Glimcher, 2004; Dayan and Daw, 2008; Shadlen et al., 2007; Yu, 2007). Reframing the SPRT under a fully general POMDP framework, as suggested by Dayan and Daw (2008) might help to clarify regions of disgreement, and strengthen the interpretation of these neural activities in terms of a Bayesian decision algorithm (see page 51).

In Bayesian formulations of a sensory decision, a posterior belief distribution is computed (expressed in $p(o_j \mid d_i)$), and then combined with externally determined loss functions (expressed in $p(U \mid o_j)$) to yield a decision that maximises expected utility. This suggests a sequential process (see also Figure 3.1b), but the neural implementation might well be non-sequential. One intriguing possibility is that value could alter the sensory representation itself via feedback connections. There is existing evidence that top-down signals can modulate processing in sensory regions, when attention (Johnson et al., 2007; Reddy et al., 2007; Vuilleumier and Driver, 2007; Wojciulik et al., 1998), emotion (Hsu and Pessoa, 2007; Vuilleumier et al., 2004, 2001), task set (Summerfield et al., 2006a), and conflict (Egner and Hirsch, 2005) are manipulated. Indeed, changes in reward schedule have been reported as expressing attention-like effects on early visual areas in rodents (Maunsell, 2004; Shuler and Bear, 2006).

2.3.4 Action Selection

So far we have considered how various sources of information about the likely value of different decisions might be combined in the brain. However, our understanding of the mechanism by which values are compared – represented by the **argmax** operation in Figure 2.5 – is slim. The diffusion-to-bound model for implementing a SPRT (see Gold and Shadlen, 2007, for review) makes a decision when the evidence in favour of one option exceeds a threshold. Bogacz and Gurney (2007) have proposed a neurobiologically inspired (see Behrens et al., 2003) network model, in which the BG implement threshold crossing for a multiple SPRT test. In this model, output nuclei that issue motor commands combine sensory evidence, received from cortex via the striatum, with a signal expressing the conflict between alternatives, represented in the subthalamic nucleus (STN) and golbus pallidus

(GP). When the combination of these signals for a particular action exceeds a (negative) threshold, the relevant motor command is disinhibited. In a related model, Lo and Wang (2006) argue that local dynamics in the superior colliculus (SC; known to be involved in generating saccadic eye movements) produce burst responses that signal threshold crossing, and that the level of the threshold is set by the strength of cortico-striatal synapses (see also Frank, 2006; Simen et al., 2006).

This picture is consistent with views of the basal ganglia as a switch or arbiter that acts to resolve competition between multiple cortical and subcortical systems that vie for control of behaviour (Redgrave et al., 1999; McHaffie et al., 2005; Mink, 1996; Frank, 2006). However, the same caveats discussed in the previous section about the generality of the SPRT and its link to neural variables apply here – in particular, it is unclear how utility should affect evidence accumulation and threshold crossing. Ding and Gold (2008) found that neurons in the caudate nucleus of the dorsal striatum reflect value-related variations in reward expectation, threshold crossing and bias, but such results are not specific enough to draw firm conclusions, especially in the absence of flexible computational models of the decision process.

Another important question is how decision variables interact with motor plans – in primates prefrontal and parietal areas whose activity reflects likelihood ratios overlap with those involved in selecting, planning, and implementing motor responses (Hernandez et al., 2002; Romo et al., 2004). This supports the view that the motor system is an integral part of a decision-making process that ultimately finds expression in a limited number of effectors (Cisek, 2007; Wyss et al., 2004; Verschure and Althaus, 2003). However, these studies all ask the monkey to report their perceptual decision via a motor response – the two are functionally linked. In human fMRI studies, similar tasks in which perceptual decisions are reported with saccadic eye movements found decision-variables in FEF and ventrolateral PFC, again regions thought to be involved in regulating motor commands (Sereno et al., 2001; Heinen et al., 2006). However, when Heekeren et al. (2006) varied response modality they found a network of left posterior dlPFC and cingulate cortex, left IPS, and left fusiform/parahippocampal gyrus that correlated with the strength of sensory evidence independent of whether responses were given with button presses or eye movements. More generally, there is evidence that dlPFC is involved in selecting responses on the basis of context and sensorimotor contingencies, not just in gating a particular motor response (Thoenissen et al., 2002). This raises important issues about how flexible and multifarious fronto-parietal decision-variables are, how they are linked to specific actions, and whether there are significant species differences.

2.3.5 Outcome Evaluation and Learning

In most situations, the value of decisions must be learnt, and the prerequisite for learning is reflection on experience – in Figure 2.5, evaluating $U(\hat{o})$. Human fMRI studies have found that the medial OFC correlates with subjective reports of positive outcomes, for primary reinforcers such as food (e.g. Kringelbach et al., 2003) and secondary reinforcers such as monetary reinforcement (e.g. Knutson et al., 2001). The medial OFC also shows a reduction of activity that parallels the reduction in gustatory reward value after people are fed to satiation, suggested that it tracks subjective value (e.g. O'Doherty et al., 2000). Relatedly, Gottfried et al. (2002) reported that the OFC also reflected learned associations of positive vs. negative odors with neutral face stimuli, suggesting that it has chemosensory response properties. Correlates of negative outcomes such as pain have been found in the insula and ACC (e.g. Davis et al., 1997), and a primate electrophysiology study found outcome value signals in the dorsal ACC (Seo and Lee, 2007). Interestingly, the correlation of medial OFC activity with subjective reports of value is cognitively penetrable – it increased when participants were told that a smell belonged to an expensive rather than a cheap bottle of wine (Plassmann et al., 2008), or to cheese rather than smelly socks (de Araujo et al., 2005). McClure et al. (2004) tried to pull apart the contributions of cultural and sensory evaluations of Pepsi® vs. Coca Cola®, and found that activity in ventromedial PFC predicted purely sensory preferences, and that brand knowledge recruited additional regions in the hippocampus, dlPFC, and midbrain.

Once outcomes have been evaluated, the comparison between the expected utility of an action and what was actually obtained; ϵ_r, can be used to drive learning. There is much that the system could change – which actions it considers, various components of expected utility, or mechanisms of action selection. An area in which computational approaches to value-based decision making have been particularly successful is the use of reinforcement learning (RL) models to capture the habitual learning of action values, uncovering neural substrates via matching model parameters to neural data (e.g. Montague et al., 1996; Hare et al., 2008). At the heart of RL is the idea that changing the value of actions in proportion to transient prediction error signals causes the system to converge on values that optimise reward (Sutton and Barto, 1998; Montague et al., 2006; Sutton, 1988; Montague et al., 1996). Neurons in primate midbrain dopaminergic nuclei show activity profiles that correspond closely to prediction errors (Schultz et al., 1997; Montague et al., 1996), and in human fMRI studies prediction error correlates with the BOLD signal in ventral striatum regions targeted by midbrain dopaminergic neurons (e.g. Hare et al., 2008; Breiter et al., 2001; Yacubian et al., 2006; O'Doherty et al., 2003; Seymour et al., 2004). This role for the basal ganglia in reinforcement learning is not inconsistent with the role in threshold

mechanisms discussed above, as these mechanisms need to be shaped during learning (see Bogacz and Gurney, 2007, for discussion).

In a fast-changing world, always choosing the action accorded the highest EU by your current assessment is not necessarily optimal in the long-term – if expected utilities change, you risk missing out on a better option in the future. The habitual learning of values could lead to dangerous inflexibility, and animals must therefore balance the need to exploit habitual action-value contingencies with the need to explore a dynamic world (Cohen et al., 2007; Daw et al., 2006). In this context, we might expect to see signals with longer timescales reflecting performance monitoring and changes of cognitive set that alter the value landscape. More specifically, tracking components of value other than the magnitude of reward, including other moments such as variance or risk, uncertainties in $p(U \mid o_j)$ and $p(o_j \mid d_i)$, and their combination into EU/PT signals, could drive flexible learning.

In a recent human fMRI study, Preuschoff et al. (2008) found correlates of risk prediction and risk prediction error in the insula, and Dreher et al. (2006) found that the ventral striatum correlated better with sustained reward uncertainty (related to risk) than to transient prediction error. This signal has a longer timescale, and could be a useful index of when exploration is advantageous, perhaps being passed to fronto-parietal control regions. Daw et al. (2006) found that, in humans, the intraparietal sulcus (IPS) and frontopolar cortex were specifically associated with decisions to explore rather than exploit. In primates, the supplementary eye fields (SEF) and rostral cingulate motor area have been implicated in the general need to adjust behaviour in the face of performance changes (Ridderinkhof et al., 2004; Stuphorn et al., 2000; Ito et al., 2003), which is interesting in the context of the overlap between decision variables and motor planning regions (see page 61).

The question of how different sources of uncertainty might drive different kinds of learning relates to the question of how different valuation systems can compete for behavioural control (Rangel et al., 2008; Dayan et al., 2006; Daw et al., 2005). Finding answers will require a greater understanding of the neuroanatomical correlates of the (de-correlated) components of valuation, supported by theoretical models whose performance can be compared to neural data (see e.g. Yu and Dayan, 2005; Dayan and Daw, 2008; Gold and Shadlen, 2007). In this thesis we do not directly address learning, but we explore the anatomical basis of variables important for learning such as risk, uncertainty, and reward feedback (see Chapter 5). As discussed above in the context of neural coding models for Bayesian inference, considering how representations change with learning and over time is critical to having a comprehensive picture of the constraints on them and the roles they need to play.

2.3.6 Searching for Synthesis

The picture that emerges from considering value-based and perceptual decision-making is of a distributed network of cortical and subcortical regions. This is perhaps unsurprising – a decision is charged with integrating a huge variety of states, drivers, and sources of information into a single motor act (see e.g. Cisek, 2007). For value-based decision making the representation of potential actions and the diverse range of relevant internal and external states is poorly understood. However, valuation itself has been extensively investigated, especially in the context of goal-directed control (see Rangel et al., 2008). Determining the value of an action involves the integration of many sources of information, thought to occur partly in the basal ganglia. The basal ganglia are connected to regions of PFC involved in integrating value with emotional context and with monitoring performance, including the ACC and OFC, and to the dlPFC, which is implicated generally in cognitive control (Wood and Grafman, 2003). All these regions reflect components of valuation, with some suggestive evidence that $p(o_j \mid d_i)$ is encoded in the BG (e.g. Yin et al., 2005) and $p(U \mid o_j)$ in the OFC and dlPFC (e.g. Wallis and Miller, 2003; Hare et al., 2008), with integrated EU/PT signals reflected across the network (Tom et al., 2007; Rolls et al., 2008; Balleine et al., 2008). Haber (2003) has argued that integrative networks within the basal ganglia are ideally suited for channelling information from limbic, to cognitive, to motor circuits, effecting the integration of value and motivational context into action planning. However, exactly where the different contributors to valuation are integrated is unknown (see Rangel et al., 2008), and the question of how different valuation systems compete for the control of action is unresolved (Dayan et al., 2006; Daw et al., 2005).

Perceptual decision-making involves the representation of external states of the world in specialised regions of the occipital and temporal lobes, and uncertainty in these representations is an important determinant of EU via $p(o_j \mid d_i)$. How this overlaps with other contributions to $p(o_j \mid d_i)$ is unknown. Decision variables – quantities that reflect the likelihood of taking a particular decision – are found in the fronto-parietal cortex, particularly in primate LIP neurons that reflect the integration of sensory evidence according to a diffusion-to-bound model that implements the SPRT, with the striatum implicated in signalling and regulating threshold-crossing (Gold and Shadlen, 2007). This again implicates a cortico-striatal circuitry, and the contributions of perceptual uncertainty to expected utility would be expected to impact on the BG, concordant with their proposed role in integrating value information. However, restricted SPRT models do not speak to the integration of changes in utility, prior bias, or motivational state, contributing to controversy about the interpretation of evidence that LIP neurons also reflect the magnitude (Platt and Glimcher, 1999) and probability (Yang and Shadlen, 2007) of reward. The neural implementation of mechanisms

for decision or action selection, across both perceptual and value-based decision making, is an open question – threshold-crossing in a diffusion-to-bound mechanism is only one option.

In most scenarios, the brain must *learn* about valuation, and adjust the parameters of representation, action selection, and learning in order to maximise reward in a constantly changing environment. Work on habit learning has found correlates of prediction errors in midbrain dopaminergic neurons and in their ventral striatal target, suggesting that the integration of reward information is affected directly by prediction error. Correlates of risk prediction error have also been found in the insula, supporting the idea that the brain monitors a comprehensive model of the sources of value and adjusts the parameters of decision-making appropriately (Preuschoff et al., 2008). This chimes with the Bayesian perspective, in which uncertainty is critical for inference and learning, but raises questions about the relation between the uncertainty intrinsic to sensory representation and abstract uncertainty signals that directly modulate other components of the decision-making circuit. Relatedly, it is not known to what degree the representation of prediction errors for perceptual decisions overlaps with the correlates seen in dopaminergic midbrain regions and ventral striatum for economic decisions. A recent fMRI study of face detection found evidence for predicted perception instead in the medial frontal cortex, and for increases in top-down connectivity from the frontal cortex to FFA for face decisions, consistent with the comparison of predicted and observed information (Summerfield et al., 2006a).

2.4 Attention and the Bayesian Coding Hypothesis

A serious limitation in evidence for the BCH is the simplicity of the situations it deals with – in its probabilistic formulations, in behavioural and electrophysiological experiments, and in the neural coding models that implement representation and computation. Most studies consider inference with regard to one or at most a handful of objects in the focus of attention, in stark contrast to the multitude of spatially distributed features that make up real world scenes (see Figure 2.2). There are clearly practical reasons for studying simple paradigms (Rust and Movshon, 2005), including constraints on training animals, the limited specificity of neural recordings, and technical barriers to modelling highly complex inference. Relatedly, studying behaviour in very simple paradigms that allow observers to achieve optimality enables researchers to argue that Bayesian uncertainty must be taken into account. This is an important project, and provides a tractable substrate for linking neural coding models to neural data, but presents a very restricted picture of the brain. Simple introspection tells us that inference is not optimal for all elements of a complex

scene, and the improved processing of objects in the focus of attention implies that it is not possible for the brain to represent everything optimally at the same time.

Attention research has a long and tangled history, encompassing an astonishing range of phenomena; from arousal and vigilance to highly focused processing of a particular feature dimension. In this thesis we are concerned with top-down, selective attention – in the oft-cited words of William James, the "taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought". The idea of selection implies the existence of a limited neural resource or process, to which attention grants selective access. However, attempts to delineate the nature of this limited resource, and the mechanisms and effects of selection, have floundered in a sea of heterogenous effects. The failure to theoretically unify these effects has led to assertions that a single neural resource allocated by attention is not a useful concept (Driver, 2001; Zelinksy, 2005; Itti et al., 2005), and to a focus on *when and how* selection occurs, rather than *why*.

In Chapter 6 we present a framework that places the unifying limited resource on the **computational** and **algorithmic** levels, informed by **implementational** constraints. We claim that the computational goal of performing inference on complex joint posteriors is intractable, and that attention locally improves the impoverished representations that result. The intractability that constitutes a limited resource comes in part from fundamental constraints on neural response properties and the architecture of cortex. Most basically, there is a limited number of neurons in the brain, and the idea that this might not be enough to represent every complex state of the world is an old one. In the context of the BCH, the idea is that neural coding models for representing huge correlation matrices might demand more units than are available. The other fundamental processing limitation comes from functional specialisation (Zeki, 1976, 1978; Wade and Bruce, 2001; Maunsell and Newsome, 1987) – different features are processed in different cortical regions and neurons have limited receptive fields (see page 57 and Section 2.4.3 below). Further, the loose cortical hierarchy that supports increasingly abstract representations results in increasingly large and specific RFs, presenting limitations in the fidelity with which relationships between different features can be represented. In the context of the BCH, this relates to constraints on representing correlations, and can be thought of as breaking a full posterior into factored, pseudo-independent chunks (see Chapter 6).

The goal of selection can be defined in different ways, spanning several traditional categories in selective attention research, and in exploring the literature it is important to distinguish these different ideas. First, selection for representation has sometimes been subsumed into selection for action – even with a perfect representation of the world, the brain must still select one saccade plan, one position of the body, one behavioural response. There is some evidence for an anatomical overlap between action-selection and selecting

a single focus of attention (see e.g. Rizzolatti et al., 1987), but as we discussed above, the progression from sensory epithelia to motor command is unlikely to be characterised by either a complete identity, or by two entirely discrete processes (see pages 21 and 61). Prior information about what is important, or what might be about to happen, is critical for selecting actions, but also appears to have an impact on representation (see Itti et al., 2005). In Chapter 6 we consider this relationship in a Bayesian framework, discussing how the computational goal of matching a necessarily approximate posterior as closely as possible to the true distribution improves representation in a way that is often tightly tied to action-relevant information.

In terms of selection for representation, the notion that there is a discrete operation allowed by attention, or a discrete 'bottleneck' through which only attended information can pass, has dominated the field. However, it is arguable that most such models have attempted to map a behavioural definition of a capacity limit too directly onto the form of the algorithm – informed by technological metaphor rather than implementational constraints. The latter point to a limited resource that is in fact distributed across the brain and manifested in many different neural properties, rather than a single bottleneck or filter. Another prominent way of thinking about selection for representation, and one that is instead directly *inspired* by implementational constraints, is selection for 'binding' – the idea that attention is required to combine information represented in separate cortical regions. But here it is arguable that this maps an coarse description of an implementational constraint (rather than a cognitive metaphor) too directly to the algorithmic level. All of these models have struggled to encompass the evidence for degrees of representational capacity without attention, and degrees of representational improvement with it. In the remainder of the literature review we survey these different approaches to selection in more detail, arguing that a reciprocal loop through Marrian levels is essential to tie down a phenomenon that has so many different goals, and involves every system of the brain.

2.4.1 Searching for Bottlenecks

The experimental study of attentional selection began in the 1950s, when the dominant metaphor for the mind was a serial information processing device. Models of the limited resource allocated by attention were thus dominated by discrete 'bottleneck' metaphors, such as Broadbent's influential Filter Theory, which conceptualised attention as a filter that chose which information would pass through to later stages of processing (Broadbent, 1958). Such models proved unable to encompass behavioural evidence that attention selects at different stages of processing in different contexts, characterised by an ultimately futile debate between 'early' and 'late' selection (see Cherry, 1953; Moray, 1959; Treisman, 1960;

Deutsch and Deutsch, 1963; Norman, 1968; Duncan, 1980). It also proved impossible to find a neural correlate of a discrete, single bottleneck or filter. Neither of these problems would be surprising from a consideration of the implementational constraints discussed above – the 'bottleneck' is distributed across the brain, and as such is likely to support different limitations in different processing domains.

The single filter approach was challenged by 'two-process' theories (Neisser, 1967) in which a pre-attentive, parallel processing stream can guide the allocation of an attention-demanding, deeper processing stage, and attenuation theories in which inputs have a graded likelihood of passing through to later processing stages (Treisman, 1960, 1969). The latter comes closer to embodying the kind of selection we propose, and also raise the important question of how much processing can be done without attention – theoretical studies have suggested that preattentive operations in early cortical regions can perform surprisingly complicated analysis (see e.g. Li, 2002). However, these models still embody algorithmic distinctions not supported by the implementational substrate.

Another influential proposal for reconciling evidence for early and late selection is the perceptual load theory of Lavie and Tsal (1994), which states that attention can only select when limited perceptual resources are overloaded. Experiments supported this idea, suggesting that the widespread challenges to early selection were in part due to the low perceptual load of the paradigms used to search for it (Lavie and Tsal, 1994; Lavie, 1995). However, there is something a little strange about the lack of control in low perceptual load, especially given that observers can guide attention far more finely in other situations, questioning how much this theory identifies the role of attention as opposed to the inferential consequences of having a limited perceptual resource (Dayan, 2008a; Yu et al., 2008). By thinking about the limited resource in Bayesian terms, we can make clearer distinctions between representational limits prior to attention, the limits of attentional processes themselves, and how the effect of attention depends on the interplay between the two.

Attention researchers have gradually moved towards a picture in which a serial 'filter' is just an abstract metaphor for a variety of selection processes (for reviews see Driver, 2001; Kinchla, 1992; Wolfe, 1998). Selection can occur on the basis of low-level physical attributes, or high-level semantic attributes deemed pertinent by memory or conscious control. 'Bottom-up' processing can influence 'top-down' processing, but there is not always a clear distinction between the two – neural processing is distributed and recurrent rather than purely serial and feedforward. Information can be processed to a variety of 'depths', and attention gates access to different kinds of processing – from simple physical analysis to conscious awareness – in different situations. This reflects the distributed constraints we reviewed above, but without being properly pinned down by them.

2.4.2 Psychophysical, Neurophysiological and Anatomical Effects

Cognitive models of selection were built on behavioural observations, but dealt in a restricted class of algorithms uninformed by evidence for the neural source of processing limits. More recently, there has been a focus on gathering more direct psychophysical and neurophysiological data on how attention affects stimulus representation – constraining algorithms in a more bottom-up way. A rich collection of psychophysical studies reveals improvements in the discrimination, detection, and identification of attentionally cued objects, but does not reveal a single, limited resource allocated by attention (see e.g. Itti et al., 2005; Driver, 2001). Concurrently, electrophysiologists have identified changes in the response properties of individual neurons when attention is directed to stimuli within their receptive field – changes which should underlie the observed psychophysical effects. Such effects have been reported throughout the brain, but are various, and variously interpreted (see Treue, 2001; Maunsell and McAdams, 2000; Dayan and Zemel, 1999; Gilbert et al., 2000; Ito and Gilbert, 1999; Motter, 1993; Reynolds et al., 2000; Roelfsema et al., 1998; Ghose and Maunsell, 2008). They include the observations that attention modulates local competition (Desimone, 1998), increases contrast gain (McAdams and Maunsell, 1999), and sharpens tuning functions (Spitzer et al., 1988). Ongoing attempts to reconcile these findings imply a desire to find a universal mechanism at a lower level, and their failure suggests yet again that attentional effects have a heterogenous basis in the brain.

One of the central debates in this area is whether neural contrast-response functions are *shifted* or *multiplied* when a stimulus within the neuron's receptive field is attended (see McAdams and Maunsell, 1999; McAdams and Reid, 2005; Williford and Maunsell, 2006). This relates to a debate in the cognitive domain about whether attentional improvements can be best characterised as noise reduction (Baldassi and Burr, 2000, 2004; Pashler, 1998), or as enhancing the contrast of a stimulus (Carrasco et al., 2004; Carrasco, 2005). SDT models have been used to argue first, that many effects attributed to the allocation of a limited resource are in fact due to changes in distractor noise, and second, that attention itself can be characterised as excluding distractor noise or irrelevant spatial locations (e.g. Dosher and Lu, 2000; Palmer, 1994; Morgan et al., 1998). The SDT approach is closely allied to the probabilistic, Bayesian decision-maker, but is restricted to discrete distributions for each object in the scene. The Bayesian extension to SDT uncertainty reduction we present in Chapter 6 helps to reconcile apparently contradictory behavioural findings, and shows how uncertainty reduction (potentially informed by prior information that guides action - see page 67) is coupled to improvements in representation. As with the SPRT model for decision-making discussed above (see page 60), an overly restrictive algorithmic formalism makes it hard to draw conclusions when the limiting assumptions of the model break down.

Anatomical investigations using fMRI concur with physiological evidence for attentional modulations throughout the brain, right down to striate cortex (Poghosyan and Ioannides, 2008; Brefczynski and DeYoe, 1999; Kastner et al., 1999, 1998; Ress et al., 2000; Somers et al., 1999). This again supports the picture gleaned from behaviour, and from considering the implementational source of processing limitations, of a mechanism that is distributed and continuous, rather than discrete and categorical. Attentional control signals have been identified with frontal and parietal areas (Corbetta and Shulman, 2002; Knudsen, 2007; Kastner and Ungerleider, 2000; Huddleston and DeYoe, 2008; Green et al., 2008) both for spatial and featural attention (Poghosyan and Ioannides, 2008), supporting the concept of top-down control of attention. However, there is ongoing debate about the exact role of these fronto-parietal areas and the sequence of their activation (see e.g. Green and McDonald, 2008; Sommer et al., 2008; Itti et al., 2005).

2.4.3 The Binding Problem and Feature Integration Theory

Looking at the psychophysical and neural effects of attention promotes tighter constraints on the kinds of algorithm that can be proposed for attentional selection and the benefits it offers. But the models derived from such observations have arguably suffered from the absence of a general computational level description – in particular they often fail to make it clear why the selection that produces attentional effects is necessary in the first place. Above we discussed more fundamental constraints on neural representation itself – constraints that necessitate selection rather than constraints on models of attentional effects. The idea that functional specialisation leads to a problem with representing relationships between separated cortical regions, and restricted RFs, has been described as a 'binding problem' that might be solved by attention.

In its most general form, the binding problem asks how the spatially distributed neural processing of different components of a task can be coordinated (Gray, 1999). In the visual system, this means asking how features represented in different cortical areas can be appropriately integrated or 'bound' into composite objects (see Robertson, 2005; Treisman, 1998). In 1980, Treisman and Gelade made the influential proposal that the role of attention was to solve the binding problem in a local 'spotlight'-like region – gluing together features into objects. Experimental support came from illusory conjunction experiments, in which observers will misbind simple features when their attention is distracted, for example reporting a green 'T' when presented with a red 'T' and a green 'L' (Treisman and Schmidt, 1982; Prinzmetal, 1981; Nissen, 1985). Visual search experiments found that the time taken to search for a target defined by a *conjunction* of features was linearly dependent on the number of items in the display, whilst the time taken to search for a single feature showed

no such dependence (e.g. Steinman, 1987). This was interpreted as revealing a sequence of attentional 'shifts', in which attention serially binds different items, in order to decide whether or not they are the target (Treisman, 1977, 1982, 1988; Treisman and Sato, 1990).

Implicit in Treisman's 'feature integration theory' (FIT) is the belief that what visual search and illusory conjunction experiments reveal is a genuine feature-binding problem instantiated in the architecture of the visual system. This is supported by anatomical and physiological evidence for cortical specialisation – different features are processed in different regions of the brain (Zeki, 1976, 1978; Wade and Bruce, 2001; Maunsell and Newsome, 1987), such that sorting out which features belong to the same objects is non-trivial. In this context, the feature maps of Treisman's theory were thought to correspond to cortically specialised, retinotopically arranged areas (see Treisman and Gelade, 1980), and the spotlight of attention to some kind of top-down signal from fronto-parietal areas (see Itti and Koch, 2001, for discussion of this correspondence). However, this again has the flavour of mapping implementational constraints too directly onto somewhat metaphorical algorithms.

And just like Broadbent's Filter Theory (Broadbent, 1958), FIT was dogged with an increasing list of exceptions, as researchers found conjunctions that were processed preattentively, and difficult feature searches that seemed to require attention. The distinction between serial, attention demanding search and parallel preattentive analysis has been repeatedly challenged (Pashler, 1987; Palmer, 1995; Geisler and Chou, 1995; Eckstein, 1998), as search behaviour has been found to depend on target-distractor similarity (Duncan and Humphreys, 1989; Palmer, 1994; Verghese and Nakayama, 1994), eccentricity (Carrasco et al., 1995), and lateral inhibition and masking (Wertheim et al., 2006). Questions have also been raised about the appropriate analysis of illusory conjunction experiments, and the role of memory and report mechanisms in apparent failures of binding (Ashby et al., 1996; Butler et al., 1991; Johnston and Pashler, 1990; Saarinen, 1996a,b; Neill, 1977). Theoretical studies have also shown how much of what attentional 'binding' was proposed to achieve could arise from simple neural networks inspired by the architecture of striate cortex (see e.g. Li, 2002).

To deal with the evidence for some limited binding without attention, Wolfe proposed 'Guided Search Theory', an extension of FIT in which an initial pre-attentive feature 'bundling' stage precedes binding (Wolfe et al., 1989). It is again difficult to motivate a clear distinction on the cognitive level – here between bundling and binding rather than between bound and not-bound. The notion of degrees of binding finds a natural expression in a computational, Bayesian approach, where binding judgements can be more or less accurate, and the effects of attention can boost accuracy at any point along the continuum from free floating to tightly bound features. SDT models have been used to argue that set-size effects can in fact be explained in terms of the increasing level of decision

noise as more distractors are added to the scene (Palmer, 1994). But there are examples of both simple (Cameron et al., 2004) and complex (Palmer, 1994) search tasks for which the appropriate SDT model leaves an unexplained increase in accuracy or reaction time that can be attributed to the allocation of an attentional resource. As mentioned above, a Bayesian extension of SDT ideas marries uncertainty reduction to the improvement of stimulus representation, potentially reconciling these effects.

The 'breaking' of the FIT algorithm as a description of behaviour is, as for filter theory, paralleled by a failure to find a discrete neural correlate of a binding operation. The idea that binding might occur by synchronising neuronal oscillations has been extensively discussed and investigated (e.g. von der Malsburg, 1995; Womelsdorf et al., 2007). In parallel, synchronous oscillations have also been proposed as a mechanism for attentional selection, with experimental evidence for increased synchrony between neurons that respond to an attended stimulus (e.g. Fries et al., 2001, 2008; Engel et al., 2001). Whether synchrony plays a role in both intra-feature attentional enhancement and inter-featural binding has not been resolved, and there is ongoing debate about the feasibility of a precise timing-based mechanism and its read-out (see Shadlen and Movshon, 1999; Salinas and Sejnowski, 2001; Averbeck et al., 2006, and page 47). A competing suggestion is that attention operates to decrease the RF of neurons such that only features belonging to the same object are contained within a particular cell's remit and are thus bound (Reynolds and Desimone, 1999), but it is not clear how this would function across different feature dimensions and in cortical regions with very large RFs. Ultimately, it seems likely that a constellation of neural mechanisms contribute to behaviourally defined 'binding' and 'selection' operations, and the distinction between binding vs. grouping for selection may be one without a clear neural correlate (see Engel et al., 2001).

Another approach to the difficulties encountered by FIT is to argue that the binding problem doesn't exist (see Ghose and Maunsell, 1999). This is based on the observation that cortex, rather than consisting of a parallel array of simple feature maps as FIT would suggest, embodies a hierarchy of increasingly complex representations that at the top can correspond to complex objects such as a person or tool (Barlow, 1972). Researchers using this principle to build models of invariant object recognition have claimed that their success implies that high level representations for all objects would suffice, leaving no role for attention in binding (see for discussion Riesenhuber and Poggio, 1999; Treisman, 1995). Perhaps for some restricted class of objects this could work, but the classical "grandmother cell" objection becomes pertinent as soon as you extend to the number of objects encountered in the world. Setting aside computational arguments against a punctate representation, under most representational schemes there simply aren't enough neurons to code for all possible feature combinations (Engel et al., 1992; Singer and Gray, 1995). One could argue that

since we discriminate objects, not features, this would avoid such a combinatorial explosion, but behavioural evidence speaks to great subtlety in our ability to categorise bindings of slightly different features, and to limitations on this ability (Wolfe and Cave, 1999).

Despite enduring problems with the interpretation of classic visual search and illusory conjunction tasks, it seems clear that there is a neural resource limitation, and that attention assists with judgements about colocated features (Desimone and Duncan, 1995). However, the search for a discrete 'binding' operation failed, suggesting that an attentional bottleneck cannot be characterised as the ability to bind free floating, simple features in one spatial location – yet again, we find that it is hard to locate a single mode of processing exclusively allowed by attention, and to identify the capacity of its selection. It seems that attention instead *improves* representations in different ways in different circumstances. With a probabilistic model that admits of degrees rather than distinctions, it is perfectly possible that the pre-attentive representation might sometimes be sufficient to explain performance, as in experiments suggesting that the ventriloquism effect is unaffected by attention (see Vroomen et al., 2001; Bertelson et al., 2000, and Section 2.1.2).

2.4.4 Modelling Selection

Computational models of bottom-up selection have had great success in predicting the eye movements of observers looking at natural scenes and videos (Itti and Koch, 2001; Itti et al., 1998; Torralba et al., 2006; Shipp, 2004), and are often assumed to imply that such overt selection might be mirrored in covert attentional shifts (see page 67). These models are inspired by the anatomical observations embodied in FIT (Treisman and Gelade, 1980), and by neurophysiological data on the responses of cells with and without attention (Treue, 2001). They compute stimulus salience on the basis of simple physical properties processed in separate feature-maps, then aggregate these signals to determine a winning location for the allocation of attention. What these models fail to explain is how the representation of a stimulus is altered by attention – they rather say which stimulus might be most demanding of top-down attention, whatever effect attention might then have.

The 'biased competition' framework of Duncan and Desimone also considers the mechanism of attentional selection, but rather than determining the most salient stimulus on the basis of bottom-up properties, their model starts from the idea that stimuli are constantly competing for representation in the visual system and for control of behaviour (Desimone and Duncan, 1995; Desimone, 1998). Many different neural mechanisms work to resolve this competition, and attention is seen as an emergent property of these mechanisms, essentially serving to bias the competition. The response properties of simple neural units in models

of how biased competition might work mimic some of the electrophysiological observations described above, and there is some direct experimental evidence for competitive interactions during attentional selection (e.g. Balan et al., 2008; Beck and Kastner, 2008). What is slightly unclear in this framework is the distinction between competition for *representation* and competition for *attention* – for example, models of competitive interactions fail to satisfactorily explain why only some neurons can shift or scale their tuning curves, and therefore why the performance enhancements that inhere on these changes can only occur for attended stimuli.

These theoretical approaches both face questions about the relationship between bottom-up and top-down allocation. Bottom-up salience modellers have recently 'bolted on' top-down influences on the competition for attentional allocation (e.g. Peters and Itti, 2007), though this again leaves it unclear what either signal does to stimulus representation. The biased competition model also allows for biasing signals to come from multiple sources (Desimone and Duncan, 1995; Desimone, 1998), but with multiple sources of attentional direction and a lack of clarity in distinguishing capacity limits in representation and attention it is difficult to see how they can explain a single focus of attention. In our probabilistic computational framework we try to make explicit the distinction between capacity limitations in representation and attention, and how various source of 'biasing' signal can be integrated into a coherent attentional mechanism. This expands on existing models of attention as a Bayesian prior (Dayan and Zemel, 1999; Rao, 2005), extending SDT insights about how noisy representations are affected by knowing which noise sources are irrelevant.

2.4.5 So Where's the Limited Resource?

In sum, research into attentional selection has tended to focus on when and how selection occurs, rather than on why it is necessary. This is perhaps a consequence of a phenomenon that is not only incredibly diverse, but needs to be simultaneously pinned down on all Marrian levels. Allowing behavioural data, or limited interpretations of it, to directly dictate the form of an algorithm can be dangerously metaphorical, as evidenced by the fate of cognitive bottleneck models and feature-integration theory. Constraints from detailed psychophysical data and from neurophysiological effects of attention can guide models of selection, but due to their local focus have tended to ignore the computational *goal* of the mechanisms they describe and why it is necessary. A consideration of the constraints that necessitate selection is key, encompassing fundamental properties of cortical architecture and neural receptive field properties. However, there is again a danger of mapping these constraints too simplistically onto the form of an algorithm, as with the proposal of a single spotlight of attention that is the only route to featural binding. If the question of why

selection is necessary is framed in computational terms, this danger is ameliorated, and we also have a way to address the nature of representation *without*, as well as *with* attention (see e.g. Li, 2002; Dayan, 2008a; Yu et al., 2008). SDT models take this approach, but in a restricted notation that cannot convincingly encompass signal enhancement effects. A Bayesian approach to limitations on probabilistic representation and inference explicitly lays out what is limited and why, and can be informed by implementational constraints. It also constitutes an important step towards expanding our picture of the Bayesian brain towards a more comprehensive theory of inference and decision-making in real-world contexts.

2.5 Conclusion of Literature Review

In the first half of this literature review we considered evidence for the Bayesian Coding Hypothesis. This evidence is promising, but limited, and falls short of the integrative trinity schematised in Figure 1.3. To tie together all three elements of this triangulation at once, simple behavioural tasks in domains with clearly identifiable neural substrates are needed. However, much behavioural evidence for Bayesian optimality involves cue combination or the integration of biased priors, often with relatively complex visual features that don't map easily onto simple PPCs. In Chapter 3 we develop a paradigm that can be used to demonstrate optimal inference with regard to the uncertainty in a single sensory posterior, by employing an external loss function. This allows us to demonstrate near-optimal behaviour for a simple offset stimulus likely to be supported by early visual processing, strengthening the evidence for ubiquitous processing of uncertainty and providing a potential substrate for future integrative experiments.

Simple optimality studies may provide the most watertight evidence for neural processing of uncertainty, but restrict the BCH to unrealistically simple settings – we argue throughout the thesis for extending the BCH to more complex domains. This raises serious challenges – how to measure and model Bayesian behaviour when it is not optimal, how to integrate Bayesian inference with cognitive functions such as reward-learning, attention, and memory, and how to build neural coding models that can encompass more complex, and approximate, inference and representation. In Chapter 5 we ask how the components of Bayesian decision theory map onto the neuroanatomical correlates identified in neuroscientific studies of decision making, specifically how sensory uncertainty is integrated with externally manipulated utility. To do so, we use the same paradigm as in Chapter 3, but with complex face-house mixtures rather than simple offset stimuli. Suboptimal behaviour in this task, reported in Chapter 4, raises critical questions about when Bayesian inference is optimal, and how we can characterise non-optimal or approximate Bayesian inference.

In Chapter 6 we explicitly consider what happens when the brain can't represent the full posterior over a complex scene, linking this to traditional concepts of a limited capacity resource. We then consider how selective attention might operate in this framework, arguing for the utility of a computational perspective in guiding theories of attention. Bayesian methods are used in machine learning to approach highly complex inference problems – here we make some initial steps towards bringing this approach to the BCH.

3

Bayesian Decision Making with Simple Visual Uncertainty

In this chapter we attempt to extend the evidence for Bayesian behaviour in unimodal sensory perception, and maximise the ability of this evidence to argue for the BCH. We show that human observers performing a simple visual choice task, under an externally imposed loss function, approach the optimal strategy, as defined by Bayesian probability and decision theory (Cox, 1961; Berger, 1985). To behave optimally requires observers to utilise an estimate of their internal uncertainty, rather than simply a modal estimate of the uncertain stimulus. In concert with earlier studies, this suggests that observers possess a model of their internal uncertainty, and utilise this model in the neural computations that underlie their behaviour (Knill and Pouget, 2004). However, unlike most previous studies we look at a simple quantity likely to be represented in early visual cortex, rather than cue combination scenarios, or the computation of higher-level variables such as motion or depth. This suggests that representations of uncertainty are widespread throughout the cortical hierarchy, and can be propagated to decision-making regions. Crucially, observers in our study approach optimal behaviour even when denied the opportunity to learn adaptive decision strategies based on immediate feedback. Alongside the observation that behaviour responded to changes in uncertainty over experimental sessions, this supports the conclusion that representations of uncertainty are also pre-existing and flexible, playing an online role in perception. It also provides a potential substrate for making tight integrative links between Bayes-optimal behaviour and neural data.

3.1 Introduction

As discussed in the literature review, evidence for the BBH consists in showing that Bayesian formulations of perceptual or sensorimotor computations provide an accurate description of behaviour (see Section 2.1). There are two main strategies for making this evidence stronger – the first is to show that people produce optimal behaviour when this requires the use of information about uncertainty rather than simply modal estimates (usually the mean) of the stimulus or motor outcome (e.g. Ernst and Banks, 2002; Knill and Saunders, 2003; Trommershauser et al., 2003b), and the second is to explain apparent suboptimal biases via ecological Bayesian priors (e.g. Jacobs, 1999; Stocker and Simoncelli, 2006a) or changes in the likelihood (e.g. Stocker and Simoncelli, 2006b). Taken at face value, these studies argue strongly that certain elements of human behaviour are well described by Bayesian inference. What the proponent of the BCH would like is for behaviour to provide evidence for *neural implementation* of probabilistic representation and computation.

Behavioural evidence can't be directly linked to neural firing – intervening neural coding models are required – but it can be used to argue indirectly for neural representations of Bayesian uncertainty. Critically, we can rule out alternative strategies by which the brain could mimic Bayes-optimal behaviour without having to use uncertainty in its computations. Feedback-driven incremental threshold adjustment is one such strategy, and not all behavioural studies have considered this possibility (but see for example Trommershauser et al., 2003b). Here we provide only periodic cumulative feedback, thus ruling out the use of at least simple adaptive threshold schemes.

In broadening the picture of the Bayesian brain it is important to map out where, and when, behaviour is optimal, which will both improve our understanding both of the constraints on optimal inference, and provide a richer group of neural substrates to investigate in integrative experiments. Much existing evidence in the perceptual domain concerns cue combination studies, in which signals from different modalities or different functionally specialised cortical regions must be integrated to reach a conclusion about their common cause (see Section 2.1.1 and e.g. Jacobs, 1999; Ernst and Banks, 2002), or from the integration of prior biases with likelihoods (see Section 2.1.4 and e.g. Stocker and Simoncelli, 2006a). Both these approaches are founded in showing that belief distributions can be combined according to the uncertainty in each - either that two distributions with different uncertainties can be weighted appropriately, or that the impact of the prior is appropriate to the uncertainty in the likelihood. A key question is whether uncertainty in individual sensory distributions is used in perception, but there is little evidence for Bayesian inference with respect to a single likelihood. This is primarily because the optimal estimate in such infer-

ence is the mean of the distribution - the same as in a system that ignores uncertainty (but see Landy et al., 2007).

We addressed this issue by adapting an asymmetric loss function paradigm historically used in SDT psychophysics to explore the Receiver Operating Characteristic (ROC)[1], and more recently employed to probe the use of motor uncertainty in sensorimotor computations (see Section 2.1.3 and e.g. Trommershauser et al., 2003b; Kording et al., 2007). In our task, the observer had to make a simple offset or 'Vernier' discrimination (Westheimer, 1979), and we imposed an asymmetric loss function on their judgements – answering 'left' incorrectly could carry a different penalty to answering 'right' incorrectly. To behave optimally, observers must combine an estimate of sensory uncertainty, at least in the form of a likelihood ratio, with knowledge about the external loss function (Cox, 1961; Berger, 1985). In this study the stimulus was very simple, and perceptual uncertainty was due to sensory noise and subsequent processing rather than to external manipulations. This enabled us to ask whether even for a very simple visual task the brain has an estimate of internal uncertainty available to guide behaviour (see also Schwartz et al., 2006), and suggests that information passed to decision-making regions consists of more than a modal estimate of the stimulus. In the notation of Figure 2.5, optimal behaviour involves a bias whose extent is dependent on uncertainty, but where this bias arises from $p(U \mid o_j)$, rather than from the contribution of a biased prior to $p(o_j \mid d_i)$. In Chapter 4, we use the same paradigm to ask whether Bayesian inference operates for stimuli at the other end of the complexity spectrum, using a stimulus axis composed of varying blends of face and house stimuli.

Related to the need to rule out incremental learning strategies, indirect behavioural evidence for the BCH is made stronger if uncertainty can be shown to be flexibly represented in a ubiquitous and online fashion, rather than learnt in a limited and 'model-free' way for particular tasks. In the present study observers took part in two experimental sessions on different days, and the majority achieved optimal performance in both sessions despite changes in sensory uncertainty. The present results provide strong evidence that uncertainty, at least in simple perceptual contexts, is represented online even for early featural analysis, and is available to decision-making regions of the brain.

[1]The ROC plots the proportion of correct 'yes' answers against the proportion of incorrect 'yes' answers in a signal detection task. The area underneath the ROC curve gives the observer's sensitivity, and points along the curve are plotted out by changing the decision criterion, for example by changing the loss function (see Green and Swets, 1966).

3.2 Methods

3.2.1 Observers

Four participants (2 male, 2 female) took part in the experiment. They had a mean age of 25.5 years, were all right-handed, and had normal or corrected-to-normal vision. Three were entirely naive with respect to the aims of, and theory behind, the experiment; and the fourth (observer 4) was the author of this thesis (LW).

3.2.2 Stimulus and Equipment

The stimulus consisted of a pair of vertically arranged Gabor-like patches, in which a sinusoidal grating with a spatial frequency of 0.03 cycles/mm (0.21 cycles/° of visual angle), was multiplied by a Gaussian envelope with a characteristic width (2σ) of 29.9 mm, truncated at a full width of 100 mm (14.4° of visual angle) in the horizontal direction, and a rectangular envelope with a width of 10.3 mm (1.48° of visual angle) in the vertical (see Figure 3.1a). The pixel intensity in the two patches ranged from 0 to 255 (black to white) against a grey background of intensity 128. The separation of the two patches was 5.67 mm (0.813° of visual angle), and the stimulus appeared with the centre of the upper patch located 66.7 mm (9.56° of visual angle) either to the left or right of the fixation cross in a pseudorandomised order. On each trial the entire lower patch (both envelope and grating) was displaced relative to the upper patch by one of 20 pseudorandomised values, ranging from -15 to +15 pixels (positive numbers indicating offsets to the right). Each pixel corresponded to 0.333 mm (0.0478° of visual angle).

Before each block of main trials, observers were given a short block of practice trials in which the stimulus duration was 500ms. In main trials the stimulus duration was 160ms, which is shorter than the latency for initiating a saccade (Carpenter, 1988; Hodgson, 2002), and was chosen to avoid fixation of the stimuli. There was a randomized delay period of 750-1250 ms between the time of response and the time of presentation of the next stimulus.

The experimental program was written in MATLAB (The Mathworks Inc, Natick, MA), using the Psychophysics Toolbox extensions (Pelli, 1997; Brainard, 1997).

3.2.3 Procedure

Observers sat at a table in front of a computer screen, and placed their head on a chin rest such that the perpendicular distance from their eyes to the screen was 400 mm. During the experiment, observers fixated a central cross on the screen, and were asked to make simple Vernier judgments (Westheimer, 1979), reporting whether the lower of the two Gabor patches was offset to the right or left of the upper one. The task is depicted schematically in Figure 3.1a. Responses were provided using a computer keyboard.

We imposed an asymmetric loss function to probe the observers' representation of uncertainty. On each trial observers were awarded points for a correct answer ('rewards'), or had points taken away for an incorrect answer ('costs'). Observers were instructed, and given an incentive, to maximise the cumulative number of points scored during the experiment. The loss function varied between blocks of trials — the rewards for correctly answering 'right' ($\mathcal{R}_r$) or 'left' ($\mathcal{R}_l$) were constant and equal, but the cost for incorrectly answering 'right' ($\mathcal{C}_r$) could be different from that for incorrectly answering 'left' ($\mathcal{C}_l$). A similar asymmetric penalty approach has been used to probe uncertainty in recent studies of motor planning (Trommershauser et al., 2003a, 2005). When the penalty for answering 'left' incorrectly is greater, a reasonable strategy would be to answer 'right' more often when uncertain of the answer. This would result in a psychometric curve shifted in the direction of the higher penalty, yielding a higher overall score (see Figure 3.2a). This loss function maps deterministically from a particular outcome to an externally defined reward or 'utility' – under the notation introduced in Figure 2.5, sensory uncertainty in the mapping from categorical decision to correct vs. incorrect outcome; $p(o_j \mid d_i)$, is combined with a loss function consisting of deterministic mappings; $U(o_j)$.

We used five different sets of costs and rewards, listed in Table 3.1. The final column in this table shows, qualitatively, the relative shift we would expect for each such loss function according to the strategy just described. In the Results section we describe an optimal Bayesian observer analysis which confirms, and quantifies, the optimality of this 'curve shifting' strategy, by computing the psychometric function that gives the maximum expected utility (EU; see Equation 2.9). The different loss functions were presented in a counterbalanced, pseudorandomised block design. This was repeated in two experimental sessions on separate days, to increase the amount of data collected. Visual inspection of the data suggested that the level of observers' internal uncertainty differed in the two sessions, which was confirmed by Bayesian model comparison and may have been due to perceptual learning, consolidation (e.g. during sleep - see Maquet (2001)) or extrinsic factors. This then served as a further test of our hypothesis — if observers can behave optimally in two

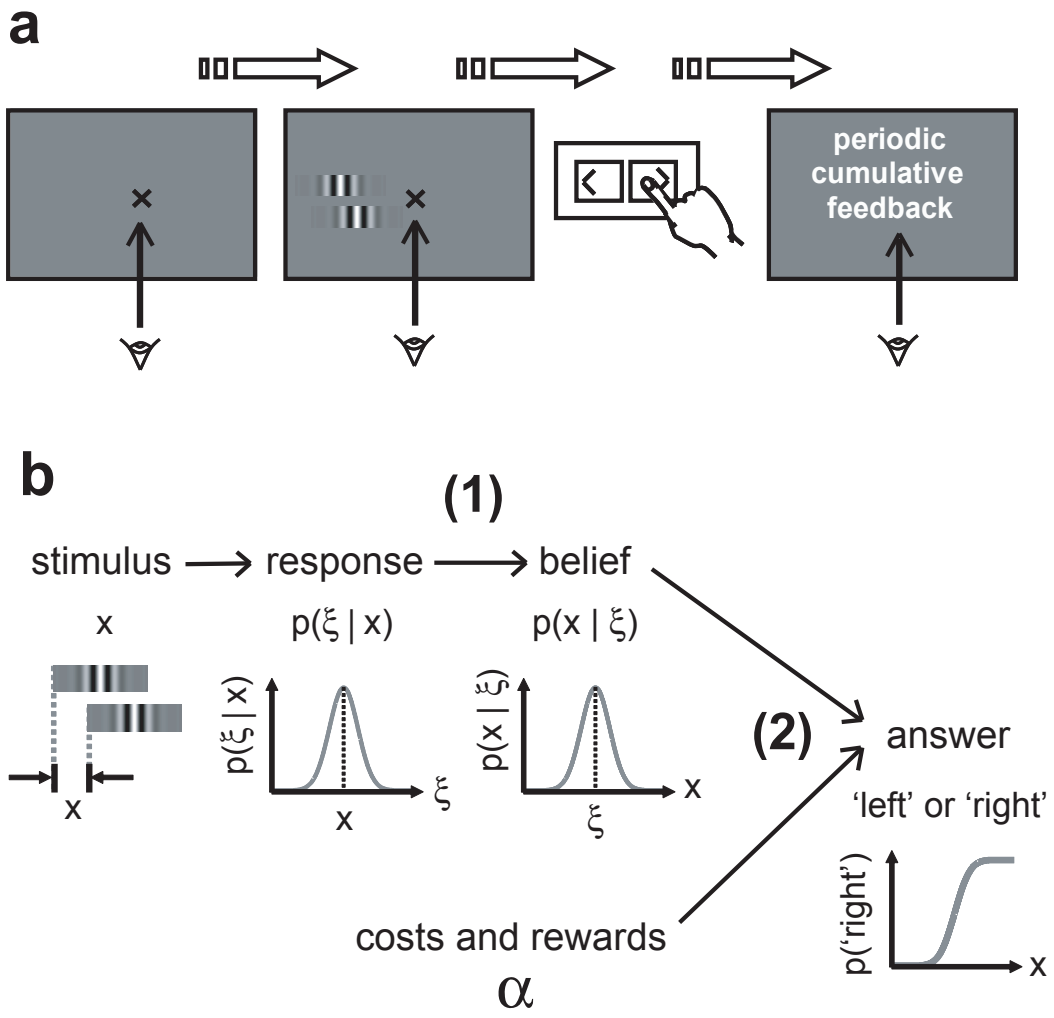

Figure 3.1: ***Experimental design.*** ***a,*** *On each trial a stimulus consisting of two vertically arranged Gabor patches was briefly flashed, and the observer pressed a key to say whether the lower patch was offset to the left or right of the upper patch. Participants were asked to maximize their score, with varying numbers of points being awarded for a correct answer ('reward') and deducted for an incorrect answer ('cost'). They received only periodic feedback about their performance, in the form of a cumulative score every 15 trials.* ***b,*** *Schematic of the quantities and transformations involved in the construction of the Bayes-optimal observer. The stimulus x produces a stochastic neural response. The observer transforms this neural response into a belief distribution (see (1)), and this belief is then combined with a loss function that expresses the cost or reward of each outcome, and decisions make in order to maximise expected utility (see (2)).*

Table 3.1: ***Values of α corresponding to costs and rewards.*** *This table gives the five different loss functions, in which the costs $\mathcal{C}$ and rewards $\mathcal{R}$ for answering 'right' and 'left' were manipulated as described by the quantity α. According to Bayesian decision theory, and as illustrated by the arrows in the final column of the table, this prompts a shift of the psychometric function in the direction of the lower cost, and by an amount proportional to the penalty asymmetry.*

$\mathcal{R}_r = \mathcal{R}_l$	$\mathcal{C}_l$	$\mathcal{C}_r$	α	curve shift
+20	-10	-50	0.3	$\longrightarrow$
+20	-20	-40	0.4	$\rightarrow$
+20	-30	-30	0.5	0
+20	-40	-20	0.6	$\leftarrow$
+20	-50	-10	0.7	$\longleftarrow$

sessions with *different* noise distributions, this supports the claim that they carry a flexible representation of current internal uncertainty.

Each block consisted of a short practice session (60 trials), and a main session (260 trials). In the main session feedback on performance was provided only every 15 trials, when observers were shown the total score they obtained in the last 15 trials, as well as their cumulative total in the block so far. The sparseness of this feedback made it unlikely that observers could learn an optimal internal threshold by an adaptive threshold adjustment strategy. Control analyses (see Section 3.3.7) support this view. In the practice session, observers received trial-by-trial feedback to familiarize them with the cost values for that block, and encourage consistent performance in the main trials. However, the practice stimuli were presented for 500ms rather than 160ms, which made the task much easier. As the effective internal noise should therefore be different for the practice stimuli, feedback in the practice session could not be used to adaptively learn a response threshold relevant to the main-block trials. In addition, the easier stimulus meant that there were very few trials on which observers failed to give the correct answer, implying that there should have been very little uncertainty in their belief. This further limits the likelihood that they could use the practice session to test alternative strategies for dealing with the loss function.

An instruction screen appeared at the beginning of each block, and after each feedback screen, reminding observers of the task and costs for that block. After the experiment was finished, observers were debriefed using a simple questionnaire. Participants were paid according to standard UCL protocol, with a score-related bonus in gift vouchers to motivate concentrated performance and encourage observers to try to maximise their total score.

3.3 Results

3.3.1 Bayesian Optimal Observer Analysis

The final column of Table 3.1 shows the relative shift of the psychometric function we might expect for different settings of the costs and rewards, under an intuitive strategy for maximizing score in which observers shift their psychometric function in the direction with the higher penalty. A quantitative Bayesian decision theory analysis was used to confirm and quantify the optimality of this strategy. Figure 3.1b depicts the quantities involved in this analysis. The visual stimulus, with a Vernier offset x, evokes a stochastic neural response, on the basis of which the observer constructs an internal belief about the value of the stimulus offset (step (1) in Figure 3.1b). This belief is then used to guide a decision about the appropriate response (step (2)).

An individual observer's responses to repeated presentations of the same visual stimulus may vary, bringing stochasticity to $p(o_j \mid d_i)$ – the probability of being correct vs. incorrect when answering either side of the categorical offset boundary. We assume that this variability arises from at least two separate sources of noise. The first source perturbs the observer's sensory estimate of the Vernier offset by a random additive increment. This creates a noise distribution centred on the stimulus, the uncertainty due to which is reflected in the observer's belief. The second source affects the observer's decision directly, in a way that is independent of the value of the stimulus offset. We may think of this as 'decision noise', or as the result of motor errors or of lapses in attention. We do not expect this source of variation to affect the observer's internal belief about the value of the offset, and so it is neglected in the theoretical development below. When modelling experimental responses, however, we introduce a separate 'lapse rate' parameter, so that these stimulus-independent errors do not corrupt our estimate of the stimulus-centred noise. Note that we do not distinguish between stimulus-centred *sensory* noise, and any stimulus- or estimate-centred *decision* noise which might, for instance, arise as sensory information is integrated with the loss function. Our definition of Bayesian optimality in decisions thus refers to all stimulus-centred variation. In concert with earlier analyses, we do however assume that the majority of this variation is 'sensory noise', and so treat and refer to it as such.

In keeping with the standard psychophysical treatments of sensory noise, our model assumes that the internal estimate of the Vernier offset, ξ, is normally distributed with constant variance σ^2 around the true stimulus offset: $p(\xi|x) = \mathcal{N}(\xi; x, \sigma^2)$ (Thurstone, 1927; Green and Swets, 1966). We test this assumption below, showing that the psychometric curves were all well fit by cumulative normal functions, with a constant slope parameter for

each observer in each session. However, our observers each displayed a systematic bias in their responses; this will be addressed later.

In the Bayesian view, the observer's belief about the Vernier offset x is not limited to a single estimated value ξ. Instead, ξ parameterises a belief distribution over all possible values of x that are consistent with the sensory evidence. The optimal form for this belief distribution is given by Bayes' rule:

$$p(x|\xi) = \frac{p(\xi|x) \cdot p(x)}{p(\xi)} \ . \tag{3.1}$$

We assume that the prior belief about x is uniform, which implies that this optimal belief will also be Gaussian, with the same variance as the sensory noise distribution, and mean given by ξ: $p(x|\xi) = \mathcal{N}(x; \xi, \sigma^2)$ (we might also have assumed a broad zero-centred Gaussian prior, although then the variance of the posterior belief would have been slightly smaller than that of the sensory noise, for which there was no evidence in the data). In fact, if observers are able to learn the true distribution of x, their prior (and therefore posterior) belief should take the form of a series of delta functions located at each discrete offset value used. In addition, for extreme values of x, the stochastic response ξ may fall outside the range of possible values, distorting the posterior. However, variability in decisions at the extremes was minimal, so that any divergence from normality at those points would have little impact on estimates of sensory variance. And within the central range, where decisions did vary, the values of the stimulus offset used were very closely spaced and we saw no evidence that observers were aware of the discretisation.

The observer must base his or her response on the belief distribution (step (2) in Figure 3.1b), and Bayesian decision theory gives the optimal form of this response (see Berger, 1985; Maloney, 2002; Yuille and Bulthoff, 1996, and Equation 1.3). The observer should answer 'right' if and only if, on the basis of his or her belief, the expected reward or utility for answering 'right' is greater than that for answering 'left': i.e. if $\mathbf{EU}[\text{'right'}] > \mathbf{EU}[\text{'left'}]$. In this simple case, expected utiltiy is obtained by adding together the product of the probability of the answer being correct times the corresponding reward, and the probability of the answer being incorrect times the corresponding cost. These two probabilities express the degree of the participant's belief that the lower patch fell to the right or left of the upper patch, and are given by the areas under the belief distribution $p(x|\xi)$ that fall of to

the right and to the left of 0 respectively.

$$
\begin{aligned}
P(\text{answer 'right'}) &= P\left(\mathbf{EU}[\text{'right'}] > \mathbf{EU}[\text{'left'}]\right), && (3.2)\\
\mathbf{EU}[\text{'right'}] &= \int_0^\infty p(x|\xi)\cdot\mathcal{R}_r \; dx + \int_{-\infty}^0 p(x|\xi)\cdot\mathcal{C}_r \; dx, && (3.3)\\
\mathbf{EU}[\text{'left'}] &= \int_0^\infty p(x|\xi)\cdot\mathcal{C}_l \; dx + \int_{-\infty}^0 p(x|\xi)\cdot\mathcal{R}_l \; dx. && (3.4)
\end{aligned}
$$

With some rearrangement, and combination of the integrals, we arrive at an expression in which the observer's decision about whether the Vernier displacement was to the right or left on a particular trial is given by a Heaviside function (H) that compares their belief distribution to a single quantity, α, which includes all the cost and reward terms (the values of α for each set of costs and rewards used in our experiment are given in Table 3.1);

$$
A_r(\xi,\alpha) = H\left[\int_{-\infty}^0 p(x'|\xi)\; dx' \; - \alpha\right]; \qquad \alpha \equiv \frac{\mathcal{C}_l - \mathcal{R}_r}{\mathcal{C}_l - \mathcal{R}_r + \mathcal{C}_r - \mathcal{R}_l}, \tag{3.5}
$$

where we have introduced the notation $A_r(\xi,\alpha)$ for the Bayesian decision given a particular ξ and α: $A_r = 1$ corresponds to the observer answering "right", and $A_r = 0$ to the observer answering "left", and x' is a dummy variable of integration over the observer's belief.

It is useful at this point to introduce a normal density function which has the same width as $p(\xi|x)$ and $p(x|\xi)$, but zero mean: $f_\sigma(\zeta) = \exp(-\zeta^2/2\sigma^2)/\sqrt{2\pi\sigma^2}$. Thus $p(\xi|x) = f_\sigma(\xi - x)$ and $p(x|\xi) = f_\sigma(x - \xi)$, and the corresponding cumulative function is $\Phi_\sigma(\zeta) = \int_{-\infty}^{\zeta} f_\sigma(\zeta')d\zeta'$. Then,

$$
\begin{aligned}
A_r(\xi,\alpha) &= H\left[\int_{-\infty}^0 f_\sigma(x'-\xi)\; dx' - \alpha\right] && (3.6)\\
&= H\left[\int_{-\infty}^{-\xi} f_\sigma(\zeta)\; d\zeta < \alpha\right] \qquad [\text{where } \zeta = x' - \xi] && (3.7)\\
&= H\left[\Phi_\sigma(-\xi) - \alpha\right] && (3.8)\\
&= H\left[\xi - \Phi_\sigma^{-1}(\alpha)\right]. && (3.9)
\end{aligned}
$$

The probability that the observer answers "right" for a particular stimulus x and cost structure α can then be found by integrating the assumed sensory noise distribution:

$$P\left(A_r = 1|x, \alpha\right) = \int_{-\Phi_\sigma^{-1}(\alpha)}^{\infty} p(\xi|x) \;\; d\xi. \tag{3.10}$$

If we again insert Φ_σ (and exploit its symmetry) we obtain the following easily computed expression for the optimal probability with which the observer should answer 'right' for a given α and x value;

$$\begin{aligned} P\left(A_r = 1|x, \alpha\right) &= \int_{-\Phi_\sigma^{-1}(\alpha)}^{\infty} f_\sigma(\xi - x) \; d\xi & (3.11)\\ &= \int_{-\infty}^{x+\Phi_\sigma^{-1}(\alpha)} f_\sigma(\zeta') \; d\zeta' \qquad [\text{where } \zeta' = x - \xi] & (3.12)\\ &= \Phi_\sigma(x + \Phi_\sigma^{-1}(\alpha)). & (3.13) \end{aligned}$$

The only unknown quantity in Equation 3.13 is the standard deviation, σ, of the zero-mean cumulative Gaussian Φ_σ. This plays two roles in our analysis; it is both the width of the sensory noise distribution, and, under the assumed uniform prior, the width of the consequent belief distribution. Under the symmetric cost condition ($\alpha = 0.5$) the observer's decision reflects only whether the mean of his or her belief lies to the left or right of 0 (according to sensory noise), and is independent of the width of the belief distribution. Thus, following the standard psychometric approach, we estimate the variance of the noise by fitting a psychometric function based on a cumulative Gaussian to the behavioural data, with the slope of the function providing an estimate of σ (as illustrated in the upper grey box in Figure 2.5).

The Bayesian decision theory analysis expressed in Equation 3.13 makes two predictions: as α changes, the psychometric curves (1) retain the same cumulative-normal shape, with the same width parameter, and (2) translate by an amount $\Phi_\sigma^{-1}(\alpha)$. Figure 3.2b shows an example of the psychometric function fit to the data for one observer in one session, and shifts with changing α are clearly visible.

The fitting procedure, and the methods used to test these predictions, are detailed below. Briefly, we first verified that the shape of the psychometric function did not change with α via Bayesian model comparison (BMC). We then tested the agreement of the observed curve translations with those predicted by the optimal Bayesian analysis. We fit psychometric functions to the data and measured the centre μ of each, i.e. the value of x at which the

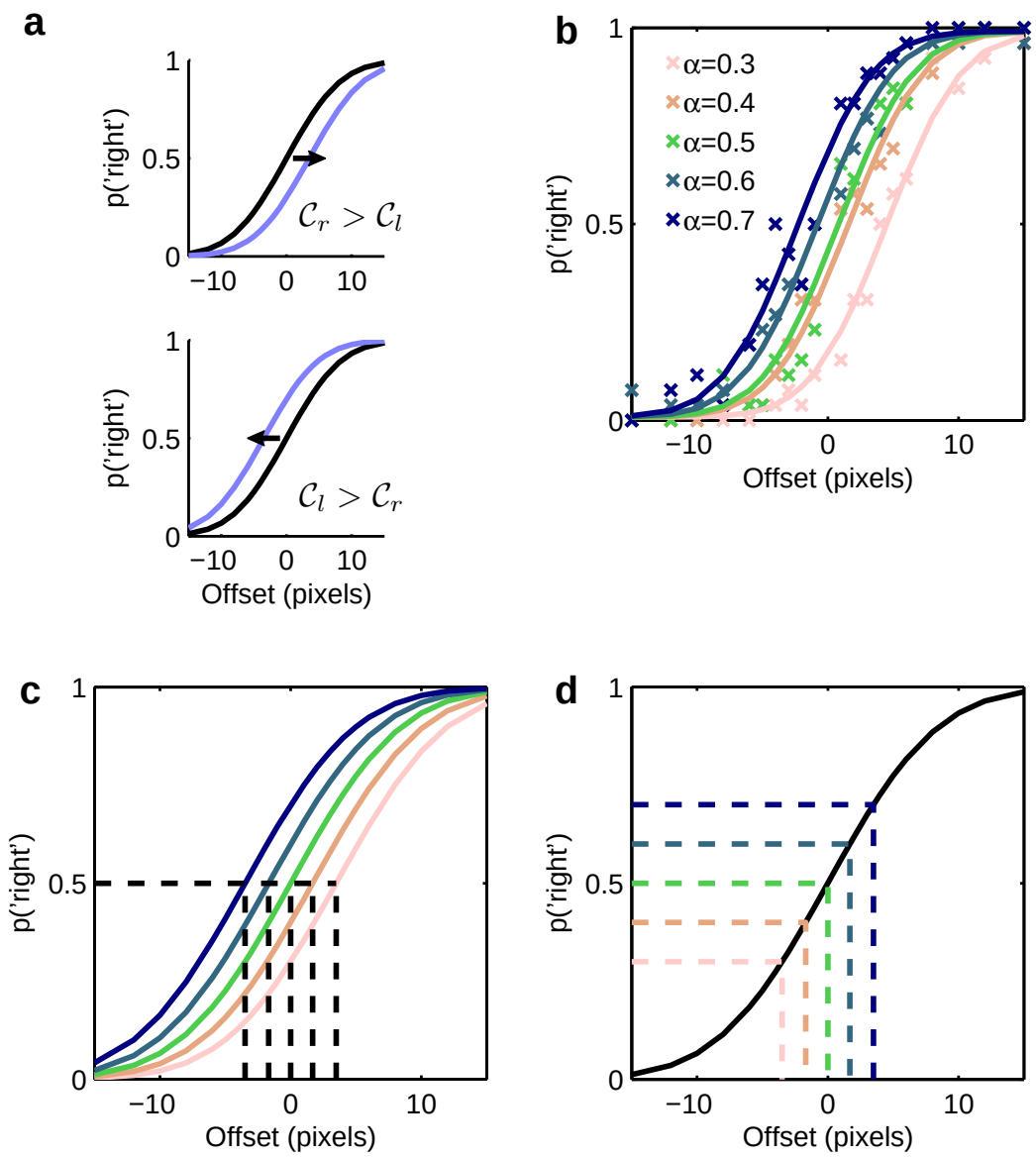

Figure 3.2: ***Evaluating behavioural optimality.*** ***a,*** *Illustrates the qualitative prediction for maximising reward – observers should give the answer with the lower penalty when uncertain, resulting in a shift of the psychometric curve in the direction of the higher cost.* ***b,*** *Example data from one observer in the five different α conditions. Crosses show data points, and the smooth lines show psychometric functions fit to the data, with the slope constrained to be the same for each α condition* ***c,*** *Illustrates the procedure for measuring observed shifts, once psychometric functions have been fit to the data from the five α conditions.* ***d,*** *Illustrates the procedure of taking the inverse value of the psychometric function at the values of α used in the experiment. The optimal shift between two psychometric curves is then given by the difference between the two corresponding inverse values.*

fitted psychometric curve gave equal probabilities of each answer, given by the mean of the underlying Gaussian (see Figure 3.2c). We then used the maximal slope of the psychometric functions as a measure of σ, and inverted Equation 3.13 to recover the predicted *optimal* values of the centre μ_j^* for each cost asymmetry value α_j (see Figure 3.2d).

$$\begin{aligned} 0.5 &= \Phi_\sigma(\mu_j^* + \Phi_\sigma^{-1}(\alpha_j)), \\ \mu_j^* &= \Phi_\sigma^{-1}(0.5) - \Phi_\sigma^{-1}(\alpha_j), \\ \mu_j^* &= -\Phi_\sigma^{-1}(\alpha_j). \end{aligned} \tag{3.14}$$

3.3.2 Fitting the Psychometric Function

The pattern of observers' responses was modelled by a cumulative normal psychometric function incorporating a random *lapse* term (see for example Wichmann and Hill, 2001), and binomially distributed response counts. We used 20 different true offsets x_i, and 5 different cost asymmetries α_j, with N_{ij} trials in each condition. The number of trials n_{ij} in which observers answer 'right' for stimulus offset x_i and cost distribution α_j, is assumed to be drawn from a binomial distribution:

$$P(n_{ij}) = \binom{N_{ij}}{n_{ij}} p_{ij}^{n_{ij}} \, (1 - p_{ij})^{N_{ij} - n_{ij}}. \tag{3.15}$$

In the absence of lapses, the optimal probability p_{ij} should be given by $P\left(A_r = 1 | x_i, \alpha_j\right)$ in Equation 3.13, which has a cumulative normal form. To fit the data, we therefore also assumed an underlying cumulative Gaussian shape, parameterised in terms of the standard error function, such that the parameters μ_j and ρ_j gave the centre and maximal slope, respectively, of the curve under the jth value of α.

$$p_{ij}^{\textit{no lapse}} = \frac{1 + \mathrm{erf}(\sqrt{\pi} \cdot \rho_j \cdot (x_i - \mu_j))}{2}; \qquad \mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt. \tag{3.16}$$

However, it is likely that observers occasionally make errors due to stimulus-independent (but possibly cost-structure-dependent) sources such as 'decision noise', motor response errors, or moments of inattention (Green and Swets, 1966; Wichmann and Hill, 2001). In this case they might give either answer with equal probability, effectively setting p_{ij} in such

cases to $\frac{1}{2}$ rather than the value given above. We took the probability of such an event occurring in any trial to be ϵ_j (the 'lapse rate' parameter referred to above), leaving the probability that the response was instead based on the cumulative Gaussian function as $1 - \epsilon_j$;

$$p_{ij} = (1 - \epsilon_j) \cdot \left[\frac{1 + \mathrm{erf}(\sqrt{\pi} \cdot \rho_j \cdot (x_i - \mu_j))}{2} \right] + \epsilon_j \cdot \tfrac{1}{2} \ . \tag{3.17}$$

There are thus three parameters, all of which potentially depend on α: the centre μ_j and slope ρ_j of the cumulative Gaussian, and the random error or lapse rate ϵ_j. An estimate of the slope parameter ρ_j provides an estimate of the width of the underlying Gaussian, according to;

$$\sigma = \frac{1}{\sqrt{2\,\pi}\,\rho} \ . \tag{3.18}$$

3.3.3 Shape of the Psychometric Curves

The Bayesian analysis predicts that, as the loss function varies, the psychometric curve will shift, but will retain both the cumulative Gaussian shape, as well as the same maximal slope. We tested both of these predictions.

To ask whether the cumulative Gaussian model with allowance for lapses (Equation 3.17) was appropriate for the data at all values of α, we examined the residual error between the measured response data and the best fit psychometric curve. Figure 3.3 shows the deviance residuals (McCullagh and Nelder, 1989; Wichmann and Hill, 2001) for all four participants, for each of the two sessions. The deviance residual is used to measure discrepencies in terms of the underlying likelihood model; in effect, it rescales the error by the locally predicted variance. Based on the total deviance, the cumulative normal model could not be rejected by a degrees-of-freedom-adjusted χ^2-test, nor by a Monte-Carlo-based exact-binomial test (Wichmann and Hill, 2001) ($p > 0.3$ and $p > 0.8$ respectively, after correcting for multiple tests; in neither case could the distribution of p-values over the multiple tests be distinguished from uniform; Kolmogorov-Smirnov test, $p > 0.05$).

There is also no systematic trend evident in Figure 3.3 to suggest that the *shape* of the psychometric function was inappropriate for any value of α, for any observer. This was confirmed using a runs test for randomness of the sign of the residuals, by which the hypothesis that the scatter of residuals was random could not be rejected ($p > 0.7$ after multiple-test correction, p-values uniform by Kolmogorov-Smirnov test, $p > 0.05$).

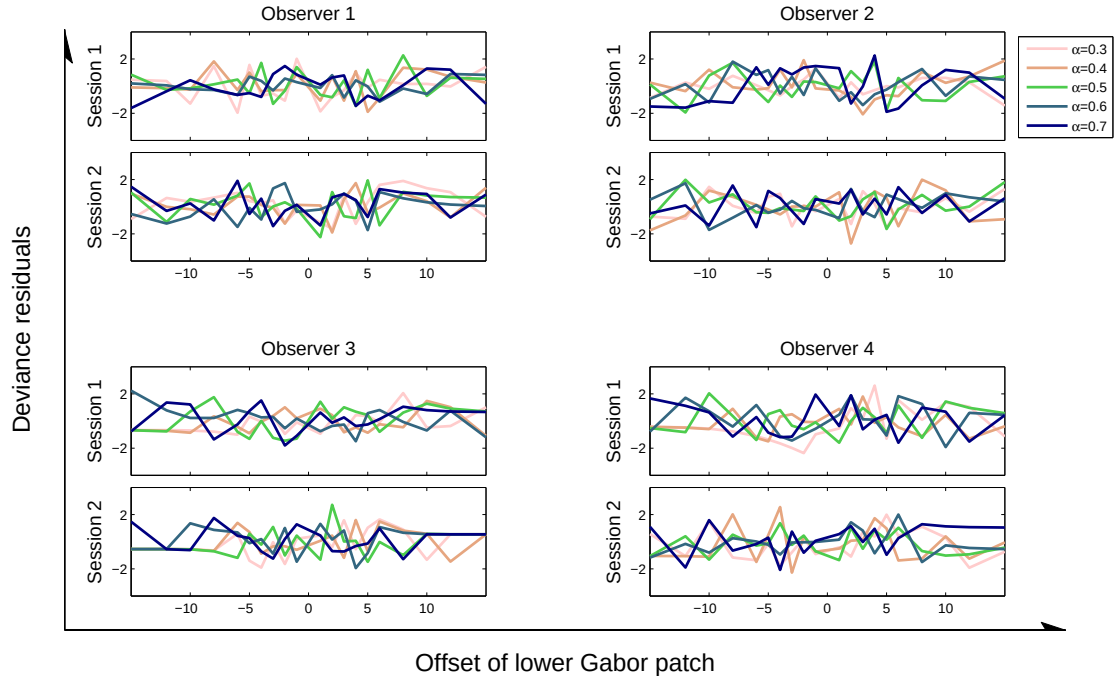

Figure 3.3: ***Deviance residuals between model and data.*** *Deviance residuals between the model fit to the behavioural data and the data points, for each observer in each session.*

The second prediction was that the slope of the psychometric curve is also independent of the value of α; that is, the parameters ρ_j in Equation 3.17 are, in fact, all the same. Visual inspection of the data supported this assumption (see Figure 3.2b for an example). To assess this quantitatively, we fit models with either shared or varying slope and lapse parameters within and between sessions, using gradient descent to find the MAP parameter values under a non-informative prior. We then employed BMC to ask which of these models was best supported by our data, by comparing the log marginal likelihood that our data came from each model independently of particular parameter settings (see page 15). This approach to choosing an appropriate model originates with Jeffreys (1939), and incorporates an Occam's razor-type penalty for models with more parameters (Gull, 1988; Kass and Raftery, 1995; Mackay, 2004).

For each observer, we thus fit a model to the data from both sessions, and separate models to the data from each session. For each of these cases, we then fit models with individual ρ_j and ϵ_j parameters for each α, models with the values of ρ_j and ϵ_j restricted to have the same value for all α, and models with one parameter restricted while the other was allowed to vary. The centres μ_j always varied with α, as all data sets showed clear shifts. We then wanted to compute the marginal likelihood for each model, which in the absence of prior bias, is proportional to the probability that the data came from that model. For the

models fit separately for each session, the total log evidence is the sum of the log evidences obtained for each session. However, the marginal likelihood is rarely analytically tractable, and so we used the MAP parameters to compute a Laplace approximation for each model. The Laplace approximation results from taking the first three terms of a Taylor expansion about the MAP parameters (Mackay, 2004). In the equations below, Δ refers to the data, m to the model, θ to the vector of all parameters, θ^* to the MAP parameters, and k to the number of parameters in the model. The matrix A is the Hessian of the log posterior, i.e. the matrix of second partial derivatives of $\log P(\theta|\Delta, m)$ with respect to θ, evaluated at θ^*.

$$\log P(\Delta \,|\, m) = \log \int d\theta \, P(\Delta, \theta \,|\, m), \tag{3.19}$$

$$\log P(\Delta \,|\, m) \approx \log P(\Delta \,|\, \theta^*, m) + \log P(\theta^* \,|\, m) + \frac{k}{2} \log 2\pi - \frac{1}{2} \log |A|. \tag{3.20}$$

The values of the Laplace approximation for each of the four models are shown in Table 3.2. The highest evidence, corresponding to the 'best model', is highlighted in bold for each observer and each case. In accordance with our assumption, the best model had a single ρ parameter for all α, whether or not this slope was the same across sessions. A single lapse parameter was best for two observers and separate lapse parameters for the other two. As mentioned above, the lapse rate is incorporated in the model to account for decision noise, motor errors, and moments of inattention, and it seems reasonable that, whilst the internal uncertainty is the same for each α value, such random lapses might vary. In addition, the model with different parameters for the two sessions was always preferred, suggesting that observers' sensory noise changed between the two experimental sessions (see Section 3.3.6 below). The values of the parameters fit to the best model for each observer, in each session, are given in Table 3.3.

3.3.4 Optimal and Observed Shifts

Consideration of the various models thus showed that the behaviour of each observer in each session was best modeled by a family of curves of the same shape and slope, but with centres depending on α. We next asked whether the observed shifts in the curve centres were aligned with the predictions of the Bayesian decision theory analysis.

In at least one regard, observers were not optimal. The Bayesian prediction for the curve centre in the $\alpha = 0.5$ condition is always 0. However, for all observers, the curve centres for the $\alpha = 0.5$ condition were non-zero. Two observers showed a rightward bias

Summed Laplace approximation for individual session models				
	single ρ		separate ρ	
observer	single ϵ	separate ϵ	single ϵ	separate ϵ
1	1995	**1997**	1970	1975
2	1602	**1638**	1613	1619
3	**2351**	2338	2346	2335
4	**2061**	2033	2055	2041
Laplace approximation for pooled session model				
	single ρ		separate ρ	
observer	single ϵ	separate ϵ	single ϵ	separate ϵ
1	**1851**	1841	1827	1830
2	1433	**1487**	1446	1441
3	**2200**	2185	2197	2181
4	1944	**1948**	1941	1929

Table 3.2: ***Results of Bayesian model comparison.*** *Laplace approximation to marginal log likelihood for each of four models for each observer. Bold text shows the model with the highest log likelihood for each participant. It should be noted that each unit difference in log likelihood corresponds to an e-fold ratio of model probabilities.*

in both sessions, and two showed a leftward bias in both sessions (see Table 3.3), and we found no evidence that the bias was absent in the asymmetric penalty conditions. A similar directional bias has been reported widely in psychophysical studies (Green and Swets, 1966). In the analysis below we treat the directional bias as a constant constraint on observers' computations, and attempt to separate this form of non-optimality from the novel question of whether observers were able to integrate correctly the loss function with an estimate of internal uncertainty (see Section 3.4 for further discussion of this point). Thus we compute shifts as *relative to the biased centre for* $\alpha = 0.5$, yielding 'predicted relative shifts' for the other four α conditions;

$$\Delta\mu_j^* \;=\; \mu_j^* - \mu_{0.5} \;=\; \Phi_\sigma^{-1}(\alpha_j) - \Phi_\sigma^{-1}(0.5) \;=\; \Phi_\sigma^{-1}(\alpha_j). \tag{3.21}$$

The comparison between observed and predicted relative shifts for the two sessions is shown in Figure 3.4a and b. Note that both observed and predicted shifts derive from the same set of data, as the predicted shifts are based on an estimate of internal uncertainty derived from the slope of the psychometric curve. Thus estimation errors due to limited sampling may be correlated, and independent error bars for the two quantities cannot be drawn. Instead, we employed a bootstrap procedure to estimate the covariance of the errors in the two derived

		Session 1				Session 2			
obs.	α	μ (pix.)	ρ (1/pix.)	σ (pix.)	ϵ (prob.)	μ (pix.)	ρ (1/pix.)	σ (pix.)	ϵ (prob.)
1	0.7	-4.0	0.072	5.6	0.052	-3.8	0.086	4.6	0.031
	0.6	-3.5			0.025	-3.3			0.00002
	0.5	-1.9			0.13	-2.1			0.017
	0.4	2.5			0.16	0.55			0.031
	0.3	3.1			0.045	0.63			0.072
2	0.7	-7.2	0.049	8.2	0.019	-7.3	0.055	7.2	0.013
	0.6	-7.5			0.0019	-5.1			0.00012
	0.5	-2.1			0.11	-1.8			0.10
	0.4	2.0			0.021	3.1			0.10
	0.3	4.2			0.0098	5.8			0.016
3	0.7	-2.3	0.087	4.6	0.017	-0.92	0.13	3.1	0.011
	0.6	-0.79				-0.60			
	0.5	0.82				0.28			
	0.4	1.6				0.44			
	0.3	4.5				0.76			
4	0.7	-3.5	0.075	5.3	0.0069	-3.6	0.079	5.1	0.042
	0.6	-2.3				-2.1			
	0.5	0.45				1.3			
	0.4	4.3				4.2			
	0.3	4.8				5.0			

Table 3.3: ***Results of model fitting to experimental data.*** *Centre (μ), slope (ρ), and lapse (ϵ) parameters for each observer in each α condition and each session (values given to 2 significant figures). For each observer there is a separate μ for each α condition, representing the centre of the psychometric function in pixels. However, there is only a single ρ for all α conditions, representing the fact that the observer's internal uncertainty is the same regardless of the value of α. The Gaussian standard deviation in pixels corresponding to these values of ρ is given in the next column. Two observers have a single ϵ for all α, and two have separate ϵ for each α. These constraints on parameter values were determined via Bayesian model comparison (see Table 3.2)*

quantities, shown by the ellipses in Figure 3.4a and b. A linear fit to the observed shifts was computed by minimizing weighted squared error in the plane with respect to these covariances, and is also shown in Figure 3.4a and b.

There is a strong qualitative match between measured and predicted relative shifts. Observers shift their psychometric curves in the right direction, and by an amount that is proportional to the size of the penalty asymmetry. This is in contrast to the simple strategy verbally reported by all observers, which was to give the answer with the lower penalty whenever they were unsure. Interference from this cognitive strategy might help to explain the non-linearity of the observed shift plot in Figure 3.4a and b — the two leftward and two rightward shifts are more similar than in the quantitatively optimal scenario. However, if observers' behaviour was dictated by this simple strategy, we would expect the shifts to be of the same magnitude regardless of the size of the penalty asymmetry. In fact the shifts are significantly larger ($p < 0.01$ under a 2-tailed paired-samples t-test) for the greater cost asymmetries.

It has been observed previously that observers are reluctant to behave optimally when this entails an extreme bias in their responses (Green and Swets, 1966). Such effects could be characterised via a Bayesian prior that expresses observers' expectations about what they are expected to do – it seems strange to give the 'wrong' answer. However, it is difficult to determine such biases in a well constrained manner, and so we chose asymmetries that demanded a relatively small shift in the psychometric curve. In general, observers tended to *over* compensate for the penalty asymmetry (see Figure 3.4a,b), but for most sessions a *smaller* over-compensation was seen for larger α values, as would be predicted by such an effect. This could also have contributed to the non-linearity in the observed shifts plot.

3.3.5 Optimality of Achieved Score

Although all observers showed the predicted pattern of shifts, the quantitative match was not exact. This is perhaps unsurprising given the requirement to integrate implicit knowledge of internal uncertainty with high level cognitive instructions, but also raises the issue of which behavioural measure should be used to statistically test for optimality. In the present study, observers are asked to maximize their *point score*, not to work out what the optimal shift of the curve should be. It is possible that the function relating curve shift to total score is relatively flat in the region of the maximum score obtainable, such that there is little benefit in terms of expected utility from an exact quantitative match to the predicted curve shifts.

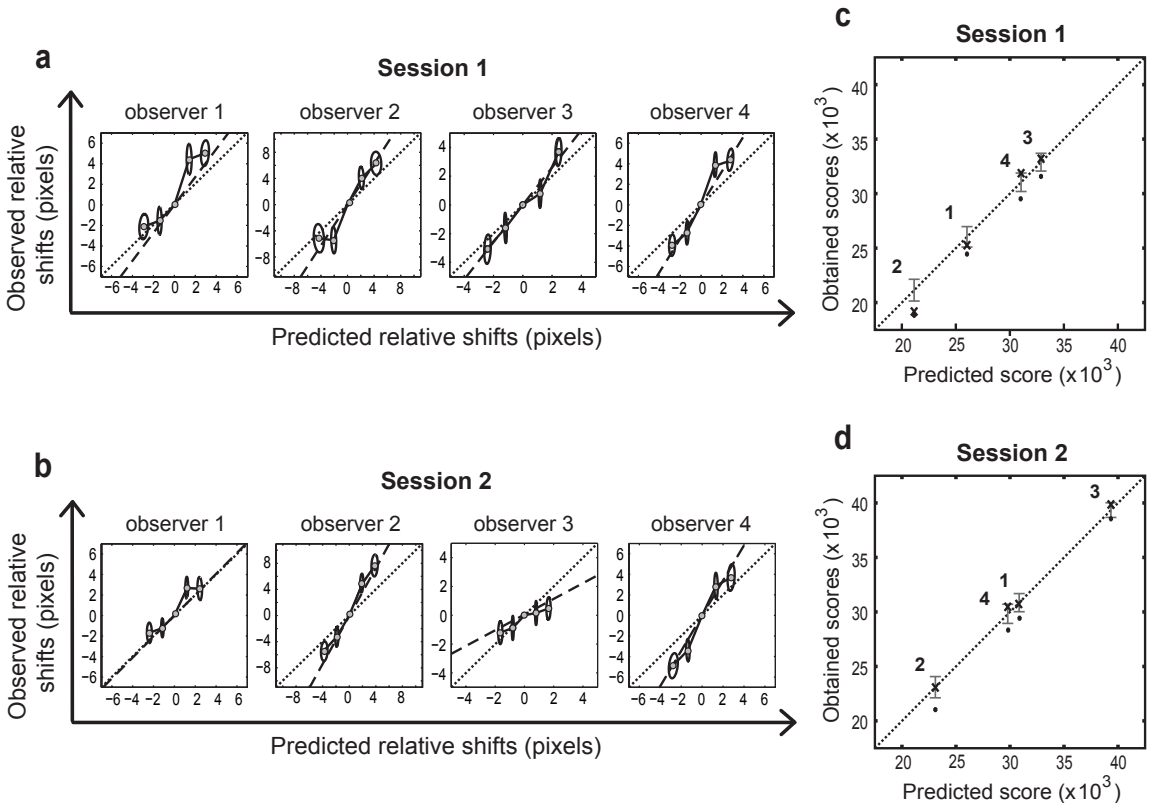

Figure 3.4: ***Comparison of predicted and observed behaviour. a, b,*** *Predicted and observed relative shifts in the centres of the psychometric curves, for the first (**a**) and second (**b**) sessions. If performance is quantitatively optimal (up to a constant bias), the data points (grey circles) should lie on the dotted identity line. The ellipses show the 2σ covariance expected due to sampling errors, and the dashed line is a linear fit to the data points, computed by minimizing the weighted squared error in the plane with respect to these covariances. All observers showed the predicted pattern of shifts, but were not quantitatively exact.* ***c, d,*** *The mean and variance of the score that would be obtained if each observer behaved optimally given the directional bias was calculated for the first (**c**) and second (**d**) sessions. Crosses plot predicted against observed score, with observers numbered as in **a** and **b**. The identity line again represents optimal performance (given the directional bias), and the vertical bars show one standard deviation from the mean. All points are within this range except for one observer in the first session. Filled circles show the mean score expected if observers failed to shift the centre of their psychometric curves from the biased centre of the curve for $\alpha = 0.5$, and all such points lie outside the predicted range.*

Our model of the psychometric curve was a composite function based on an underlying cumulative Gaussian, which gives the probability of answering 'right' for a particular value of the stimulus x and α (see Equations 3.15 & 3.17). We used this model to compute the mean and standard deviation of the score observers would have obtained had they shifted their psychometric curves by the optimal amount from the biased mean of the curve for the $\alpha = 0.5$ condition. We were then able to ask whether the true score was statistically distinguishable from this predicted value. The total score ('reward') for one session is obtained by summing the scores for each α value;

$$\mathcal{R}_{total} = \sum_j \mathcal{R}_j. \tag{3.22}$$

The score for each α value is given by the sum, over the different possible stimulus offsets x_i, of the number of trials on which the observer answers 'right' and 'left' correctly and incorrectly, multiplied by the appropriate reward or cost parameter. Using the same definitions of N_{ij} and n_{ij} as in Equation 3.15, this is

$$\mathcal{R}_j = \sum_{i \mid x_i<0} \left(\mathcal{R}_l \left(N_{ij} - n_{ij}\right) + \mathcal{C}_r \, n_{ij}\right) + \sum_{i \mid x_i>0} \left(\mathcal{R}_r \, n_{ij} + \mathcal{C}_l \left(N_{ij} - n_{ij}\right)\right). \tag{3.23}$$

Under our model, n_{ij} is binomially distributed with mean $N_{ij}\, p_{ij}$, where p_{ij} is given by the psychometric function (Equation 3.17). To obtain the expected score under the optimal strategy (constrained by the observed bias), we evaluated Equation 3.16 for each offset, using the measured value of ρ, but using the optimal relative value of μ_j obtained by adding the optimal relative shift to the observed bias in the symmetric condition (i.e. , $\mu_{0.5} + \Delta\mu_j^*$). Calling these optimal relative values p_{ij}^*, the expected score under the constrained optimal strategy is,

$$\langle \mathcal{R}_{total} \rangle \;\; = \;\; \sum_j \langle \mathcal{R}_j \rangle \,, \tag{3.24}$$

with

$$\langle \mathcal{R}_j \rangle = \sum_{i \mid x_i<0} N_{ij} \left(\mathcal{R}_l \left(1 - p_{ij}^*\right) + \mathcal{C}_r \, p_{ij}^*\right) + \sum_{i \mid x_i>0} N_{ij} \left(\mathcal{R}_r \, p_{ij}^* + \mathcal{C}_l \left(1 - p_{ij}^*\right)\right). \tag{3.25}$$

To compare the measured scores to this value, we must also calculate the variance in the score that is to be expected as decisions vary due to sensory noise. As the score in each

condition is independent, this variance is given by

$$Var(\mathcal{R}_{total}) = \sum_j \langle \mathcal{R}_j^2 \rangle - \langle \mathcal{R}_j \rangle^2 \,. \tag{3.26}$$

The second moment above can also be computed in closed form, using the expression for $\mathcal{R}_j$ in Equation 3.23, the binomial mean as before, and the binomial second moments:

$$\langle n_{ij} n_{i'j} \rangle = N_{ij}\, N_{i'j}\, p^*_{ij}\, p^*_{i'j} + \delta_{ii'}\, N_{ij}\, p^*_{ij}\,(1 - p^*_{ij}) \,, \tag{3.27}$$

where $\delta_{ii'}$ is the Kronecker delta. We obtain

$$\begin{aligned} \langle \mathcal{R}_j^2 \rangle \;\; &= \;\; \langle \mathcal{R}_j \rangle^2 \;+\; (\mathcal{C}_r - \mathcal{R}_l)^2 \sum_{i \mid x_i < 0} N_{ij}\, p^*_{ij}\,(1 - p^*_{ij}) \;+\; \ldots & (3.28) \\ &\qquad (\mathcal{R}_r - \mathcal{C}_l)^2 \sum_{i \mid x_i > 0} N_{ij}\, p^*_{ij}\,(1 - p^*_{ij}), & (3.29) \end{aligned}$$

and so the expected variance in score is

$$Var(\mathcal{R}_{total}) = \sum_j \left((\mathcal{C}_r - \mathcal{R}_l)^2 \sum_{i \mid x_i < 0} N_{ij}\, p^*_{ij}\,(1 - p^*_{ij}) \;+\; (\mathcal{R}_r - \mathcal{C}_l)^2 \sum_{i \mid x_i > 0} N_{ij}\, p^*_{ij}\,(1 - p^*_{ij}) \right). \tag{3.30}$$

The corresponding standard deviation is shown in Figure 3.4.

Scores for all observers fell within one standard deviation of the predicted score in the second session (Figure 3.4d), as did all but one in the first session (Figure 3.4c). The failure of this one observer to obtain a score in this range in the first session could be due to cognitive interference or motivational issues. To test the sensitivity of total score as a measure of optimality, we computed the mean score that would have been obtained had participants failed to shift their psychometric curves from the biased central point. All such points lay outside the predicted range, as shown in Figure 3.4c and d, though not by a dramatic amount. This analysis suggests that, whilst not quantitatively optimal, the observed shifts were sufficient to obtain a score well within the predicted range.

3.3.6 Changes in Performance

As described above, each observer participated in two experimental sessions on different days, in order to increase the amount of data collected. Inspection of the data suggested

that uncertainty differed between the two sessions, which provided a further test of the hypothesis that observers' behaviour is driven by internal beliefs that accurately reflect their sensory noise – if that sensory noise were to change, their beliefs, and thus their behaviour under the asymmetric loss function, should change concomitantly.

We first established that the level of observers' sensory noise did, in fact, appear to change, as would be reflected by a change in the slope of the psychometric curve. We used BMC to compare models with the same slope in the two sessions to models in which the slope could differ. Table 3.2 shows that the model with different slopes in the two sessions was overwhelmingly preferred in all cases, although, *within* each session, the model with the same slope for different loss functions was still the most probable. Thus, despite the apparent change in sensory noise, the basic prediction that the shape of the psychometric curve is unaffected by the loss function is confirmed.

In general, the slope of the psychometric curve was steeper in the second session, and observers' behaviour altered in accordance with the predictions of the Bayesian analysis. This can be seen as a trend towards smaller shifts in the second session (compare Figure 3.4a and b), and towards higher scores (compare Figure 3.4c and d). In particular, the three observers whose scores were in the predicted range in both sessions maintained this performance in the face of a clear change in apparent sensory noise. Furthermore, had the three observers who showed substantial changes in accuracy between the two sessions retained the same relative shifts in the second session as in the first, their expected scores would have fallen outside the optimal ranges shown. This suggests that observers were indeed adopting an efficient strategy, taking into account both the level of uncertainty and the external loss function.

As discussed in Section 3.2, we did not attempt to distinguish between stimulus-centred *sensory* noise, and any stimulus-centred *decision* noise not modeled by the stimulus-independent lapse-rate parameter. However, we assumed that the majority of this stimulus-centred variation was in fact due to sensory noise, and treated it as such. If this assumption is incorrect, the measured slope may incorporate stimulus-centred 'decision' noise associated with integrating the loss function with the true uncertainty, and thus an increase in slope might reflect an improvement in task performance rather than a change in internal uncertainty. However, inspecting Figure 3.4a and b shows that for only one observer did the slope of the linear fit to performance (the dashed line) get *closer* to the identity line in the second session, supporting the assertion that the internal uncertainty was changing, rather than the ability to perform the task. Indeed, the observer whose fit improved was the same observer who obtained a score outside the predicted range in the first session, and it seems possible that she *did* change her strategy.

3.3.7 Controls for Feedback

In order to use an ideal observer analysis to argue, with the BCH, that observers represent and compute with the relevant uncertainty, it is crucial to rule out alternative strategies for obtaining optimal performance that do not require such knowledge. In the present task, it is possible that trial-by-trial feedback, had we provided it, would have allowed observers to incrementally adjust a internal threshold (perhaps in proportion to the size of the penalty) until their payoff was optimized. This could have led to psychometric curves that looked very much like those we predict from the analysis above. Indeed, classic Psychophysical studies have used a similar paradigm with trial-by-trial feedback to demonstrate this kind of 'optimal' criterion selection (Green and Swets, 1966).

Previous studies of uncertainty that have used trial-by-trial feedback have dealt with a similar potential confound by looking for evidence of incremental threshold adjustment in the data (Trommershauser et al., 2003a, 2005). The alternative strategy, that of withholding feeback, was adopted by (Kording and Wolpert, 2004) in a sensorimotor task, although without any asymmetry in costs. Here, we chose to provide only occasional (every 15 trials) cumulative feedback during the testing blocks (see page 83). This provided motivation, but prevented observers from behaving optimally via trial-by-trial threshold adjustment.

However, even such scarce feedback does provide some limited information about sensory noise, so we performed control analyses to confirm that the magnitude of the cumulative feedback had no measurable effect on behaviour. First, we fit a psychometric curve to all data which followed 'good' feedback (i.e. a cumulative score for the preceding 15 trials which fell above the 75th percentile), and to all data which followed 'bad' feedback (i.e. a cumulative score for the preceding 15 trials which fell below the 25th percentile). There were no feedback-related trends in the data (data not shown). To test for effects that might have been lost in averaging in this technique, we then examined whether 'good' feedback reinforced the direction of any changes in threshold, and whether 'bad' feedback reversed the direction of any such changes. If observers were modifying their behaviour in this way, we would expect a positive correlation in threshold changes following 'good' feedback, and a negative correlation following 'bad' feedback. However, we observed only a slight negative correlation in both cases (data not shown).

3.4 Discussion

Uncertainty inescapably affects almost all neural processing, arising due to variability in the external world, due to the under-constrained nature of many problems of perceptual inference and motor planning, from variability in motor execution, and due to noise in sensory processing (see Section 1.3.2). A long-standing and fundamental question in neuroscience is whether, and if so how, the brain takes account of this uncertainty in the course of perception, decision making, action and learning. The BCH states that not only do observers take into account uncertainty, they do so according to the optimal prescriptions of Bayesian decision theory, and further, that these probabilistic computations are implemented in the brain (see Knill and Pouget, 2004; Doya et al., 2007).

Evidence that people follow Bayesian prescriptions for perception and action is an important source of support for the BCH, but to maximise its strength it is necessary to show that optimal performance cannot be achieved by other means, and is flexible and ubiquitous. Here, we attempt to achieve this in the context of a simple visual offset judgement, showing that observers possess an internal model of sensory uncertainty, and that they use this model to guide their decisions. Crucially, observers' decisions are sensitive to uncertainty even when they do not receive significant feedback about their accuracy or score. This indicates that the uncertainty-sensitive decision strategy is not learnt by adaptive-threshold adjustment during the experiment, but is instead based on a pre-existing, implicit model of current internal uncertainty, that is presumably available at all times. In addition, observers' decisions, and thus their models of internal uncertainty, tracked the changes in uncertainty which were associated with varying levels of sensory noise in the two experimental sessions[2]. Taken together, these observations suggest that the processing of uncertainty is a fundamental aspect of sensory computation. Furthermore, in our experiment observers must combine knowledge of this uncertainty, rather than simply a modal estimate of the stimulus, with an externally imposed loss function to perform well. Our results therefore also indicate that information about early sensory uncertainty, at least in the form of a two-alternative likelihood ratio between the models for left and right displacement, is propagated across multiple cortical layers to decision-making regions of the brain.

The sensory processing that supports a Vernier judgement is likely to occur early in the visual pathway, perhaps principally in relatively well-understood striate cortex (V1). Combined with the existing evidence for optimality in crossmodal, motor, and visual cue combination experiments, our results support the claim that uncertainty is represented

[2]As discussed below with regard to distinguishing sensory and decision noise, future work could directly manipulate the level of sensory uncertainty, also providing a more rigorous version of observers' ability to maintain optimal strategies as their level of uncertainty changes.

throughout the brain, even for simple, low-level visual quantities. Furthermore, such a task is a strong candidate for future integration with physiological data via neural coding models (see Section 2.2.4). Such 'triangulation' of methodologies is best approached in a very well constrained domain, where the task can be performed by primates as well as humans, and where the cortical substrate is reasonably well defined (see Section 2.2.4). Our demonstration of a Bayes-optimal strategy in a simple visual domain has all these properties. Another key property of our task is that the set of visual stimuli was fixed, so that stimulus-related uncertainty arose almost exclusively from visual processing, corresponding to 'internal noise' in psychophysical experiments (Green and Swets, 1966). This is in contrast to many previous studies of sensory uncertainty, where variability was driven by external manipulations, such as the random placement of dots or the addition of corrupting noise (but see Stocker and Simoncelli, 2006a). Using a fixed stimulus set thus strengthens the conclusion that the mechanisms exposed are fundamental to sensory processing, rather than being limited to strategies for dealing with uncertainty in the external world.

There are many technical issues with Bayesian optimality experiments, which make deviations from optimality hard to interpret – it can be hard to discriminate between suboptimal performance, a failure to identify the observer's ecological prior, and non-Bayesian strategies. For example, with monetary loss functions observers often demonstrate over compensation, a reluctance to make extreme shifts, and failure to keep track of the current loss function (Green and Swets, 1966; Landy et al., 2007). This relates to the body of evidence that human economic reasoning is characterised by heuristics and biases (Kahneman and Tversky, 1979; Glimcher and Rustichini, 2004), whose 'rationality' is still under debate. These might in fact constitute contributions to an ecological prior that people struggle to leave behind in an experimental setting (see page 34), but we wanted to avoid this complication. We therefore used values of α that demanded relatively small shifts, and used practice trials and a blocked design with regular reminders of the current cost values. Recent work using a similar loss function approach in an unspeeded visual orientation estimation task (Landy et al., 2007) found evidence for optimality, but also for a variety of suboptimal strategies that might reflect heuristics and biases. It may be that unspeeded adjustment tasks are more vulnerable to such influences, compared to our simple, speeded forced-choice categorization task. However, as discussed above we do see some evidence for over-compensation, and for relatively smaller over-compensation for larger cost asymmetries. Using more extreme loss functions would have risked exacerbating these effects, but might also have resulted in an optimal score range that was more sensitive to quantitative curve shifts (see page 98).

The optimal Bayesian decision maker for this task would have set the mean of their curve in the equal penalty condition to 0 – maximising expected utility clearly requires

observers to answer 'left' when the offset is left of centre and 'right' when it is right of centre. However, we found a biased mean for the equal penalty condition, and defined optimal performance in terms of shifting psychometric functions by an optimal amount relative to this biased centre. This embodies an assumption that the bias is a fundamental feature of the observer's perception – supported by behavioural evidence (see Green and Swets, 1966) and by the consistency of the bias across sessions for each observer. In general, this highlights the difficulty of pinning down simple settings in which perfect optimality, as defined by a Bayesian decision theory analysis with unbiased priors and linear utility function, can be demonstrated.

In our analysis we did not attempt to distinguish between stimulus-centred *sensory* noise, and any stimulus-centred *decision* noise not modeled by the lapse-rate parameter. However, we assumed that the majority of this stimulus-centred variation was due to sensory noise. This assumption was supported by examining the results across sessions – the ability of observers to choose optimally in the face of asymmetric costs did not change as their ability on the task, measured by the slope of the psychometric function, did (see Figure 3.4a and b). Using external manipulations to produce randomly intermixed uncertainty levels on each trial would allow us to separate more explicitly any stimulus-centred decision noise from uncertainty due to sensory processing. However, Landy et al. (2007) found greater suboptimality when levels of uncertainty were randomly intermixed rather than blocked by session as in our experiment, and it is not clear *why* this should be the case. It could reveal limits on the ability to perform online Bayesian processing, or alternatively arise from cognitive 'interference'. In the present study we were not interested in trying to delineate these factors, and so used a blocked design which retains the property of having stimulus uncertainty arising internally.

In the present study we aimed to demonstrate minimal conditions for Bayes-optimal behaviour under uncertainty. We showed that observers approach the quantitatively optimal strategy given a directional bias, and score within the predicted range, in a simple unimodal visual task that requires them to integrate a model of their internal uncertainty with an external loss function. The assumptions of the model, and the predictions that arise from them, were tested, and we took care to rule out alternative strategies for achieving the observed behaviour. Our results therefore support the assertion that the processing of uncertainty is a fundamental aspect of sensory computation, and can be used to inform subsequent decision-making processes. In Chapter 4 we report the results of a psychophysics paradigm that (a) asks whether optimal performance is also seen when the stimulus to be categorised has a complex, multidimensional structure, and (b) is amenable to fMRI analysis, enabling us to investigate the anatomy of Bayesian decision making in Chapter 5.

4

Bayesian Decision Making with Complex Stimuli and Labile Value

The class of perceptual judgements in which observers take their uncertainty into account, and do so in a Bayes-optimal way, is unknown. In Chapter 3 we found that even for a very simple, unimodal visual judgement, people could score near optimally given changes in an externally imposed loss function. This provides indirect evidence for the BCH, implying that uncertainty about stimuli represented in early visual areas is available online, and can be combined with information about reward value in decision-making regions of the brain. In this Chapter, we ask whether observers can behave optimally when categorising complex, multidimensional face-house stimuli, and when the loss function changes more rapidly. Observers again exhibited the qualitatively optimal strategy, but performance was quantitatively suboptimal. We discuss possible interpretations of this suboptimality, raising important issues about approximate Bayesian inference. As well as further delineating the limits of Bayesian optimality, this study was designed to provide ideal parameters for an fMRI investigation of the anatomical basis of the integration of sensory uncertainty with externally imposed loss functions, reported in Chapter 5.

4.1 Introduction

In Chapter 3 we showed that for simple visual stimuli, uncertainty could be integrated with externally imposed loss functions, in a way that approached Bayesian optimality. This paradigm used a unimodal Vernier offset discrimination with a flat prior, implying that uncertainty about even very simple features is represented and can be transmitted to decision-making areas of the brain. The avoidance of trial-by-trial feedback, and the preservation of optimality in the face of periodic changes in uncertainty, strengthened the conclusion that this kind of neural representation is flexibly and ubiquitously available.

Evidence for Bayes-optimal perception has tended to focus on visual features thought to be represented at intermediate stages of visual processing, such as size, depth, and motion direction – perhaps because they are most amenable to the cue combination and biased ecological prior paradigms typically used to probe uncertainty (see Knill and Pouget, 2004, and Section 2.1). If the representation of posterior belief distributions over stimulus features is truly ubiquitous, all visual areas should carry uncertainty information, not just those that analyse simple features. In Chapter 3 we asked about the representation of uncertainty in a single posterior over a simple visual quantity, and here we ask the same question for high dimensional, semantically rich stimuli. Bayesian optimality analysis is far more readily performed when the noise model – the form of the likelihood distribution – is simple and analytically tractable. High dimensional stimuli can be hard to characterise along a single axis, and we therefore developed a stimulus dimension that consisted of face/house mixtures, running from 100% face to 100% house (see Section 4.2.2 for details). This subsumes the multiple dimensions along which such an object can be characterised into a '% face' axis. Observers were asked to categorise stimuli as face vs. house, again under an asymmetric loss function, yielding a psychometric curve that should be shifted in the direction of the lower penalty, and by an amount determined by a Bayesian decision theory analysis (see Equation 3.3.1).

In this study[1] we also changed the loss function every two trials, rather than every block, to investigate whether observers can flexibly shift their decision boundary with rapid changes in value, even without the opportunity to learn from trial-by-trial feedback. In order to integrate uncertainty with external monetary loss in this task, the potential losses must be represented and combined with the stimulus difficulty and level of uncertainty to determine the answer on each trial. Failure to do so when the loss function changes rapidly could result from limitations in the speed with which the brain can adjust its valuation machinery, from

[1] This work was the result of a collaboration between myself and Dr Maneesh Sahani at the Gatsby Unit, Mr Stephen Fleming, Prof. Ray Dolan, and Prof. Chris Frith at the Wellcome Trust Centre for Neuroimaging, and Dr Oliver Hulme at the Institute for Ophthalmology. See page 13 for details of contributions.

restrictions on learning, or from the interference of simple cognitive strategies imposed by observers to deal with an apparently complex task. As discussed in Section 3.4, it can be hard to distinguish 'interference' from the effects of an experimentally-inappropriate but ecologically sound prior, and the removal of feedback to strengthen evidence for the BCH could obscure other, still-Bayesian, learning-dependent strategies. (see Section 4.4 and page 144 for further discussion).

The changes to the paradigm that enabled us to ask about optimality with regard to complex stimuli and labile value were also motivated by the desire to investigate the neural basis of perceptual decision-making with fMRI (see Section 2.3 of the literature review, and Chapter 5). We wanted to ask where in the transformation from epithelial activity to motoric report the impact of externally determined value would be observed, and how far uncertainty due to perceptual processing overlaps neuroanatomically with the representation of probabilistic contingencies in value-based decision making. In the decision-making terminology laid out in Figure 2.5, we wanted to investigate the neural correlates of manipulating a deterministic loss function; $U(o_j)$, asking specifically whether it affects areas correlated with the representation of sensory uncertainty in the decision-outcome mapping; $p(o_j \mid d_i)$. In order to do so, we needed a stimulus axis whose two extremes were represented by anatomically distinguishable regions – any topographical separation between representations of left and right Vernier offsets that does exist is unlikely to be resolvable with fMRI – and so a face-house dimension was a natural choice (see Heekeren et al., 2004; Kanwisher et al., 1997; Epstein and Kanwisher, 1998). In addition, slower timings and frequent changes in the loss function increased the efficiency of our experimental design, making analysis of the imaging data more sensitive to the hypotheses we wanted to test.

4.2 Materials and Methods

4.2.1 Participants

Nineteen right-handed participants participated in the psychophysics (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. The study was approved by the Institute of Neurology (University College London) Research Ethics Committee.

4.2.2 Stimuli

Face and house stimuli are popular in fMRI studies due to the anatomically distinguishable visual regions that correlate strongly with their processing (Kanwisher et al., 1997; Epstein and Kanwisher, 1998). However, they aren't natural candidates for two ends of a stimulus continuum, which we require in order to plot psychometric functions for observers' categorisation performance. When considered on the level of visual features the mapping from face to house embodies a multi-dimensional, non-linear function, so we circumvented this problem by using mixtures of the phase components of fast Fourier transforms of the face and house images to produce a single '% face' axis. This is similar to the approach used by Heekeren et al. (2004), though they combined phase matrices with various degrees of random noise to produce a noise continuum, rather than mixing phase matrices to produce a continuum between face and house identity.

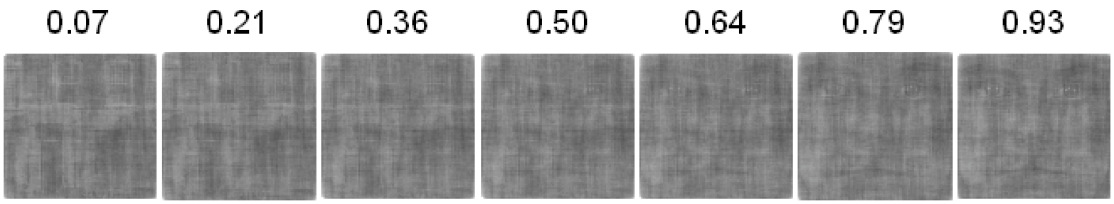

Figure 4.1: ***Face-house stimulus continuum.*** *Illustrates an exemplar Fourier phase transition from a single house image to a single face image. Numbers above each image indicate the proportion of "face" phase in the stimulus. In the experiment, stimuli were created from a face and house randomly drawn from the total image set on each trial. Seven phase proportions are shown here; in the experiment, there were fifteen equally spaced mixtures between pure face and pure house stimuli.*

10 neutral faces (5 male, 5 female; taken from the KDEF face set of Lundqvist et al. (1998)) and 10 houses (photographed by Stephen Fleming) were cropped to be of equal size and converted to grayscale. Fast Fourier transforms 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. The resulting phase matrix was then recombined with the average magnitude matrix of the whole stimulus set using an inverse Fourier transform. This procedure produced a smooth transition between noisy faces and houses, and ensured that

each stimulus had identical frequency power spectra. Finally each image was normalised to have equal luminance relative to the screen background and constant root-mean squared contrast. Figure 5.4 shows an exemplar Fourier phase transition from a single house image to a face image. In the experiment, 15 stimuli were presented, equally spaced from 100% face to 100% house, allowing us to plot a full psychometric function.

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 Dell monitor running at a refresh rate of 60 Hz, situated in a dimly lit room. All images subtended 4° of visual angle at a viewing distance of 60cm.

4.2.3 Procedure

Observers were not informed of the image continuum, and were simply asked to categorise each briefly presented stimulus as either a noisy face or house. They found this task natural and were unaware of any blend between the two image categories when debriefed. Face and house responses were made using left and right-hand key presses respectively. Before introducing a reward component to the task, each observer completed 540 trials of simple face/house discrimination. This served two purposes – first, to provide training and allow task performance to saturate. Second, unlike the Vernier task where a 0 offset defined the categorical boundary between 'left' and 'right', it is not clear which phase mixture should define the categorical boundary between face and house. We therefore fitted a psychometric function to the training data (see below), and used the point of subjective equality (PSE) for each observer to define face and house categories for the main task where reward was introduced – thus optimal shifts are automatically relative to the mean of the curve in the neutral condition as expressed in Equation 3.21. The average PSE across the group was $53.9 \pm 9.15\%$ face phase.

The main task involved further face/house discrimination under a potentially asymmetric loss function. For the Vernier task reported in Chapter 3 we awarded points for a correct answer and subtracted points for incorrect answers, with total points at the end being converted into a monetary bonus. The number of points subtracted for incorrect answers was varied, so that in four of the five conditions the loss due to answering "left" incorrectly was different from that due to answering "right" incorrectly. In this study, rather than losses limiting the amount of points observers could gain from correct answers, observers started each block with an endowment of £10 and could only lose money for incorrect answers. This manipulation was chosen in part to avoid the well-known bias in which people weight losses more highly than gains – i.e. the slope of the utility function that maps external

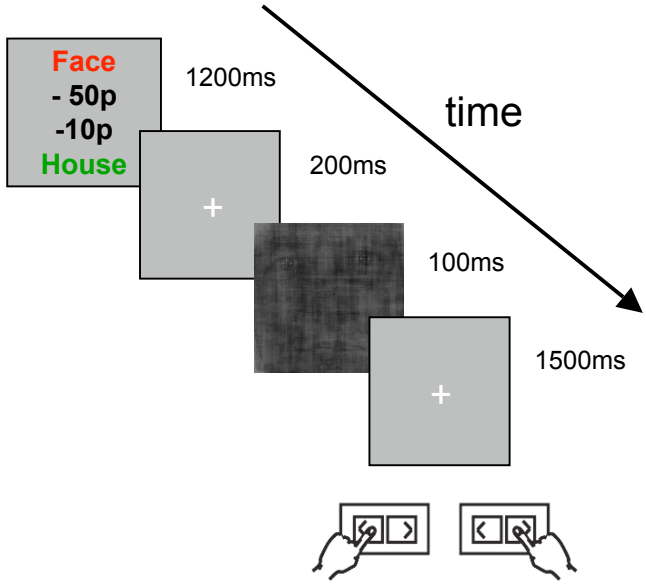

Figure 4.2: ***Experimental procedure.*** *After training, participants completed 1260 trials as schematised here. On each trial, a cost signal screen informing observers of the potential losses for an incorrect face or house categorisation was displayed for 1200ms. This was followed by a fixation cross displayed for 200ms, then a stimulus randomly drawn from the face-house phase continuum was displayed for 100ms. Observers made a face vs. house categorisation response, pressing a button with their left vs. right hand to indicate their decision. Feedback was given only every 15 trials.*

reward to internal utility is steeper for negative than for positive rewards as described in Prospect Theory (PT) (Kahneman and Tversky, 1979) – which could not be measured in the Vernier task. This also ameliorated potential problems with discriminating anatomical correlates of positive and negative cost in the fMRI study (see Rangel et al., 2008).

In the present study we used only three value conditions rather than the five used in Chapter 3, in order to garner enough data for each condition given limited scanner time. In the neutral value (NV) condition observers lost 30p for either error, in the face value (FV) condition they lost only 10p for an incorrect "face" answer compared to 50p for an incorrect "house" answer, and vice versa in the house value (HV) condition. Participants were introduced to the task through a series of practice blocks with decreasing stimulus presentation time and stimulus onset asynchrony (SOA) until integrating the cost and stimulus information became natural and performance saturated. They then completed nine experimental blocks of 140 trials each, lasting around 3 hours including breaks as desired.

In these main experimental blocks, stimulus timings were as illustrated in Figure 4.2 – the penalty instruction screen was presented for 1200ms, followed by a 200ms fixation cross, and then a brief 100ms presentation of a face-house stimulus. Observers then had 1500ms to issue their button-press response before the next trial began. As in the Vernier task, feedback was given only periodically to avoid incremental learning of stimulus parameters that could proceed on the basis of trial-by-trial feedback. A screen displaying the current total remaining from the current block's endowment was presented every 15 trials. The value condition changed every two trials (maximising fMRI efficiency – see Chapter 4).

4.3 Results

Exactly the same intuitive strategy applies to this task as to the Vernier discrimination – when the cost for answering "face" incorrectly is less, it makes sense to answer "face" more often when unsure, which should result in a qualitative shift of the psychometric curve towards face answers (and vice versa for the house condition; see Figure 3.2a). As in Chapter 3, we assessed whether observers followed this strategy by fitting psychometric functions to their data, and then using BMC to test whether the best model of the data was one with different means but the same slope for the different value conditions – a prerequisite for then computing quantitatively optimal shifts (see Equation 3.21) and scores (see Equation 3.25). We used gradient descent to fit a set of three composite binomial error functions to each observer's data (see Equations 3.15 and 3.17), for each combination of shared and separate mean (μ), slope (ρ), and error rate (ϵ) parameters. Looking at the raw data (see Figure 4.3) suggested that, unlike in the Vernier task, observers didn't always change the mean of their psychometric function with changes in value, so we tested the number of means as well as the number of slopes and error rates.

Figure 4.3 shows the full, nine parameter model fit to each observer's raw data, and it appears that for some observers the curves did not shift very much, and that the slope was not always the same across conditions. To quantify this, we computed the Laplace approximation to the log marginal likelihood for each of the eight models, to determine which was the best explanation of the data (see Equations 3.19 and 3.20). The results are reported in Table 4.1, with the highest value of the Laplace approximation highlighted in bold. The parameters for the best model are given in Table 4.2 for each observer.

As shown in Table 4.1, the model that provided the best fit to each observer's data was highly variable. The summed Laplace approximation given in the final row shows that for the group, the best model was one with separate means and slopes for each value condition, violating the shared-slope assumption of the ideal observer analysis reported in Chapter 3.

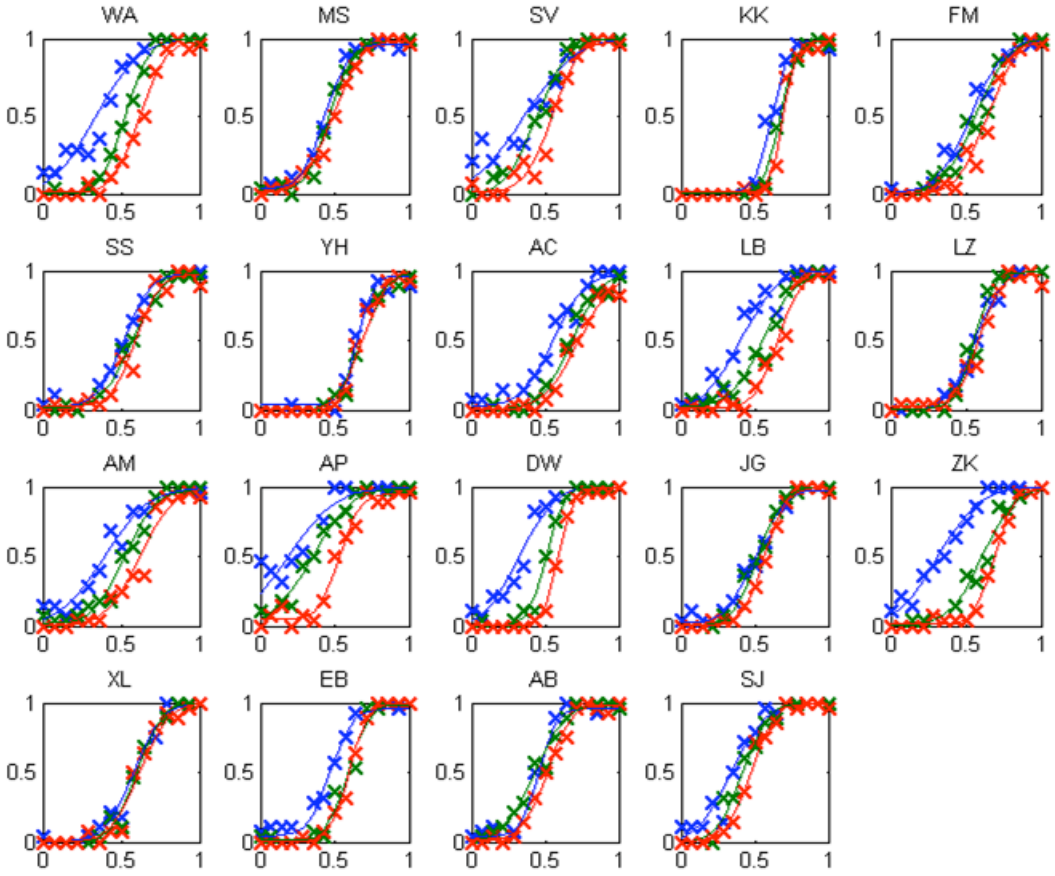

Figure 4.3: ***Behavioural data for individual observers.*** *Individual choice probability data from the 1260 trials of the out-of-scanner psychophysics. On each figure, the abscissa represents the proportion of face phase in the image, and the ordinate the proportion of "face" responses to that stimulus in the different value conditions. Blue = face value (FV); green = neutral value (NV); red = house value (HV). The curves plotted through the data represented composite binomial error functions (see Equations 3.15 and 3.17 in Chapter 3), fit using gradient descent. These functions were fit with each combination of shared and separate mean (μ), slope (ρ), and error rate (ϵ) parameters for each of the three value conditions, but displayed here is the full, nine parameter model for each observer.*

Laplace approximation to log marginal likelihood									
		single μ				separate μ			
		single ρ		separate ρ		single ρ		separate ρ	
	obs.	sing. ϵ	sep. ϵ	sing. ϵ	sep. ϵ	sing. ϵ	sep. ϵ	sing. ϵ	sep. ϵ
1	WA	-151.0	-139.6	-138.3	-139.2	**-95.35**	-98.83	-87.75	-90.45
2	MS	**-79.30**	-83.55	-80.11	-83.73	-87.10	-91.60	-88.15	-92.25
3	SV	-126.4	-118.1	-119.6	-118.9	-116.3	-112.7	-105.8	**-105.3**
4	KK	-75.61	-81.18	**-74.36**	-79.01	-75.90	-81.36	-76.54	-81.93
5	FM	-91.14	-94.89	-92.93	-96.67	**-89.08**	-92.24	-90.56	-93.24
6	SS	**-86.88**	-92.05	-88.61	-93.23	-91.91	-97.19	-93.45	-98.12
7	YH	**-75.22**	-77.78	-76.66	-79.00	-86.34	-89.45	-88.00	-88.48
8	AC	-115.0	-115.8	-115.4	-115.5	**-96.65**	-100.4	-98.14	-101.1
9	LB	-144.5	-148.5	-144.8	-148.8	-101.6	-106.0	**-100.8**	-103.1
10	LZ	**-76.44**	-80.34	-76.66	-80.45	-86.81	-90.66	-86.93	-90.50
11	AM	-127.9	-128.6	-126.6	-129.5	-102.3	-105.4	**-101.4**	-105.5
12	AP	-175.0	-175.1	-174.1	-174.9	-113.2	-117.6	-112.8	**-109.9**
13	DW	-154.9	-145.6	-136.7	-140.9	-87.63	-87.70	**-75.93**	-79.65
14	JG	-77.63	-77.69	**-77.45**	-78.69	-84.35	-84.87	-83.21	-85.07
15	ZK	-200.3	-180.2	-178.8	-180.1	-96.14	-99.88	**-89.95**	-95.04
16	XL	**-72.56**	-75.04	-74.41	-75.56	-82.38	-85.13	-84.28	-85.69
17	EB	-106.7	-100.4	-104.5	-101.8	-91.97	-90.93	**-90.34**	-92.17
18	AB	-89.97	-94.61	-89.69	-92.51	-93.94	-98.53	**-93.93**	-96.74
19	SJ	-92.85	-92.34	-91.96	-93.22	-87.40	-89.25	**-86.09**	-86.28
totals		-2119	-2101	-2062	-2102	-1766	-1820	**-1734**	-1781

Table 4.1: ***Results of Bayesian model comparison.*** *Laplace approximation for each of eight models fitted to each observer's data (4 s.f.), with the final row showing the summed approximation for each model, across all observers. Bold text shows the model with the highest value. The initials of participants who took part in the scanning study are in red.*

We augmented this Bayesian analysis with classical between-subject statistics, using $t-$ and $F-$tests to compare the means and slopes of the full, nine-parameter model in different value conditions. Figure 4.4 shows the average PSE (mean) across the group for each value condition (coloured bars; blue = FV, green = NV, red = HV). The trend was for PSE to decrease relative to neutral for the face value condition, and increase relative to neutral for the house value condition – i.e. to shift in the direction of the lower cost stimulus as predicted. To test whether this shift was significant we discretised the choice probability axis into 100 equal steps, and took the inverse of the three independently fit psychometric functions at each point, giving three stimulus value vectors; **HV**, **NV**, and **FV**. We then took the mean of the vector subtractions $\mathbf{FV} - \mathbf{NV}$ and $\mathbf{HV} - \mathbf{NV}$, giving two scalars that represent the average shift relative to neutral in the face value and house value conditions for

observer		α	μ (pixels)	ρ (1/pixels)	σ (pixels)	ϵ (probability)
1	WA	0.7	0.34	1.81	0.297	0.0067
		0.5	0.52	3.36	0.218	
		0.3	0.62	2.61	0.247	
2	MS	0.7	0.47	2.92	0.234	0.043
		0.5				
		0.3				
3	SV	0.7	0.36	1.44	0.333	0
		0.5	0.44	2.38	0.259	0
		0.3	0.54	2.81	0.238	0.038
4	KK	0.7	0.67	3.19	0.223	0.017
		0.5		4.57	0.187	
		0.3		5.26	0.174	
5	FM	0.7	0.53	2.36	0.260	0.0051
		0.5	0.56			
		0.3	0.64			
6	SS	0.7	0.55	2.91	0.234	0.048
		0.5				
		0.3				
7	YH	0.7	0.68	3.02	0.230	0.013
		0.5				
		0.3				
8	AC	0.7	0.55	2.27	0.265	0.025
		0.5	0.68			
		0.3	0.73			
9	LB	0.7	0.4	2.13	0.273	0.029
		0.5	0.55	2.16	0.271	
		0.3	0.66	3.03	0.229	
10	LZ	0.7	0.57	3.35	0.218	0.032
		0.5				
		0.3				

Table 4.2: ***Psychometric curve parameters for the best model for each observer.*** *The table lists centre (μ), slope (ρ), and lapse (ϵ) parameters for each observer in each α condition (values given to 2 significant figures). The BMC results reported in Table 4.1 were used to select which parameters were shared between alpha conditions, and which were separately determined. As can be seen, there was a large amount of variability across observers – crucially, few produced data best accounted for by a model with a single slope – an assumption necessary to the Bayesian optimality analysis. The standard deviation in pixels corresponding to the slope parameter ρ is given in the next column, and was derived according to Equation 3.18. The initials of participants who also took part in the scanning study are in red, those of participants for whom the best model had a single slope and separate means are underlined.* *Table continued on next page.*

observer		α	μ (pixels)	ρ (1/pixels)	σ (pixels)	ϵ (prob.)
11	AM	0.7	0.39	1.85	0.293	0.035
		0.5	0.52	2.53	0.251	
		0.3	0.61	2.57	0.249	
12	AP	0.7	0.19	1.51	0.325	0
		0.5	0.35	1.94	0.286	0.03
		0.3	0.53	3.69	0.208	0.11
13	DW	0.7	0.33	2.19	0.270	0.011
		0.5	0.51	4.23	0.194	
		0.3	0.59	4.67	0.185	
14	JG	0.7	0.53	2.32	0.262	0.017
		0.5		3.03	0.229	
		0.3		3.21	0.223	
15	ZK	0.7	0.32	1.90	0.289	0.023
		0.5	0.61	2.37	0.259	
		0.3	0.69	3.72	0.207	
16	XL	0.7	0.59	2.9	0.234	0.008
		0.5				
		0.3				
17	EB	0.7	0.45	2.3	0.263	0.022
		0.5	0.59	3.51	0.213	
		0.3	0.6	4.18	0.195	
18	AB	0.7	0.47	3.90	0.202	0.055
		0.5		2.60	0.247	
		0.3		3.01	0.230	
19	SJ	0.7	0.34	2.17	0.271	0.0049
		0.5	0.42	2.88	0.235	
		0.3	0.47	3.06	0.228	

Table 4.3: ***Psychometric curve parameters for the best model for each observer.*** *Continues Table 4.2.*

each observer. Paired t-tests revealed that, across the group, the curves shifted significantly in the right direction for both the FV condition ($t(18) = 5.95$, $p < 0.0001$) and HV condition ($t(18) = 4.98$, $p < 0.001$). Figure 4.4 also shows the average slope for value condition (black points). Across the group, there was a trend towards a lower slope parameter in the HV condition than NV (significant by a paired t-test; $t(18) = 2.41$, $p < 0.05$), corresponding to greater uncertainty, and a higher slope parameter in the FV condition (not significant by a paired-samples t-test; $t(18) = 1.24$, $p = 0.23$). Mean RTs did not differ between value conditions ($F(2, 36) = 0.70$, $p > 0.4$), but significantly correlated with difficulty[2] (mean $r = 0.79 \pm 0.092$, $n = 19$).

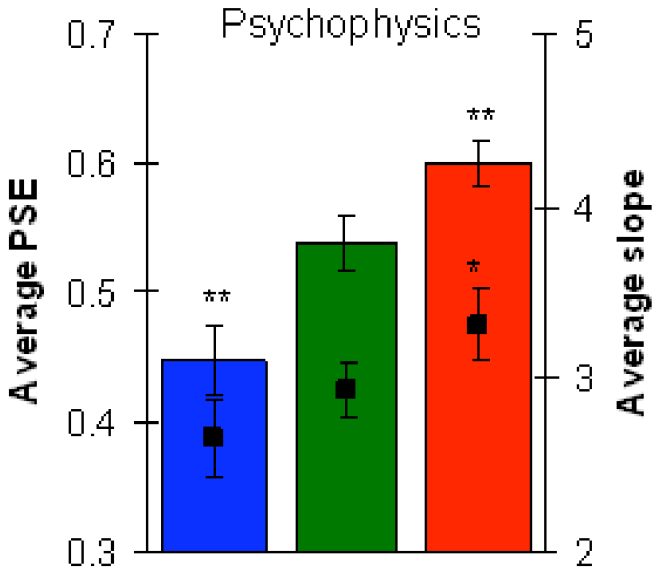

Figure 4.4: ***Average parameters of the psychometric function.*** *Bars represent the average PSE, and black points the average slope, in the FV (blue), NV (green) and HV (red) conditions (n=19). Error bars denote SEM; two asterisks, $p < 0.001$; one asterisk, $p < 0.05$ compared to NV.*

4.4 Discussion

Here, we extended the task reported in Chapter 3 to a rapidly changing loss function, and to semantically and structurally rich stimuli. This fulfilled two functions – asking

[2]Difficulty increases as you approach the PSE, and is therefore proportional to the absolute difference between 0.5 and the proportion of face answers – see Section 5.2.5 for definition.

whether Bayes-optimal decisions would be observed in this scenario, and providing a design that would be provide anatomically distinct correlates of two perceptual categories with high efficiency. In terms of the former, the data failed to show the optimal pattern of curve shifts seen for the Vernier discrimination task – BMC suggested that the best model for the data was *not* one with the same slope and different means in the different value conditions. Augmenting this analysis, classical statistical tests performed on the parameters from the full model showed that, across the group, the PSE did shift in the right direction, and only for the comparison between the neutral and house value conditions was the slope significantly different. As discussed above, one of the major benefits of simple paradigms in which optimal behaviour can be achieved is that this directly argues for the BBH (and when properly controlled, for the BCH). Failures of optimality are much harder to interpret, which relates to the classic problem with interpreting a failure to reject the null hypothesis.

In the Vernier task, the loss function changed every block, whilst here it changed every two trials, and there are a number of possible ways this could contribute to suboptimal performance. First, it might be that there are limitations in how rapidly the brain can combine uncertainty with changing loss functions – this doesn't mean, however, that posterior uncertainty is not represented and available for use in computation. Second, it could be that observers find it hard to 'keep in mind' the rapidly changing rewards, and use a cognitive strategy to deal with this that obscures more straightforward processing of uncertainty. Third, it might be that the optimal combination of uncertainty and utility occurs in a less discrete way than suggested by Equation 2.9. If utility affects sensory processing, rather than being combined with a separate representation of uncertainty (a hypothesis we consider in Chapter 5), this might not be a 'one-shot' process. This highlights the need to consider the algorithms that could support the optimal integration of uncertainty and value, rather than assuming they match the form of the computational description of the problem. In the next chapter, we take a first step towards providing some anatomical, implementational constraints that could shed some light on this question.

Recent work has configured cognitive tasks that seem to embody suboptimal reasoning in a Bayesian framework – for example, concept learning (Goodman et al., 2008), inductive reasoning (Tenenbaum et al., 2006), duration and magnitude judgements (Griffiths and Tenenbaum, 2006), and the role of coincidences in assessing causal relations (Griffiths and Tenenbaum, 2007). These studies often argue that what might be classified as heuristics and biases from a neuro-economic perspective (Kahneman and Tversky, 1979; Glimcher and Rustichini, 2004) should be viewed as constraints and priors on inductive inference, and that reasoning within those bounds is then optimal. The debate about whether human reasoning is best described as irrational, or optimal under constraints, and how this relates to genetic and cultural evolution, is ongoing (see Gigerenzer, 2002), but for our purposes

draws attention to the issue of ecological priors discussed in the literature review. Cognitive strategies for dealing with rapidly changing loss functions could perhaps be embodied in a prior that would render reasoning ecologically 'optimal', though it is important not to define such a prior by working backwards from an assumption of optimality (see page 34). However, we wanted to avoid such biases, probing the simple combination of uncertainty arising from the likelihood with an external loss function – for future work, developing paradigms that better measure or minimise ecological prior assumptions is critical.

To further complicate the picture, the loss function schedule was not the only possible source of suboptimality – we also swapped simple Vernier offsets for a complex face-house phase continuum. Bayesian analyses that rely on plotting psychometric functions have not previously used high-level object classes, in part because producing a continuous stimulus axis is difficult (see page 105). Although the psychometric function gave a good fit to the data, examining Figure 4.3 suggests a bias towards face stimuli – when observers stand to lose less money by answering 'face' they appear more ready to do so than they are ready to answer 'house' when that is the less costly option. This effect is especially noticeable when they should be quite certain of the answer – i.e. at the left hand edge of the blue psychometric curves in Figure 4.3 – and might contribute to the trend for increasing slope as you move from FV to NV to HV (see Figure 4.4). This again might represent a bias in the prior that we do not model – faces are emotionally and evolutionarily highly significant, and as we will discuss in the next chapter, top-down attentional or arousal effects to particular stimulus classes could change sensory processing and thus uncertainty.

In the Bayesian decision formalism, a sensory posterior is computed then combined with a loss function to yield optimal decisions. In Section 2.3.2, we discussed how the psychometric curve for a categorical decision expresses $p(o_j \mid d_i)$, with the y-axis giving the relative proportion of each answer for stimulus values plotted along the x-axis. However, when the loss function $p(U \mid o_j)$ makes outcomes differentially valuable following different decisions, the psychometric function observed in behaviour expresses the *combination* of $p(o_j \mid d_i)$ and $p(U \mid o_j)$. The analysis laid out in Chapter 3, based on a Gaussian noise model, enables us to 'read' this combination from the psychometric function – the slope expresses the sensory uncertainty and the shift expresses the influence of the loss function. When the psychometric functions have different slopes this mapping no longer holds, and this might be due to non-Gaussianity of the face-house dimension. An important challenge for future work is to find complex stimuli that more naturally embody simple noise distributions, or to develop sophisticated models of more complex sources of uncertainty.

This study was designed in part to address the conditions under which Bayesian inference can be optimal, but was also designed to provide a paradigm ideally suited to a scanning study for investigating the neural correlates of the combination of sensory uncertainty with

external value. As behaviour in the psychophysical study was not optimal, we are clearly not looking for correlates of *optimal* performance. However, the purpose of this study was not to conduct an fMRI version of the integrative methodology illustrated in Figure 1.3, which would present many methodological and theoretical challenges we have not addressed – not least the development of a neural coding model that maps from PPCs to BOLD signal. The purpose of this study was rather to start expanding our understanding of the functional anatomy of the elements of Bayesian decision making. This constitutes an important unanswered question even couched in non-Bayesian terms, and contributes to a 'road-map' that future integrative studies will need in order to identify neural populations for more complex tasks. As such, the qualitatively optimal strategy we observed here is sufficient, and further work on pulling apart the components of sub-optimality was set aside.

5

The Anatomical Basis of Combining Uncertainty and Value

A Bayes-optimal decision maker integrates a posterior belief distribution with a loss function when making perceptual judgements. Behavioural studies have shown that observers can follow this normative prescription (see Knill and Pouget, 2004, for a review), but there has been little exploration of how the neuroanatomical correlates of perceptual uncertainty overlap with those of value-based decision-making (see Heekeren et al., 2008; Rangel et al., 2008, for reviews of each). In this study, some of the participants from the face-house categorisation study reported in Chapter 4 repeated the task in an fMRI scanner, with psychophysical data used to parameterise analysis of the imaging data. This enabled us to ask where along the path from sensory representation to motor command the effects of external value can be observed, and whether perceptual uncertainty is reflected in the same regions as probabilistic reward contingencies. We show that the effect of external value is associated with a cortico-striatal loop previously implicated in the computation of expected utility, and is not observed in face- and house-selective regions of sensory cortex. The difficulty of the perceptual decision was reflected in the ACC, and cumulative feedback in ventral striatum and medial PFC, again consistent with analogous components of value-based decision making. Changes in sensory uncertainty were however reflected in FFA, suggestive of a posterior representation in sensory cortex that is transmitted to an action selection mechanism alongside value signals, rather than being modified by them. We consider the advantages of such an architecture, and discuss the relationship of the present results to previous neuroimaging, neurophysiological, and theoretical studies on decision making.

5.1 Introduction

In Chapter 4, we reported results from a face-house categorisation task in which optimal behaviour required the integration of sensory uncertainty with knowledge about an external loss function. This was analogous to the Vernier task reported in Chapter 3, but with complex stimuli and rapid changes in value. As in the Vernier task, observers used the qualitatively optimal strategy of shifting their psychometric performance in the direction of the lower cost stimulus. However, they did not exhibit quantitative optimality, with sensitivity changing across value conditions and therefore unable to serve as a proxy for uncertainty according to Equation 3.18 (see Table 4.1 on page 112). In this chapter[1] we report the results from an fMRI study in which some of the same participants performed another session inside the scanner, allowing us to obtain measures of regional brain activity associated with task performance over time.

The lack of behavioural optimality places a caveat on the interpretation of the imaging data – we are looking for correlates of the components of a decision-process relevant to Bayesian decision theory, rather than correlates of an *optimal* Bayesian decision. This represents a novel question from the neurobiological perspective, and also contributes to the development of an anatomical 'road-map' that future integrative studies can use to identify neural populations that might represent components of a particular Bayesian inference or decision. If we had intended to use fMRI data to argue for the BCH via a tight integrative loop of the sort illustrated in Figure 1.3, optimality would clearly have been crucial, but such a study waits on sophisticated neural coding models that map PPCs onto the BOLD signal, which may turn out to lack the specificity required to argue for the implementation of specifically Bayesian algorithms.

5.1.1 Bringing Together Perceptual Uncertainty and Value

Bayesian decision theory invokes elements of both perceptual and value-based decision making – sensory uncertainty is taken into account in the computation of a posterior belief distribution, and combined with a loss function to generate decisions that maximise expected utility. However, as described in Section 2.3, the neurobiological study of perceptual decision-making has proceeded largely separately to that of value-based choice, each focusing on different elements of the architecture schematised in Figure 2.4. Neuroeconomic

[1] This work was the result of a collaboration between myself and Dr Maneesh Sahani at the Gatsby Unit, Mr Stephen Fleming, Prof. Ray Dolan, and Prof. Chris Frith at the Wellcome Trust Centre for Neuroimaging, and Dr Oliver Hulme at the Institute for Ophthalmology. See page 13 for details of contributions.

studies of value-based decision making have focused on **valuation** in goal-directed control, and on **learning** action values that contribute to habit formation (see for a review Rangel et al., 2008). There is less known about the basis of **representation** of decision options and relevant internal and external states, and about the mechanisms of **action selection**. Neurobiological studies of perceptual decision-making have had the opposite focus, with much evidence for stimulus representation in sensory cortex and evidence accumulation in parietal and prefrontal regions (see for a review Heekeren et al., 2008), where neuronal responses are consistent with a sequential probability ratio test embodying Bayesian principles (see for a review Gold and Shadlen, 2007). There is some work showing that the magnitude of reward (Platt and Glimcher, 1999), task difficulty (e.g. Kim and Shadlen, 1999), and probability of reward (e.g. Yang and Shadlen, 2007) impacts on fronto-parietal evidence accumulation, but how value and sensory uncertainty interact anatomically is largely uncharted territory.

In this study we aimed to bring these two domains together, asking where sensory evidence (contributing to uncertainty in $p(o_j \mid d_i)$; see Figure 2.5) and external value (a deterministic $U(o_j)$ mapping; see Figure 2.5) are combined to compute the expected utility (EU) of a decision (see Equation 2.9), and whether perceptual uncertainty is reflected in the same regions as externally imposed contingencies. To achieve this, we used a paradigm that controls for other components of decision-making – we encouraged a consistent emotional and cognitive set, used a value manipulation likely to map straightforwardly onto utility, and removed trial-by-trial feedback. To come back to the Bayesian decision maker, this study also addresses the question of whether perceptual uncertainty and value are represented separately, and then integrated elsewhere, or whether value impacts on the posterior distribution itself. In the remainder of the introduction we focus on specific hypotheses about the anatomical correlates of external value and sensory uncertainty, based on the literature surveyed in Section 2.3.

5.1.2 Where Might Effects of External Value be Observed?

For both perceptual and value-based decision-making, there is evidence for cortico-striatal circuitry involved in integrating different value signals to determine the expected utility of decision options. Evidence from neuroeconomic studies suggests that $p(o_j \mid d_i)$ is encoded in the basal ganglia (e.g. Yin et al., 2005) and $p(U \mid o_j)$ emerges in the OFC and dlPFC (e.g. Wallis and Miller, 2003; Hare et al., 2008). Integrated EU/PT signals (see page 56) are reflected across this circuit and are also reflected in the activity of ACC and dopaminergic midbrain regions (Tom et al., 2007; Rolls et al., 2008; Balleine et al., 2008). In perceptual decision making, there is also evidence for cortico-striatal circuitry, but here the basal ganglia are often thought of as implementing threshold crossing for the evidence

accumulation process occurring in PFC neurons, which is driven by sensory processing in specialised regions of occipital and temporal cortex (Bogacz, 2007; Gold and Shadlen, 2007).

One of the key questions we were interested in was where the effects of external value would be observed. Changes in reward constitute deterministic mappings from outcome to utility; $U(o_j)$, and might therefore be reflected across the cortico-striatal valuation circuit (see Rangel et al., 2008; Lo and Wang, 2006; Bogacz and Gurney, 2007). Changes in external value also constitute changes in the variance of utility (often termed 'risk'), which has been associated with activity across a cortico-striatal circuit, particularly in striatum, insula, and medial OFC (Preuschoff et al., 2006; Dreher et al., 2006; Rolls et al., 2008; Preuschoff et al., 2008; Tobler et al., 2007).

Diffusion-to-bound models of perceptual decision making, in which neurons are thought to implement an SPRT 'decision-variable', have made famous steps in linking probabilistic models to neural variables (Gold and Shadlen, 2007). However, as we discussed in Section 2.3.3, such models are one restricted instance of a Bayesian decision algorithm, and cannot easily incorporate other modulators of value. There is some evidence that the magnitude (Platt and Glimcher, 1999) and probability (Yang and Shadlen, 2007) of reward affects primate LIP neurons, and that task difficulty is reflected in parietal and prefrontal regions both in primates (Kim and Shadlen, 1999; Shadlen and Newsome, 2001; Romo and Salinas, 2003), rodents (Kepecs et al., 2008), and humans (Heekeren et al., 2004; Ploran et al., 2007; Thielscher and Pessoa, 2007). However, the SPRT model is silent as to the effects of external value, making the interpretation of such results difficult. A more general Bayesian formulation, such as the POMDP approach proposed by Dayan and Daw (2008), has the potential to address such questions, but for our purposes we can take a broad-brush approach to where the effects of value might be observed.

First, theoretical (Bogacz and Gurney, 2007) and preliminary electrophysiological (Ding and Gold, 2008) studies suggest that the basal ganglia act to set the threshold for a diffusion-to-bound type decision process, concordant with more general observations about the role of the BG in arbitrating between different decision options (Redgrave et al., 1999; McHaffie et al., 2005; Mink, 1996; Frank, 2006). We might therefore expect changes in value to be observed in these regions, as well as in the fronto-parietal 'decision-variable' regions discussed above. Finally, it might be that external value impacts on sensory processing itself – i.e. that the components of Equation 2.9 are not represented separately and then integrated elsewhere. There is evidence from other domains that top-down signals can alter processing in sensory regions – for example when attention (Johnson et al., 2007; Reddy et al., 2007; Vuilleumier and Driver, 2007; Wojciulik et al., 1998), emotion (Hsu and Pessoa, 2007; Vuilleumier et al., 2004, 2001) and task set (Summerfield et al., 2006a)

are manipulated. Indeed, changes in reward schedule have been reported as expressing attention-like effects on early visual areas (Maunsell, 2004; Shuler and Bear, 2006).

Decision-related areas in prefrontal and parietal cortex overlap with those involved in motor planning, at least in primates (Hernandez et al., 2002; Romo et al., 2004), and significant activations might therefore reflect accumulation of sensory information or the formation of a motor plan. The question about how separable decision variables are from motor plans is still under investigation, and reflects an underlying theoretical debate about how far back up the processing stream the brain should recognise that it must choose a single action (see pages 21 and 61 and Cisek, 2007; Wyss et al., 2004; Verschure and Althaus, 2003). Evidence from human neuroimaging studies suggests that when decisions are not intrinsically linked to particular motor responses, the two are dissociable (see Heekeren et al., 2006) and we would therefore be surprised to see correlates of decision-variables and their modulation by value in regions of the motor cortex associated with button presses.

5.1.3 Where Might Effects of Categorisation Difficulty and Perceptual Uncertainty be Observed?

As explained in Section 2.3.2 and illustrated in the upper grey box of Figure 2.5, the task we use in Chapters 3 - 5 can be characterised by a psychometric function that plots the continuous stimulus axis against the relative proportion of two categorical answers – here, whether the noisy stimulus is a face or a house. Two measures of sensory uncertainty, both of which contribute to $p(o_j \mid d_i)$, can be derived from the psychometric curve, and might have different anatomical correlates. The uncertainty with regard to a particular stimulus (or stimulus 'difficulty'; D_{stim}) increases away from the categorisation boundary in both directions. We can also characterise the uncertainty across the whole axis, measured by the slope of the psychometric function, σ^2_{stim}, and corresponding to signal detection 'sensitivity' (see Equation 5.2). If behaviour is optimal under a simple Gaussian noise model, this quantity can also serve as a proxy for the width of the posterior that reflects sensory uncertainty (see Equation 3.18). This was not the case for the face-house psychophysics reported in Chapter 4, and so the definition of sensitivity here is a behavioural one.

Difficulty affects expected utility – whether or not you shift your psychometric function correctly, the probability of getting the answer right modulates the probabilistically weighted average of the reward obtained (see Equation 2.9). We therefore expect to see correlates of difficulty in regions reflecting EU/PT signals, and perhaps specifically in regions thought to encode $p(o_j \mid d_i)$. An fMRI study by Grinband et al. (2006) found correlates of difficulty (or 'outcome uncertainty') in the medial frontal gyrus, anterior insula, ventral

striatum, and dorsomedial thalamus, and a related study in which Critchley et al. (2001) measured BOLD in the delay between a categorisation decision and feedback about the outcome, found correlates in the ACC and OFC. It is important that these two studies dissociated correlates of difficulty from attentional and arousal-related networks respectively, as more difficult or riskier decisions can invoke higher levels of attention and arousal (see Itti et al., 2005, and Section 5.4.2). External value and difficulty both contribute to expected utility, and how the components of valuation play out across a cortico-striatal-thalamic circuit is still somewhat unclear. We therefore used a masking strategy to identify regions that are correlated with external value and categorisation difficulty independently, and where they overlap.

If the BCH is correct and the brain represents information about uncertainty at every stage of processing, uncertainty about the face-house continuum should be represented in the appropriate stimulus-selective regions as well as potentially modulating other components of the decision-making circuit through its effects on expected utility. We would therefore like to be able to look for correlates of the slope of the psychometric function, but since we did not explicitly manipulate sensitivity it cannot serve as a trial-by-trial regressor. However, the changes in sensitivity with changes in value that prevented us from running quantitative optimality analyses (see Table 4.1) do allow us to look for regions where the change in activation due to change in value correlates with concurrent changes in uncertainty. The caveat is of course that the slope does not serve as a direct proxy for the internal sensory noise as in Chapter 3 – we will discuss its interpretation further in Section 5.4.2.

5.1.4 Tracking other Elements of the Decision

In the present study we are interested in regions that correlate with changes in value, and with changes in sensory uncertainty and stimulus difficulty. These regions might include sensory and motor areas, as well as the decision and valuation network discussed above. We therefore perform simple analyses to identify these regions, checking that face perception correlates with the FFA and house perception with the PPA (Dolan et al., 1997; Summerfield et al., 2006b), and for button-press related activations in the primary motor cortex (M1).

As indicated in Figures 2.4 and 2.5, the evaluation of outcomes, and the comparison of what was expected to what was actually obtained, is used to drive learning about the components of valuation. Here we avoid trial-by-trial feedback, but observers still receive periodic information about their endowment, and so we might expect to see correlates of both periodic reward value and perhaps prediction error. In human fMRI studies, correlates of monetary reinforcement have been observed in medial OFC (e.g. Knutson et al., 2001;

O'Doherty, 2004), and correlates of negative outcomes such as pain in the insula and ACC (e.g. Davis et al., 1997). In our study, we might expect the regions involved in processing positive reinforcement, such as medial OFC, to be inversely correlated with the amount of money lost at every feedback screen. If participants do form a prediction about their total endowment over the period intervening between the cumulative feedback screens, we might also expect to see activity in regions that correlate with prediction error such as ventral striatum and midbrain dopaminergic nuclei (see Montague et al., 1996; Hare et al., 2008; Summerfield et al., 2006a; O'Doherty, 2004), as well as correlates of experienced reward in OFC and ventromedial frontal cortex (e.g. Knutson et al., 2001; O'Doherty, 2004).

5.2 Methods

5.2.1 Participants

Of the nineteen participants who took part in the initial psychophysics study, sixteen were scanned (5 male; 19-27 years of age; mean age, 24.2 years). One participant was excluded at this stage due to a severe change in response strategy in the scanner compared to the behavioural experiment, leaving fifteen observers in the scanning study. All had normal or corrected-to-normal vision, and no reported history of psychological or neurological illness. The study was approved by the Institute of Neurology (University College London) Research Ethics Committee.

5.2.2 Stimuli and Equipment

Face/house images were again presented for 100ms on a grey background using Cogent 2000 (www.vislab.ucl.ac.uk/cogent.php) running in MATLAB. During the fMRI experiment, stimuli were presented using an LCD projector running at 60 Hz, viewed by observers 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 (i.e. to subtend 4° of visual angle).

5.2.3 Procedure

The fMRI experiment took place within a week of the psychophysics, and employed the same task with minor alterations (see Figure 4.3). First, due to timing constraints, participants completed four runs of 105 trials. To roughly match the payoff from the psychophysics, the initial endowment for each block was increased to £12, and cumulative feedback was given every 10 trials. Second, stimulus timings were jittered in a fast event-related design – the fixation cross that preceded the stimulus was presented for between 0.1 and 3s, and the fixation cross that followed it remained on the screen for between 1.6 and 3.6s, giving an average trial length of 5.9s. Third, the decision-to-motor mapping was changed halfway through the session, so that face and house decisions were made with both left and right button presses by each participant. This allowed us to investigate whether changes in value specifically affected the motor response, by decoupling it from the decision category (Thielscher and Pessoa, 2007). To avoid switch costs, a short training run was given with the new response mapping without any imaging data being collected.

Stimuli were randomly permuted over time, so that the full phase range was covered on every ∼7 trials. Similarly, the three cost levels were cycled every 6 trials (changing every two trials, as in the psychophysics), while keeping stimulus phase and cost orthogonal. This cycling over ∼30s matched the filter properties of the canonical haemodynamic response function (HRF), thus maximising power for estimating the cost- and stimulus-related parameters in our event-related analysis.

5.2.4 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 x 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). Heart rate was monitored using a pulse oximeter, and respiration was recorded using a breathing belt. Four runs of 213 volumes were collected for each participant, followed by a T1-weighted anatomical scan and local field maps.

5.2.5 Data Preprocessing and Analysis

Functional data were analysed using SPM5 (Statistical Parametric Mapping; 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, field maps were estimated from the phase difference between the images acquired at the short and long TE (toolbox available at http://www.fil.ion.ucl.ac.uk/spm/toolbox/fieldmap). The EPI images were then realigned and unwarped using the created field map, and slice-timing correction applied to align each voxels timeseries to the acquisition time of the middle slice. Each participant's T1 image was segmented into grey matter, white matter and cerebrospinal fluid, and the segmentation parameters were used to warp the T1 image to the SPM Montreal Neurological Institute (MNI) template. These normalization parameters were then applied to the functional data. Finally, the normalized images were spatially smoothed using an isotropic 8mm full-width half-maximum Gaussian kernel.

fMRI timeseries were regressed onto a composite general linear model (GLM) containing delta (stick) functions modelling the onsets of the precue, 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-aligned delta functions were separated into three regressors depending on the value condition on each trial; FV, NV and HV.

Each was then parametrically modulated by two observer-specific functions. The first was the choice probability (CP) curve fitted to the out-of-scanner psychophysics data in the neutral value condition (see Equations 3.15 and 3.17), yielding three new regressors; FV_{CP}, NV_{CP}, HV_{CP}. The curve fit was taken from the full nine parameter model that models the value conditions independently, rather than from the best model as determined by BMC. This avoids any bias that might be induced by fitting shared parameters, and also reflects the observer's sensitivity without any affect of asymmetric value. The second was the trial-by-trial difficulty function (D), which reflects the CP function about the PSE such that there is maximal uncertainty about the answer at the PSE, and minimal uncertainty at 100% face or house phase, with the values scaled to lie between 0 and 1 (see Grinband et al., 2006, for a similar manipulation). Where CP_{ij} is the choice probability for the ith phase mixture and the jth value condition, the difficulty function is;

$$D_{ij} = \frac{1}{2}\big|\,|\,0.5 - CP_{ij}| - 0.5\,\big| \tag{5.1}$$

The D function was orthogonalised with respect to the CP function, and yielded a final set of three parametrically modualted regressors; FV_D, NV_D, HV_D. This varying difficulty of categorisation with different phase compositions is one measure of uncertainty in the decision-outcome mapping, the other being expressed in the slope of the psychometric function across the full phase axis (see page 57 and Section 5.3.6 below). The feedback-aligned delta functions were also parametrically modulated, with the amount of money lost from the endowment on the previous 10 trials.

To investigate interactions of value and response, the response-aligned delta functions were separated by value, decision and response hand, giving a 3 x 2 x 2 factorial combination. Physiological noise parameters were entered as regressors of no interest in the design matrix following decomposition into Fourier frequency components (Josephs et al., 1997; Birn et al., 2006). Motion correction regressors estimated from the realignment procedure were also entered as covariates of no interest.

5.2.6 Statistical Inference

Statistical significance was assessed using linear compounds of the regressors in the GLM, generating statistical parametric maps of t-values across the brain for each participant 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 to assess group-level significance. Cluster-based statistics (Friston et al., 1994) were used to define significant activations both on their intensity and spatial extent. Clusters were defined using a threshold of $p < 0.001$ and corrected for multiple comparisons using family-wise error correction (FWE) and a threshold of $p < 0.05$. For presentation purposes, images are displayed at $p < 0.005$ uncorrected.

Estimated time courses in regions of interest (ROIs) are plotted at seven TRs following stimulus onset using a finite impulse response (FIR) model implemented in the MarsBar ROI toolbox (Brett et al., 2002). We only plot time courses in ROIs after establishing their significance in a conventional SPM (see Figures 5.7 and 5.9).

5.3 Results

5.3.1 Behavioural Analysis

Figure 5.1 shows the behavioural data for each of the fifteen participants who took part in the scanning study (positions of each observer's data match up with those in Figure 5.1). We used the same gradient descent procedure detailed in Chapters 3 and 4 to fit composite binomial error function models to the three psychometric functions for each observer. Here we just fit the full nine parameter model, with separate mean (μ), slope (ρ), and error rate (ϵ) parameters for each of the three value conditions. This was done purely for illustration purposes, as the data are too noisy to render these fits a reliable basis for inference. The analysis of group trends in PSE and slope across value conditions, and BMC analyses on individual observers' data, was therefore conducted on the out-of-scanner data reported in Chapter 4. For the same reason, the choice probability (CP) and difficulty (D) parametric modulators were taken from the NV condition outside the scanner (see Section 5.2.5). Reaction times did not differ between value conditions ($F(2, 28) = 1.67$, $p > 0.2$), but significantly correlated with the difficulty regressor ($r = 0.56 \pm 0.21$, n = 15).

5.3.2 Signal Detection Analysis

The Bayesian Model Comparison conducted on individual observers' data from the out-of-scanner psychophysics found heterogenous mixtures of shared and separate mean and slope parameters for the best-fitting models, which is suboptimal according to the Bayesian optimality analysis we presented in Chapter 3. Paired sample t-tests on the PSE and slope from the full nine parameter models showed significant shifts of the psychometric function in the right direction, but a significant change in slope for the HV relative to NV condition (see Section 4.3). We wanted to conduct similar analyses on the in-scanner data, but without relying on psychometric function fits to noisy data. We therefore used an approximation to a 2-stimulus SDT analysis (see Green and Swets, 1966; Macmillan and Creelman, 2005, and page 21). This involved classifying stimuli either side of the NV as faces vs. houses, producing a 2 x 2 stimulus-response table in which the face category is treated as the 'signal' that the observer is trying to identify – face responses for stimuli to the face side of the PSE therefore constitute **hits**, and face responses for stimuli to the house side of the PSE constitute **false alarms** (see Table 5.3.2).

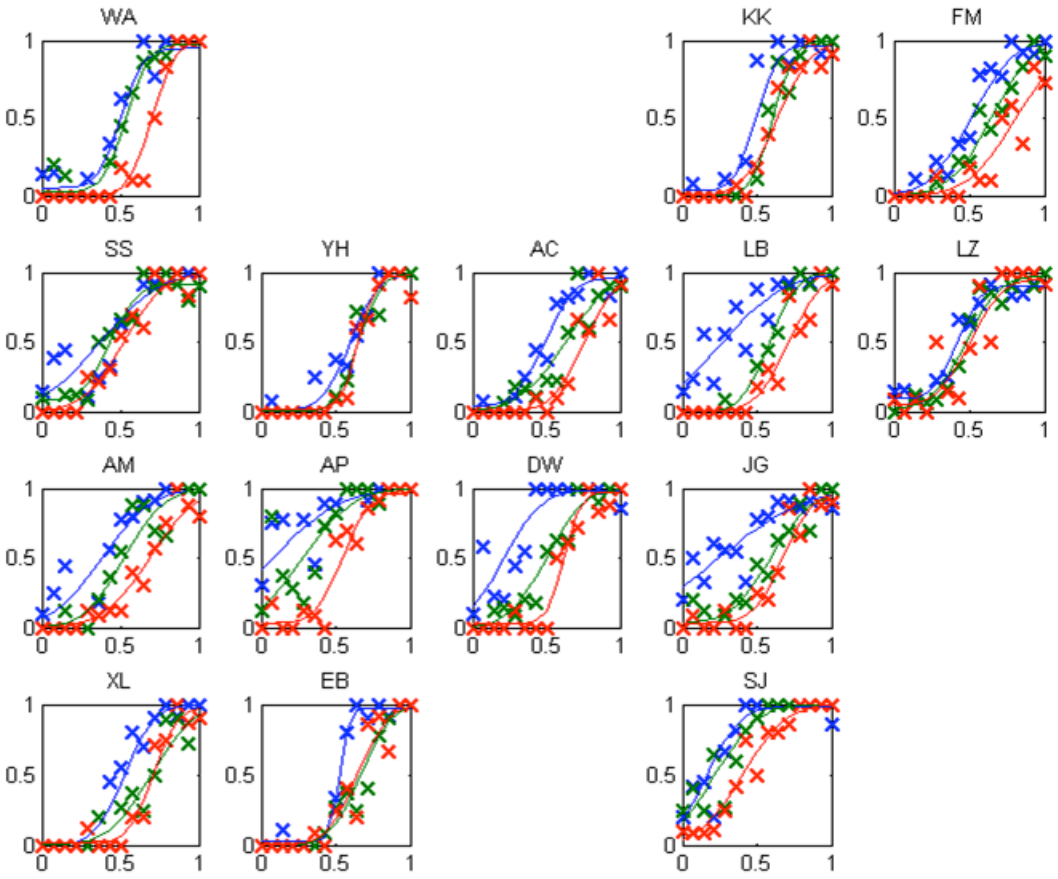

Figure 5.1: ***Behavioural data for individual observers in the scanner.*** *Shows individual choice probability data from the 420 trials of the in-scanner psychophysics, with each participant's data presented in the same grid location as their out-of-scanner data in Figure 4.3. The five empty locations correspond to those observers who took part in the out-of-scanner psychophysics but not the scanning study. On each figure, the abscissa represents the proportion of face phase in the image, and the ordinate the proportion of "face" responses to that stimulus in the different value conditions. Blue = face value (FV); green = neutral value (NV); red = house value (HV). The curves plotted through the data represented composite binomial error functions (see Equations 3.15 and 3.17 on page 89), fit using gradient descent to a full model with separate mean (μ), slope (ρ), and error rate (ϵ) parameters for each of the three value conditions. These fits are shown purely to illustrate qualitative trends, as the data are too noisy to render them a reliable basis for inference – statistical tests, BMC, and fMRI regressors were determined on the basis of the out-of-scanner psychophysics shown in Figure 4.3.*

	answer 'face'	answer 'house'
face stimulus	**hit**	**miss**
house stimulus	**false alarm**	**correct reject**

Table 5.1: ***Signal detection response types for a 2-AFC categorisation***

We can then straightforwardly compute a measure of signal detection sensitivity (d');

$$d' = \mathsf{z}\,(\text{hit rate}) - \mathsf{z}\,(\text{false alarm rate}) \tag{5.2}$$

and criterion (c);

$$c = -\frac{1}{2}\left(\mathsf{z}\,(\text{hit rate}) + \mathsf{z}\,(\text{false alarm rate})\right) \tag{5.3}$$

separately for each value condition and for each observer (Macmillan and Creelman, 2005; Green and Swets, 1966)[2], providing a robust basis for regression analysis. Sensitivity is closely related to the slope of the psychometric functions, and thus to the 'uncertainty' component of the decision (σ^2_{stim} in Figure 2.5)[3].

Figure 5.2b shows average d' and c across the group, suggesting that sensitivity remained roughly the same whilst criterion reflected the change in value as predicted. F-tests confirmed that value affected the decision criterion, c ($F(2, 28) = 52.4$, $p < 0.0001$), but not categorical discrimination ability, d' ($F(2, 28) = 0.41$, $p > 0.5$). This is in contrast to the finding that the slope for HV was significantly higher than for NV in the out-of-scanner data, and to the BMC results. The overall picture is of clear and significant shifts in the position of the psychometric functions, as predicted by the optimality analysis, but of variations in the slope that are evident in the sensitive BMC analysis and to some degree in t-tests on fitted slope parameters, but not in F-tests on the simpler categorical d' measure.

To examine how the signal detection parameters varied across the group, we plotted d' and c for each observer in each value condition in Figure 5.3. Figure 5.3b shows that criterion was consistently higher for the HV condition and lower for the FV condition compared to NV, as suggested by the group average. However, as shown in Figure 5.3a, the variation in d' across value conditions was more heterogenous. For some observers it

[2]In these equations, z is the inverse cumulative normal distribution function – for Gaussian distributions, computing d' thus corresponds to taking the difference between the areas under the 'signal' and 'noise' distribution that fall to the right of the criterion.

[3]There is a measure of signal-detection sensitivity more suitable for a continuous stimulus axis ('total d'') but this is more sensitive to outliers and does not yield a complementary measure of criterion, so we used the categorical version, and unless otherwise stated, this is what 'd'' refers to.

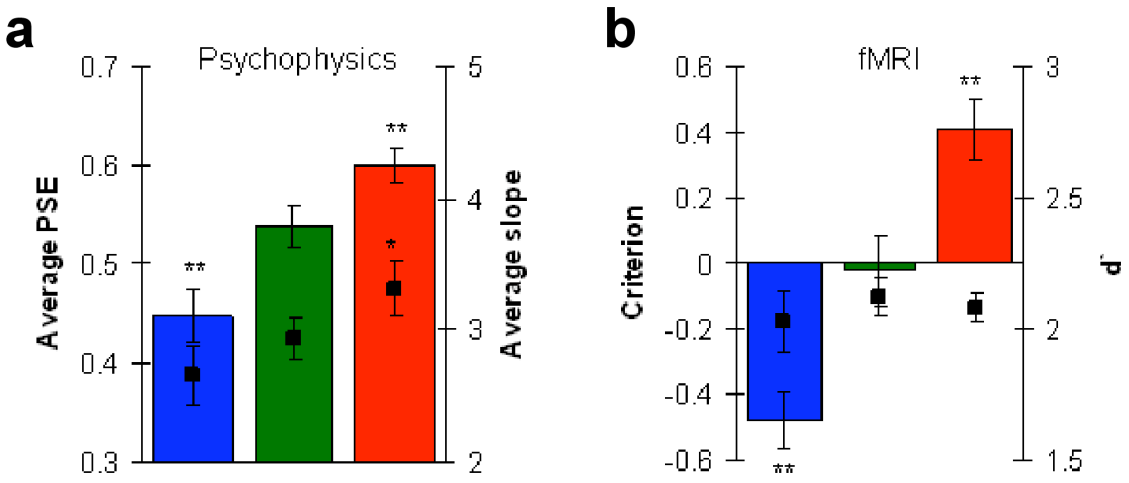

Figure 5.2: ***Decision parameters inside and outside the scanner.*** *Average parameters of the psychometric function fits to the psychophysics data, n = 19 (**a**), and for SDT analysis of the in-scanner data, n =15 (**b**). Coloured bars represent the PSE/criterion in FV (blue), NV (green) and HV (red) conditions. Black points indicate the average slope/d in each value condition for comparison. Error bars denote SEM; two asterisks, $p < 0.001$; one asterisk, $p < 0.05$ compared to NV. Panel **a** replicates Figure 4.4*

increased and for others it decreased, consistent with the Bayesian model comparison on the out-of-scanner data (see Table 4.1).

5.3.3 Stimulus-Selective Regions of Visual Cortex

We found activity correlating with increases in choice probability for faces in right FFA and right inferior occipital gyrus (IOG). Conversely, activity correlating with increased choice probability for houses was expressed in bilateral PPA, supporting previous reports that these regions respond to observers' beliefs about the relevant stimulus category (see e.g. Haxby et al., 1994; Heekeren et al., 2004; Dolan et al., 1997; Summerfield et al., 2006a). All contrasts were significant at $p < 0.001$, uncorrected, and are illustrated in Figure 5.4.

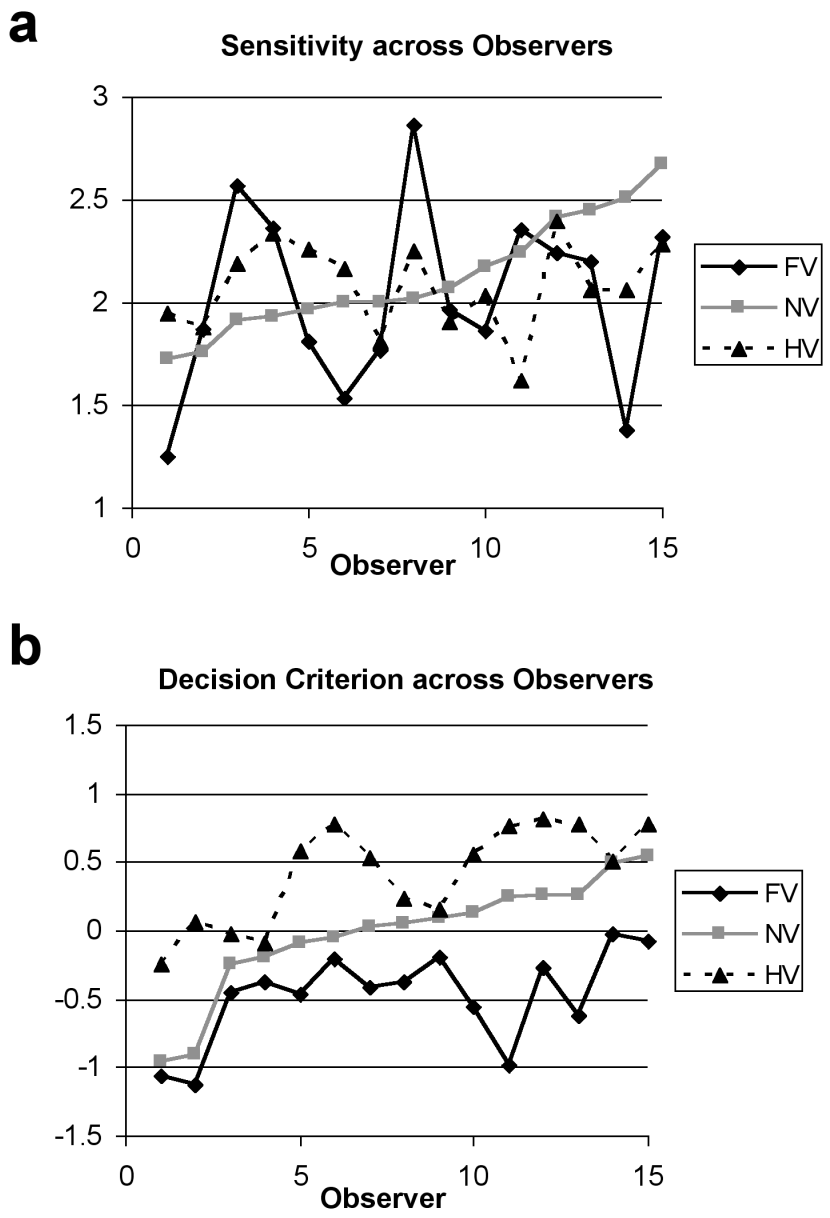

Figure 5.3: ***Parameters from signal detection analysis.*** ***a,*** *Variation in how value affects d', with observers ordered according to the NV d'. A lower average slope for FV and higher average slope for HV compared to NV is not obvious in the individual observers' data (and was not significant by F-test), but there is clearly a larger variation in d' for FV than for HV.* ***b,*** *Corresponding variation in how asymmetric value affects criterion (c), showing consistent shifts in the direction of the lower cost stimulus as predicted. Observers are ordered in terms of c in the NV condition.*

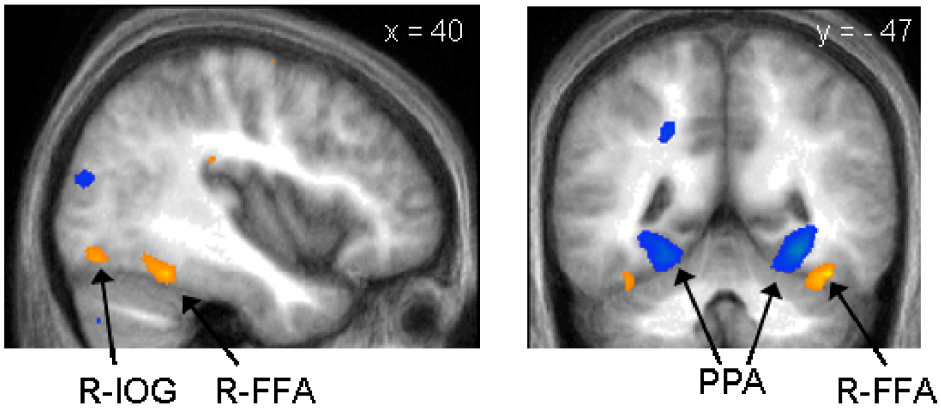

Figure 5.4: ***Category-selective activation in extrastriate visual areas*** *Saggital (x = 40) and coronal (y = -47) slices showing parametric effects of choice probability for faces (orange) and houses (blue). Effects were significant at* $p < 0.001$ *in right fusiform face area (R-FFA), [MNI coordinates (x, y, z)], 42, -48, -21 (peak* z*-score = 4.68) and right inferior occipital gyrus (R-IOG), 39, -75, -15 (*z*-score = 3.54) for face choice probability, and bilateral parahippocampal place area (PPA): left hemisphere; -24, -42, -15 (*z*-score = 3.59); right hemisphere; 33, -42, -9 (*z*-score = 5.03) for house choice probability. For display purposes images are shown at* $p < 0.005$.

5.3.4 Effects of External Value

We first wanted to test whether effects of external value would be observed in the stimulus-specific regions we identified (see Figure 5.4). Contrasts between the stimulus-aligned regressors in the two asymmetric value conditions, $[FV > HV]$ and $[HV > FV]$, failed to reveal any significant activations in the FFA/IOG and PPA respectively, even at a liberal threshold of $p < 0.01$, uncorrected. In a more stringent test, parameter estimates for the CP-modulated regressors FV_{CP}, NV_{CP}, and HV_{CP} were extracted separately from peak stimulus-selective voxels in the face and house-selective regions (see Figure 5.5). There were no significant effects of category-specific value in the response of these extrastriate areas ($F(2, 28) < 1.90$, $p > 0.17$). The lack of difference between neutral trials, and trials with the lower cost for each stimulus category, suggests that stimulus selective regions are not more active when the stimulus category they encode has higher value. The lack of difference between neutral trials, and trials with the *higher* cost for each stimulus supports this picture, and also indicates that any effect of greater attention to the riskier option is not expressed in these stimulus-selective regions (see Grinband et al., 2006).

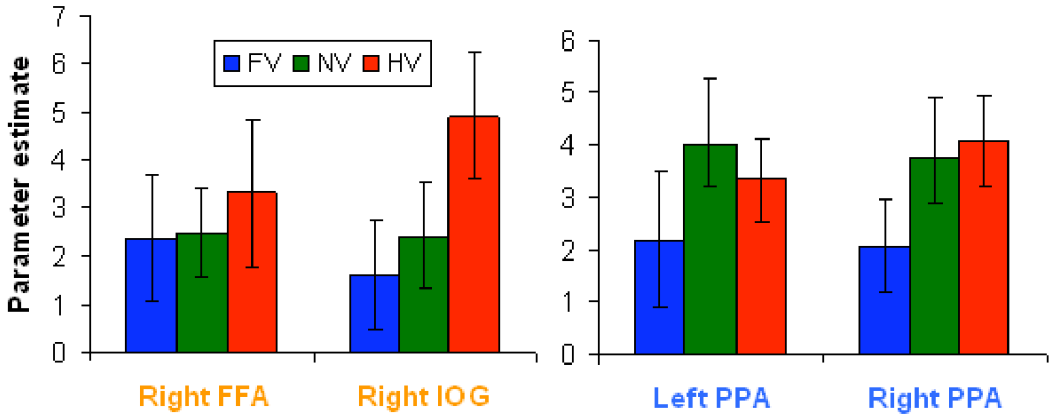

Figure 5.5: ***There is no effect of value in category-selective extrastriate visual areas.*** *Parameter estimates for the effects of choice probability in neutral and category-specific value trials, plotted at the peak voxel within each significant region identified in Figure 5.4. Post-hoc analyses of variance (ANOVAs) showed that no significant effects of value were found at any category-selective region in extrastriate cortex (all* $F(2, 28) < 1.90$*,* $p > 0.17$*). Error bars denote SEM.*

To search for the effects of asymmetric value in other regions, a whole brain analysis was conducted with the same contrasts; $[FV > HV]$ and $[HV > FV]$. No effects were found, suggesting that any regions responsive to value asymmetry are not category-specific, or else have a topological organisation whose scale is below that observable with the BOLD signal. Contrasts comparing the two asymmetric value regressors to the neutral value condition; $[FV > NV]$ and $[HV > NV]$, were therefore conducted, and revealed increases in activity on asymmetric value trials in left inferior frontal sulcus (L-IFS), bilateral dorsal premotor cortex (PMd), left caudate and left inferior parietal lobule, spanning the cortico-striatal valuation network discussed above. As shown in Figure 5.6, these activations were very consistent between the two category-specific value conditions (though note that the two contrasts are not orthogonal), and so we collapsed the two into the contrast $[(FV + HV) > 2\,NV]$. Figure 5.7 illustrates the resulting category-independent activations for all asymmetric value trials.

The consistency between the $[FV > NV]$ and $[HV > NV]$ contrasts was particularly strong for the IFS (see Figure 5.6). To gain a more detailed picture of the value-related activation in IFS we plotted the time course of the average fitted haemodynamic response for voxels in the significant cluster, showing increases in activity in both types of asymmetric value trial compared to neutral value baseline (Figure 5.8) across the time course.

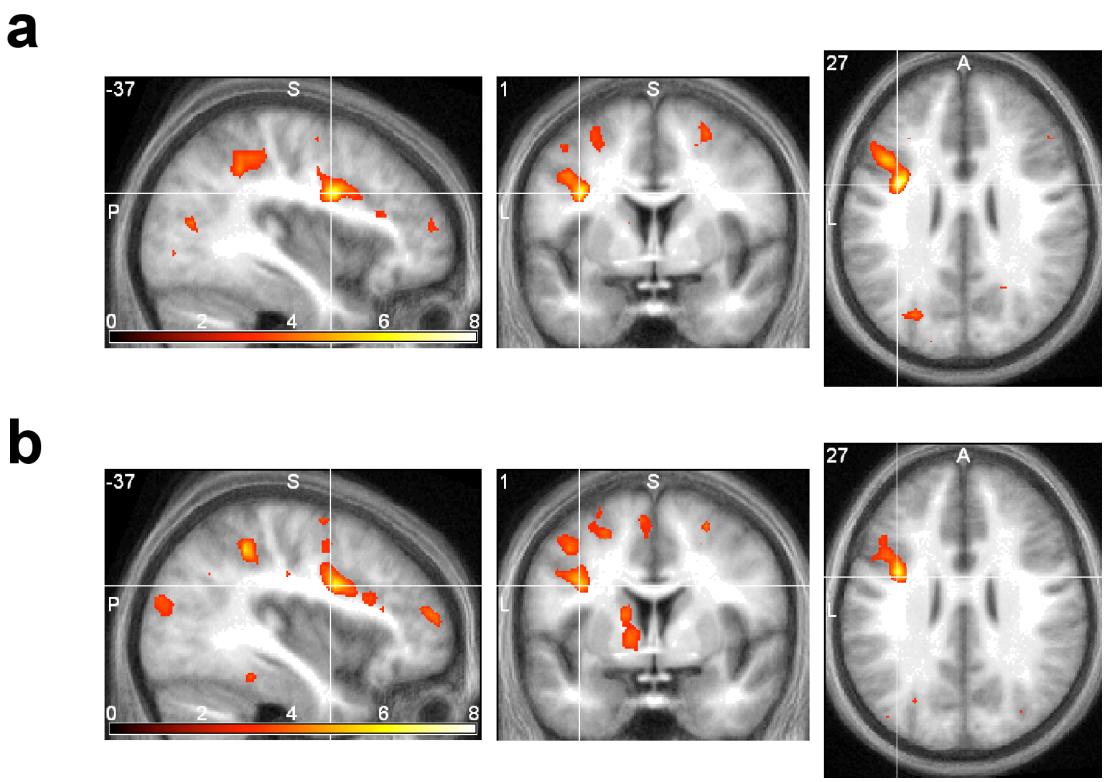

Figure 5.6: ***Effects of asymmetric value are consistent across category.*** *Images show sections through group T maps for* **a,** *[FV > NV] contrast and* **b,** *[HV > NV] contrast. Images are thresholded at* $p < 0.005$*, uncorrected. Crosshairs are located at the same point in every image (-37, 1, 27). Indicates that effects of asymmetric value are consistent for FV and NV trials, though note that contrasts are not orthogonal.*

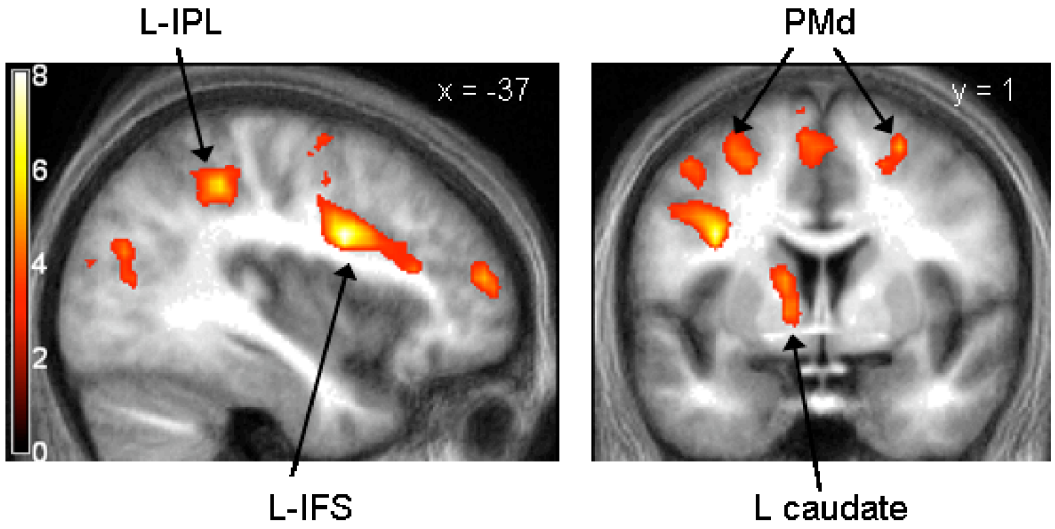

Figure 5.7: ***Effects of asymmetric value compared to neutral value trials.*** *Saggital (x = -37, x = -9) and coronal (y = 1) sections showing brain activations reflecting the main effect of asymmetric value averaged over category; [(FV + HV) > 2 NV]. Significant clusters were found in left IFS, [MNI coordinates (x, y, z)], -36, 3, 27 (peak* z*-score = 5.12); left caudate, -12, 3, 9 (*z*-score = 4.88); bilateral dorsal premotor cortex (PMd): left hemisphere, -27, -3, -51 (*z*-score = 4.64); right hemisphere, 27, 0, 57 (*z*-score, 4.03); and left inferior parietal lobule (L-IPL), -39, -42, 42 (*z*-score = 4.25); all cluster FWE corrected,* $p < 0.05$*; for display purposes images are thresholded at* $p < 0.005$*, uncorrected.*

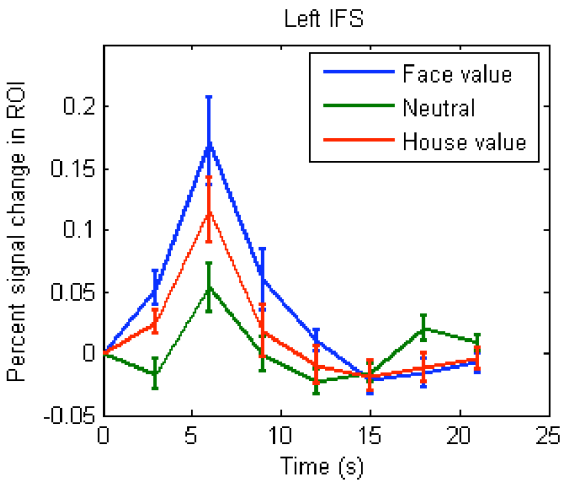

Figure 5.8: ***Consistent effects of both category-specific value conditions in IFS.*** *Haemodynamic response time courses for the three different types of value trial, plotted for the significant cluster in L-IFS (centre -37, 3, 27). Notably, both face- and house-value trials caused comparable signal changes in this area.*

5.3.5 Effects of Difficulty

To investigate trial-by-trial difficulty (see Section 5.2.5) we looked for regions correlating with the uncertainty-modulated stimulus-aligned regressors FV_D, NV_D, and HV_D. The dorsal ACC (dACC; or paracingulate gyrus), and right inferior frontal gyrus (IFG) showed significant activations (Figure 5.9), consistent with previous reports (Critchley et al., 2001; Grinband et al., 2006). As mentioned above (see Section 5.1.3), it is important to distinguish activations correlating with external value from those correlating with difficulty – both contribute to expected utility. We therefore masked each contrast with the regions that have significant correlation under the other. By exclusively masking the value contrast $[(FV + HV) > 2\,NV]$ for the regions that correlated with difficulty at a liberal ($p < 0.05$, uncorrected) threshold, we could show that L-IFS, left caudate and bilateral PMd were independently active under asymmetric value. L-IFS was the only region to meet the constraints of being both active independent of changes in difficulty, and significant in both types of asymmetric value contrast; $[FV > NV]$ and $[HV > NV]$. With the opposite mask, dACC was found to be active independently of value condition. We again plotted haemodynamic time courses for the three value conditions (Figure 5.9b), this time for the significant cluster in dACC (9, 36, 33). Unlike for the significant cluster in IFS (see Figure 5.8), there is no effect of value, supporting the identification of dACC as independently responsive to categorisation difficulty.

5.3.6 Effects of Uncertainty

We have two measures of uncertainty for each observer – one is the slope of the fitted psychometric functions for the out-of-scanner data, and the other is the categorical d' measure computed as a coarse analogue of the slope for the in-scanner data (see Section 5.3.2). Paired-sample t-tests on the change in slope between the neutral value condition and the two asymmetric value conditions found a significant difference for HV compared to NV, whilst an F-test found no significant effects of value on categorical d' across the group ($F(2, 28) = 0.41$, $p > 0.5$). The problem with the fitted slope parameter is that BMC (see Table 4.1 on page 112) indicated that for only some observers was a model with three different slopes the best fit, whereas the categorical d' measure is computed for each value condition independently, and is also a more robust measure. We therefore used categorical d' for the following analyses.

This study was designed to look for effects of external value, and we did not explicitly manipulate the observers' sensitivity, assuming that d' for each value condition remained constant throughout the experiment. Thus we cannot use d' as a first level regressor, and

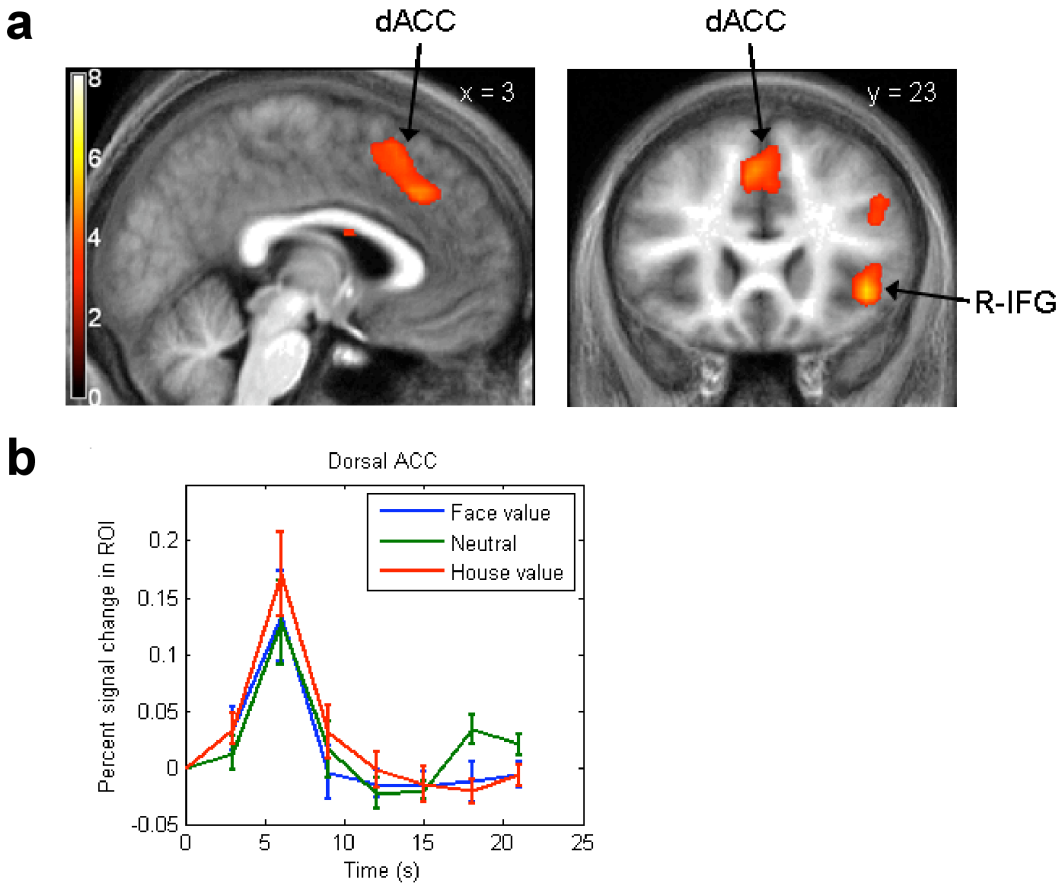

Figure 5.9: ***Effects of categorisation difficulty.*** ***a,*** *Saggital ($x = 3$) and coronal ($y = 23$) sections showing brain activations which correlate with an observer-specific D regressor, measuring how difficult the face-house categorisation task is on average for each phase-mixture. Significant clusters were found in dorsal anterior cingulate cortex (dACC) and right inferior frontal gyrus (FWE corrected, $p < 0.05$; for display purposes images are thresholded at $p < 0.005$ uncorrected).* ***b,*** *Haemodynamic response time courses for the three different types of value trial, plotted for the significant cluster in dACC (centre 9, 36, 33). While showing strong correlations with the D function regressor, this region was insensitive to changes in category value.*

are instead restricted to using it as a second-level interaction variable. In order to avoid problems with determining baseline, we are further restricted to looking at the covariation of uncertainty with *contrasts* on the first level. As described above, statistical tests revealed differences in sensitivity across value conditions that prevented us from conducting quantitative optimality analysis on the psychometric curve shifts. But this variation in d' with value also allowed us to ask where the contrast between value conditions correlated with *differences* in d'. As is evident in Figure 5.3b, d' in the face and house value conditions was higher than for neutral trials for some observers, but lower for others. We therefore used $[FV > NV]$ and $[HV > NV]$ contrasts as the first level effect, and ask where differences in d' between the two value conditions correlated with differences in the parameter estimate. This revealed a positive correlation with R-FFA ($r = 0.91$; $p < 0.0001$) and R-IOG ($r = 0.81$, $p < 0.001$) for the $[FV > NV]$ as illustrated in Figure 5.10a). This indicates that the heterogeneity across observers in how face-specific value affects sensory discrimination performance is effected via modulation of extrastriate visual areas. No effects of d' changes were found in house-selective regions, perhaps because the variations in d' for HV relative to NV were of a much smaller magnitude (see Figure 5.3b).

Contrasts with the stick functions modulated by choice probability (CP) and difficulty (D) functions seem like an appealing way to investigate interactions between uncertainty, stimulus categorisation, difficulty, and value. However, they are in fact very hard to interpret – if the magnitude of d' correlates with the correlation of the BOLD signal with the CP or D function, it does not tell you about the mapping between d' and CP or D. In addition, d' is of course related to parameters of both the CP and its rectified version D, introducing a confound into interactions. In the contrast reported in Figure 5.10 the correlation with the stick function regressor allows us to say that magnitude of d' change correlates with change in stimulus-related activity in the FFA. This is suggestive of a value-related change in sensitivity being expressed in the sensory representation, in contrast to the direct effects of value on decision criterion expressed in a fronto-striatal decision network (see Figure 5.7).

5.3.7 Correlation with Wins and Losses

As in the Vernier task, we provided only cumulative feedback on participants' endowment. From the Bayesian perspective, trial-by-trial feedback could be used to mimic optimal strategies via incremental threshold adjustment, without ever having to know about the uncertainty in the belief. In the scanning paradigm, this also avoids contamination of stimulus-locked responses with reward-related activity. However, we still expected to see anatomical correlates of periodic feedback, reflecting outcome evaluation and perhaps coarse predictions of reward generated on the basis of participants' introspective access to their

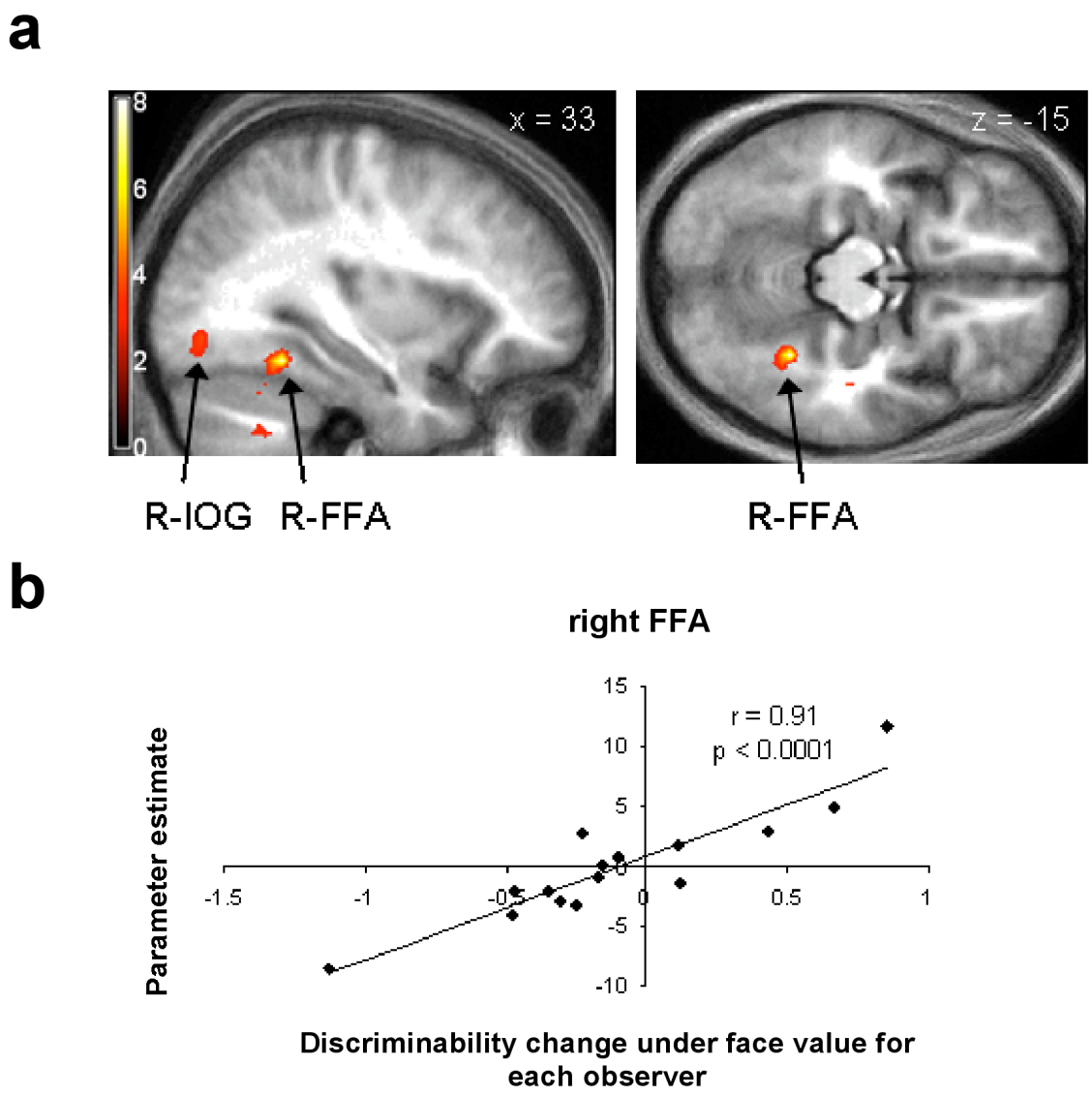

Figure 5.10: ***Individual differences in the effect of value on discrimination, a*** *Saggital (x = 33) and axial (z = -15) sections showing regions with significant correlations between the cross-observer change in category discriminability (d) in FV compared to NV trials, and the corresponding contrast image [FV > NV]. Right fusiform face area (R-FFA), [MNI coordinates (x, y, z)], 33, -48, -15 (z-score = 4.70) and right inferior occipital gyrus (R-IOG), 27, -69, -6 (z-score = 3.68), both* $p < 0.001$*. Images are shown at* $p < 0.005$ *for display purposes.* ***b,*** *Across observers, increases in R-FFA estimates are associated with better discrimination performance and decreases are associated with worse discrimination performance, compared to neutral value trials (correlation significant at* $p < 0.0001$*).*

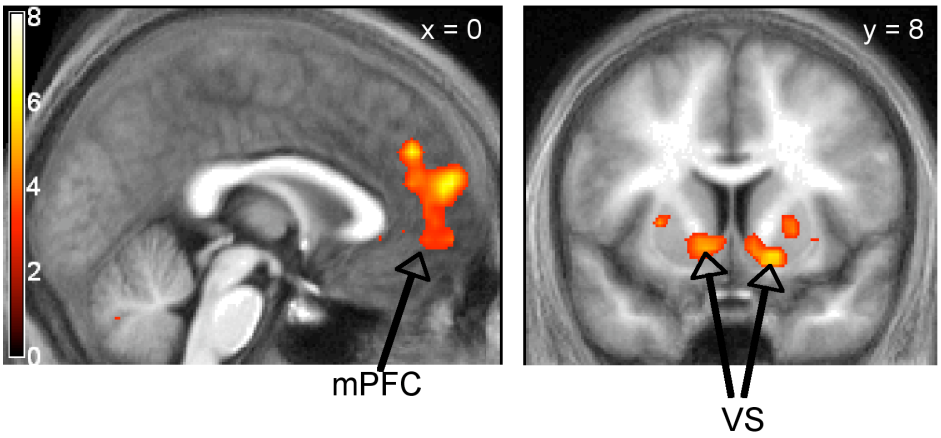

Figure 5.11: ***Brain regions inversely correlated with periodic 'loss' feedback.*** *Sagital ($x = 0$) and coronal ($y = 8$) sections show significant clusters in mPFC and bilateral ventral striatum. Clusters were significant at $p < 0.05$, FWE corrected; for display purposes images are thresholded at $p < 0.005$, uncorrected.*

performance. We found regions inversely correlated with the amount of money lost in medial prefrontal cortex (mPFC) and bilateral ventral striatum (Figure 5.11). Ventral striatum is firmly implicated in representing predicted reward and prediction error (see O'Doherty, 2004), and ventro-medial frontal regions have been previously implicated in coding reward value (e.g. Knutson et al., 2001; O'Doherty, 2004). Although in our study observers experienced only monetary losses, these were with respect to a total positive endowment, and so it is unsurprising that we did not find activations in regions such as the insula and ACC that have been found to correlate with negative outcomes such as pain (e.g. Davis et al., 1997) – it is interesting in this context that relief from pain can cause activations in midbrain and amygdala regions usually associated with positive reinforcement (Seymour et al., 2005). However, it is of course possible that for some participants, a framing effect (see De Martino et al., 2006) might have caused losses from an overall endowment to activate the insula/ACC.

5.3.8 Motor Activations and Interaction with Value

As expected, simple contrasts of motor-aligned delta functions found activity in M1, in the contralateral hemisphere to the hand that gave the response (data not shown). Although our theoretical perspective and related studies gave us little reason to expect that value signals would impact motor commands rather than the decision that generates them, we still wanted to test this possibility. We therefore assessed interactions of value and the motor

response using the contrasts: $[(FV_{left,face} + HV_{left,house}) > (FV_{left,house} + HV_{left,face})]$ and $[(FV_{right,face} + HV_{right,house}) > (FV_{right,house} + HV_{right,face})]$. This essentially asks whether there was any activity for the higher value stimulus that correlated with either motor response (left vs. right hand) or decision category ("face" vs. "house"). No such effects were found.

5.4 Discussion

In this study, we asked where BOLD activity would reflect the components of a perceptual decision under asymmetric cost: which regions would correlate with external value ($U(o_j)$ in Figure 2.5), stimulus difficulty (D_{stim} as derived from $p(o_j \mid d_i)$ in Figure 2.5), sensory uncertainty (σ^2_{stim} as derived from $p(o_j \mid d_i)$ in Figure 2.5), and cumulative feedback ($\sum_{trials} U(\hat{o})$ in Figure 2.5). Neurobiological studies of perceptual decision making have focused on evidence accumulation, and there is limited evidence for how reward value affects this process, and whether components of EU for perceptual decisions overlap with analogous components of value-based decision making. The novelty of this study from a neurobiological perspective is bringing all these components together in a human fMRI study where psychophysical performance can be simultaneously measured. From a Bayesian perspective, we wanted to contribute to the anatomical 'road-map' that will become increasingly important in gathering integrative evidence for the BCH, in more complex and multifarious settings than have so far been considered (see e.g. Knill and Pouget, 2004).

5.4.1 Behavioural Results

The out-of-scanner psychophysics reported in Chapter 4 indicated that observers, as with the analogous Vernier discrimination reported in Chapter 3, shifted their psychometric functions in the direction of the stimulus with lower cost. Paired-sample t-tests on the change in mean (μ) and slope (ρ) for the two asymmetric value conditions relative to neutral found that the means shifted significantly in the expected direction, but that there was a trend for the slope to be smaller for FV and higher for HV relative to neutral, which reached significance for the $HV - NV$ difference (see Figure 5.2a). A complementary SDT analysis was conducted on the in-scanner data (Macmillan and Creelman, 2005; Green and Swets, 1966), and F-tests found that criterion (c) changed significantly across different value conditions, but sensitivity (d') did not (see Figure 5.2b). Under this description, observers on average retained the same sensitivity across conditions, whilst shifting their decisions. However, BMC on the individual out-of-scanner psychophysics found that for only a few

observers was a model with a shared slope for the three psychometric functions the best explanation of the data, and clear individual differences in the pattern of d' variation were observed in the in-scanner data (Figure 5.3a).

Thus the picture from the behavioural data both inside and outside the scanner is of a qualitatively optimal strategy, but without the assumptions of the ideal observer model embodied in Equation 3.3.1 being supported. In Section 4.4 we discussed possible reasons for behavioural suboptimality, which included cognitive strategies that don't quantitatively reflect uncertainty, a Bayesian integration strategy that is not 'one-shot', or that behaviour reflected a more sophisticated noise model or ecological prior than embodied in our analysis. Future work will be important to tease these explanations apart, not just for this paradigm but more generally as more complex behavioural paradigms are viewed through the Bayesian lens. The anatomical results thus reflect components of a perceptual decision under sensory uncertainty and monetary risk that are brought together in the Bayesian picture, but do not represent correlates of their Bayes-optimal integration. Even if behaviour had been optimal, the potential of fMRI data to enter the integrative trinity schematised in Figure 1.3 waits on the development of appropriate neural coding models, and on confirmation that they would be sufficiently specific to distinguish the implementation of Bayesian algorithms.

5.4.2 Anatomical Results

We found correlates of stimulus processing in FFA/IOG and PPA for face and house stimuli respectively, as expected from previous studies (see e.g. Haxby et al., 1994; Heekeren et al., 2004; Dolan et al., 1997; Summerfield et al., 2006a), and found typical contralateral correlates of left and right-handed button presses in M1. No correlates of external value were found in sensory regions, suggesting that value did not change the decision threshold by biasing sensory processing. From the Bayesian decision-theoretic perspective, this is consistent with the computation of a posterior that is then flexibly combined with value in decision-making regions, rather than being itself modulated by reward contingencies.

We could argue that this makes sense for organisms living in environments where particular features can have both positive and negative valence, and where these contingencies can rapidly change. In the language of control systems, flexible action selection requires control mechanisms that do not solely depend on evidence accumulation (Stafford and Gurney, 2007). Thus, separating neural codes for sensory and value representation makes intuitive sense in an adaptable computational system (Maloney, 2002). The caveat to this discussion is of course that fMRI might not have sufficient spatiotemporal resolution to reveal subtle effects of value in sensory cortex, for example if the effect is not reflected in increases or

decreases in average activity. An electrophysiology study by Shuler and Bear (2006) found that individual neurons in rat primary visual cortex came to reflect reward timing when the animals learnt stimulus-reward associations – although this is not evidence for a change in sensitivity or criterion, it emphasises the importance of converging sources of evidence.

Correlates of external value were also absent from motor regions correlating with the button presses. In the Bayesian picture, it is hard to see how a go/no-go signal for motor commands could be optimally integrated with uncertainty, but uncertainty could be expressed through changes in 'motor' effects such as response speeding or vigor not measured in this paradigm (Bestmann et al., 2008; Niv et al., 2007).

Correlates of external value were found instead in L-IFS, PMd, caudate, and left inferior parietal lobule. This is consistent with the cortico-striatal valuation network found in neuroeconomic studies, especially with regard to the caudate activation previously associated with components of $p(o_j \mid d_i)$ (e.g. Yin et al., 2005) and with setting the threshold for diffusion-to-bound implementations of an SPRT (Lo and Wang, 2006; Bogacz and Gurney, 2007; Ding and Gold, 2008). Activation in premotor and parietal cortex is not usually observed for non-perceptual valuation (though see Roesch and Olson, 2004), and might suggest greater overlap of decision variables with motor plans for decisions involving the accumulation of sensory evidence, as observed in primate LIP studies (Gold and Shadlen, 2007). The L-IFS has been previously implicated in cognitive control, including the inhibition of motor responses (Aron et al., 2007; Schall et al., 2002), which is appropriate for a paradigm where the readiness to give particular responses is manipulated by value but perhaps conflicts with perceptual analysis. It is also part of a wider prefrontal region associated with integrating sensory evidence (e.g. Romo et al., 2004) and motivational context (Leon and Shadlen, 1999; Watanabe and Sakagami, 2007), and that reflects external value (see Rangel et al., 2008). Studies by Heekeren and colleagues using a similar paradigm found correlates of decision variables in the left superior, rather than inferior, frontal sulcus, but they did not manipulate value. In our study, finding more inferior loci, close to regions like the OFC and insula that are associated with integrating emotional and reward context, makes sense.

We found correlates of trial-by-trial categorisation difficulty (D_{stim}) in dACC and right IFG. Both are regions that reflect EU/PT signals, and the ACC has been previously implicated in representing categorisation uncertainty (Critchley et al., 2001; Grinband et al., 2006). Difficulty arises from the interplay between sensory uncertainty and the task definition, and a separately represented difficulty signal could be used to modulate various elements of the decision-making process (see also Kepecs et al., 2008). By masking out the external value and difficulty contrasts with respect to each other, we found dissociable activations in the L-IFS, PMd and caudate for external value, and dACC for stimulus difficulty. Changes in asymmetric value should have a greater impact the more difficult the

categorisation is, and at what stage in the decision-making process this interaction occurs is an open question.

We were unable to look for direct correlates of categorical d' but found that *changes* in d' for FV relative to NV were correlated with the *difference* in parameter estimates between FV and NV in the FFA. The Bayesian analysis used to demonstrate behavioural optimality in the Vernier task predicted that the slope of the psychometric function remains constant across changes in value, allowing it to serve as a proxy for sensory uncertainty. For the face-house categorisation task, BMC found that the best model for most observers was not one with a shared slope and separate means, such that we cannot interpret d' as directly reflecting sensory uncertainty. It rather indexes behavioural sensitivity, and changes in this sensitivity might be due to attentional or value driven effects. It is well known that attention can alter sensitivity (e.g. Carrasco et al., 2004), and there is evidence that this may be supported by neuronal effects in visual cortex (e.g. Vuilleumier and Driver, 2007; Wojciulik et al., 1998). However, a recent study by Pleger et al. (2008) also found evidence for an attention-independent, value-driven effect on sensitivity in primary somatosensory cortex, and in conjunction with the variety of effects value has on d' (see Figure 5.10), this argues against a simple salience-driven attentional effect in our study (see also Simoncini and Baldassi, 2008). The major effect of value, as seen in the group statistics, was a change in the PSE – the shifting of the psychometric function. The fact that the value contrast was not observed in FFA/PPA, with value correlating only indirectly with FFA via small changes in sensitivity, is suggestive of a picture in which value, like attention, can affect the fidelity of sensory processing, but where changes in criterion are implemented later in the decision-making process.

One component of value-based decision making we tried to minimise was the evaluation of outcomes and subsequent adjustments to parameters of the decision making process (see Figure 2.4 and Figure 2.5). We provided only periodic, cumulative feedback about the observer's current endowment, which prevented frequent reward activations from interfering with the main contrasts of interest. From the Bayesian perspective, this also ensures that integration of an estimate of sensory uncertainty with knowledge about external rewards cannot be mimicked by incremental threshold-adjustment. This is less pertinent in the face of suboptimal behaviour, but in conjunction with the extensive training observers underwent, suggests that correlates of value and decision making are uniform throughout the scanning session. However, we did find activations inversely correlated with cumulative losses that were consistent with the known representation of secondary reinforcement and reward prediction.

5.4.3 Conclusion

To conclude, our results constitute a step towards describing how value is flexibly integrated with sensory evidence during effective perceptual decision making. We find a high degree of overlap between the neural correlates identified here, and anatomical regions previously implicated in representing external value, stimulus difficulty, and reward evaluation. This supports theoretical suggestions about cortico-striatal decision circuitry, and questions the separability of representation, valuation, and action-selection (see Rangel et al., 2008, and Figure 2.4). There is however a suggestion that premotor/parietal decision-variables are more readily reflected in perceptual as opposed to value-based decision making. With the caveats discussed above about the sensitivity of the BOLD signal for revealing alternate implementations of a Bayesian decision, the results also support a picture of a brain that computes sensory posteriors which are then integrated with value information at a later stage in processing, though value (like attention) may affect sensitivity as well as adjusting response criteria. In future work we would like to explicitly separate sensory uncertainty from value, perhaps by externally manipulating the stimulus. We would also like to find a paradigm for which sensory correlates are anatomically separable, but where observers can still behave optimally, and to introduce priors into the picture (see e.g. Summerfield and Koechlin, 2008). In order for the BCH to become a broad framework for the perceiving, acting brain, mapping the components of probabilistic inference onto electrophysiological and neuroanatomical correlates is essential. Making the links more tightly constrained will require detailed neural coding models, and the careful triangulation of experimental, behavioural, and theoretical methodologies.

6

A New Probabilistic Framework for Selective Attention

The behavioural optimality proofs that form the backbone of evidence for the BCH are usually acquired in tasks with a small number of objects in the focus of attention, and the neural coding models that embody the underlying inferences correspondingly involve belief distributions over single features. When faced instead with complex, real-world scenes, the brain clearly fails to process everything optimally. A cornerstone of cognitive science is the idea that processing is enhanced in the focus of attention, and that this enhanced processing has a limited capacity, but we lack a good characterisation of Bayesian inference with and without attention. In this chapter we present a framework for thinking about capacity limits in the representation of complex posteriors, and extend previous ideas of attention as a Bayesian prior (Dayan and Zemel, 1999; Rao, 2005) to describe how attention locally enhances the approximate representations that result. In Chapter 5, our work on situating Bayesian inference in a neuroanatomical framework also served to address an unanswered question about how perceptual uncertainty and reward value are integrated in the brain. Here, in suggesting a Bayesian characterisation of neural capacity limits and their local resolution by attention we also help to unify apparently disparate neural bottlenecks, via a computational level description that admits multiple implementations. We demonstrate the framework with simulations in a simple, abstract model, replicating qualitative patterns of behavioural results in analogues of three key selective attention paradigms. Finally, we consider challenges for future modelling, and raise some open questions about how the framework might relate to models of hierarchical Bayesian inference in the brain (e.g. Rao and Ballard, 1999; Friston, 2005)

6.1 Introduction

Work on the Bayesian Coding Hypothesis has focused on demonstrating its plausibility – giving 'existence proofs' that observers can behave Bayes-optimally on a simple task in the focus of attention, and providing biologically plausible models for the neural substrate of probabilistic representation and computation. These existence proofs tend to deal with a single stimulus in the focus of attention, demonstrating a ceiling of inferential performance. In the real world scenes are composed of a multitude of spatially located, interdependent features, and the true posterior belief should therefore consist of a huge, highly correlated joint probability distribution. However, the resources needed to represent joint distributions grow exponentially in the extent of the correlations between the constituent variables. We suggest that the brain lacks the resources to represent the full posterior, due both to basic limitations on brain size and processing time, and to more specific features of a hierarchical, functionally specialised cortical architecture (see Section 2.4, page 66 of the literature review for further discussion). In the language of complexity theory (Papadimitriou, 1994), representing and computing over large joint distributions is *algorithmically intractable*.

In joint posteriors, correlations can come from several sources. First, the true prior should express knowledge about the fact that natural scenes tend to consist of a small number of sparsely distributed objects, made up of colocated and spatially extended features. The prior may also contain correlations between particular feature values based on the statistics of natural scenes, for example that textures corresponding to grass are likely to be green or yellow. Correlations in the likelihood can express knowledge about how the structure of firing rate observations is affected by the receptive-field (RF) properties of the neurons that generate them, and also contribute to 'explaining away' in the posterior. Explaining away is a central feature of Bayesian inference, and expresses the intuitive idea that if there are multiple possible causes of a particular observation, selecting one explanation makes the others less likely. For example, if you observed that the lawn was wet and knew it had rained last night you would be less likely to think the sprinkler had been on – in your posterior belief about causes of the lawn being wet, the two explanations are negatively correlated (this famous example appears in Pearl, 1988a). In a visual scene, for example, an object with a particular apparent size could be either small and close, or large and far away – inducing negative correlations between the possible combinations of true size and distance. Most basically, a single object cannot be in more than one place at once, making all possible locations of that object negatively correlated.

Below we will argue that the brain deals with the intractability of representing highly correlated joint posteriors by representing simplified approximations, and that attention helps to locally refine the impoverished representations that result. We will show using

simulations how this overarching framework can encompass disparate attentional phenomena, whilst also embodying many of the intuitions that have informed cognitive models. In this framework, the neural bottleneck is not to be found in a particular location, with particular functional parameters, or in a particular neurophysiological change, but is rather a fundamental and stringent constraint on computation throughout the brain. Critically, attention operates under this same constraint, avoiding the problem of conflating limitations in processing capacity with constraints on the attentional mechanism proposed to resolve them. In the final section of the chapter, we will consider the challenges for building detailed models of particular inferences under this framework.

6.2 Formalising the Attentional Framework

The fundamental premise of our framework is that the brain cannot represent full joint posteriors over real world scenes, due to a variety of implementational and computational constraints. Probabilistic descriptions of perceptual inference, and the neural coding models proposed to implement them, tend to be over single feature dimensions that are implicitly conditioned on a single object – i.e. where $p(\theta)$ is interpreted as a distribution over the orientation, θ, *of object* x (see Figure 2.2). Without this restriction, a multi-modal distribution over a particular dimension cannot distinguish multiplicity and uncertainty – two peaks in $p(\theta)$ could correspond to two orientations being simultaneously present, or to uncertainty about which of the two was present (see Figure 2.3). In Section 2.2.2 of the literature review, we discussed a new probabilistic notation derived from Sahani and Dayan (2003) that deals with this problem – using distributions over 'multiplicity functions' $p(m(\theta))$. The multiplicity function, $m(\theta)$, describes a possible state of the world – where multiple peaks explicitly correspond to multiple objects – whilst the distribution over all possible multiplicity functions represents uncertainty over these states and their values (see page 41 and Sahani and Dayan (2003) for further details).

This notation allows us to describe uncertainty and multiplicity for a single feature dimension at a single location. But this still falls short of the real-world scenario, where there are multiple different, spatially-distributed features. We thus propose an extension to the multiplicity function notation, in which each multiplicity function is over a single feature and its spatial location: $m^k\left(\mathbf{x}, \theta^k\right)$, where k indexes different features[1]. A posterior belief distribution over a scene decomposed into K feature dimensions would therefore be over K multiplicity functions, each expressing a possible spatial distribution of a particular

[1] We use a vector for space, $\mathbf{x}$, which could correspond to any number of spatial dimensions.

feature:

$$p\left(\left\{m^k\left(\mathbf{x},\theta^k\right)\right\}_{k=1}^{K}\right) \tag{6.1}$$

The question of how a 'feature' is defined is a thorny one – in the framework presented here, we simply think of features as dimensions of a visual scene that cortical neurons respond to. In the simulations reported below, we work with two abstract feature dimensions that can be characterised along a single continuous axis, and use 2D Gaussian 'tuning curves' over feature value and spatial location. However, in the visual cortex there are clearly many features that cannot be characterised in this way, and delineating complex spatio-temporal receptive fields is an important methodological and theoretical challenge (see e.g. Ahrens et al., 2008). There are also important questions about how increasingly complex representations are 'built' out of lower-level 'features', and how higher-level representations of structural features can be distinguished from the influence of environmental modulators such as lighting direction. This is an important area for future work, but the principles we demonstrate here should apply to more complex neural 'features', albeit necessitating more sophisticated probabilistic and neural coding models.

Having multiplicity functions defined over the spatial location of particular feature values has important representational, computational, and biological properties. These functions roughly correspond to the 'feature maps' invoked in computational models of attention, which provide a compact and flexible representation of a complex scene. Pairing features with space is important because the two are intimately connected – all features are spatially distributed, and it is unclear what a generic spatial multiplicity function $m(\mathbf{x})$ would say about the location of features represented in separate, featural multiplicity functions $m^k(\theta^k)$. Restricting the multiplicity functions to a *single* feature paired with space allows us to avoid the question of how we would define a common magnitude for multiple features – the magnitude of the multiplicity function reflects the 'strength' of the feature and therefore has variable physical interpretations, such as polarity contrast for orientation and opponency-channel value for colour. The probability distribution over multiple such multiplicity functions can then represent correlations between different features and their locations. This notation also corresponds to observed properties of cortical neurons – at the simplest level, they represent something about the spatial location of at least one feature. Populations that respond to more than one feature can be described as representing a distribution over several paired multiplicity functions (see Equation 6.1), and various levels of uncertainty over location and feature value can be represented by changing the form of the distribution. Sahani and Dayan (2003) present a model of how populations of neurons could encode these 'doubly distributional' representations, which future work could extend to the generalised feature-map case presented here.

6.2.1 Formalising the Resource Limitation

Using this notation we can now formalise the core proposal of the framework: that a general computational capacity limitation is the ability of the brain to represent posterior beliefs over the multitude of correlated features present in a natural scene, and that this forces the brain to approximate the true posterior:

$$q\left(\left\{m^k\left(\mathbf{x},\theta^k\right)\right\}_{k=1}^{K}\right) \sim p\left(\left\{m^k\left(\mathbf{x},\theta^k\right)\right\}_{k=1}^{K} \mid \mathbf{s}\right), \tag{6.2}$$

Here, $q(\cdot)$ is an approximation to the normalised true posterior, $p(\cdot)$. There are two features of the approximation that need to be considered – first, what form it takes and how it differs from $p(\cdot)$, and second, how it is computed. In machine learning it is common to approximate complex posteriors with factored approximations. This involves splitting the posterior up into a product of smaller distributions, each over a subset of features, which essentially pretends that the subsets are independent and therefore neglects the correlations between them (see e.g. Mackay, 2004). Another attractive feature of this approximation is that factored 'portions' of the distribution can be thought of as corresponding to functionally specialised neural areas or limited receptive fields. In what follows we will therefore use a factored approximation to demonstrate the effects of attention, which also provides an intuitive illustration of the effects of neglecting correlations. However, it is important to note that there are other possible approximations, and that the proposed effects of attention would operate similarly in other such cases. Equation 6.3 presents a simple, fully factored approximation that treats each of K multiplicity functions as independent;

$$q\left(\left\{m^k\left(\mathbf{x},\theta^k\right)\right\}_{k=1}^{K}\right) = \prod_{k=1}^{K} q\left(m^k\left(\mathbf{x},\theta^k\right)\right). \tag{6.3}$$

Neglecting both negative and positive correlations from the true posterior, as described above, makes it more likely that an inappropriate explanation of the data, $\mathbf{s}$, will be selected.

To compute an approximation that gives the best match to the true posterior, it is typical to minimise (within constraints) a distance measure between them. Here we use the Kullback-Leibler (KL) divergence $\mathbf{KL}\left[p(\cdot)\|q(\cdot)\right]$, which results in the approximation covering as much of the true distribution as possible, rather than approximating it more finely within a limited region (Minka, 2005). This seems intuitively sensible for a brain that

needs to be able to respond to stimuli in all parts of the world, and below we will describe how this interacts with the narrower focus of attention.

The unconstrained minimum of $\mathbf{KL}\left[p(\cdot)\|q(\cdot)\right]$ is achieved when $q(\cdot) = p(\cdot)$. Thus, $q(\cdot)$ is only an approximation because of constraints that prevent complete minimisation. One constraint is structural: $p(\cdot)$ will generally not fall into the class described by Equation 6.3 (in the brain, perhaps relating to implementational 'factorisation' over specialised regions and receptive fields), and so approximation $q(\cdot)$ cannot reach the true minimum. A further constraint is algorithmic. In general, when a distribution is intractable, the minimum-divergence approximation to it is also intractable. Again, appealing to machine learning, a family of algorithms have been developed to approach the minimum by instead minimising local versions of the KL divergence. These algorithms include belief propagation (BP) and expectation propagation (EP), and are reviewed by Minka (2005). Recent work has speculated about how such algorithms might be implemented in neural popualtions (see e.g. Lee and Mumford, 2003), or alternatively, the brain might learn to compute an approximate recognition model during development (see e.g. Hinton et al., 1995). In all of these cases, the prior and likelihood are encoded implicitly in an inferential machinery that approximates the posterior without ever explicitly representing it. This is a crucial point, as representation of the true posterior is exactly the intractable step that we suggest is avoided by using an approximation.

To summarise, for an approximation that factors over featural multiplicity functions, $q(\cdot)$ approximates the true posterior (Equation 6.4), which can be broken down according to Bayes' rule into the normalised product of the prior and likelihood (Equation 6.5). The normalisation constant, Z, plays a crucial role in what follows, acting as a measure of the similarity of the distributions whose product it normalises;

$$\prod_{k=1}^{K} q\left(m^k\left(\mathbf{x},\theta^k\right)\right) \sim p\left(\left\{m^k\left(\mathbf{x},\theta^k\right)\right\}_{k=1}^{K} \mid \mathbf{s}\right) \tag{6.4}$$

$$\sim \frac{1}{Z}\, p\left(\mathbf{s}|\left\{m^k\left(\mathbf{x},\theta^k\right)\right\}_{k=1}^{K}\right)\; p\left(\left\{m^k\left(\mathbf{x},\theta^k\right)\right\}_{k=1}^{K}\right) \tag{6.5}$$

The approximation is then found by minimising (within algorithmic constraints) the KL divergence between the product of the prior and likelihood, and the factored distribution:

$$\prod_{k=1}^{K} q\left(m^k\left(\mathbf{x},\theta^k\right)\right) \;= \tag{6.6}$$
$$\underset{q(\cdot)}{\operatorname{argmin}}\ \mathbf{KL}\left[\frac{1}{Z}\, p\left(\mathbf{s}|\left\{m^k\left(\mathbf{x},\theta^k\right)\right\}_{k=1}^{K}\right)\, p\left(\left\{m^k\left(\mathbf{x},\theta^k\right)\right\}_{k=1}^{K}\right) \Big\| \prod_{k=1}^{K} q\left(m^k\left(\mathbf{x},\theta^k\right)\right)\right]$$

6.2.2 Formalising the Role of Attention

Armed with a description of the limited resource and its consequences, we can now address the central questions of attentional selection. What is the effect of allocating attention, and what determines where and how it is allocated? In the framework presented here, attentional mechanisms act to locally refine the approximate posterior by imposing parameterised local 'hypotheses' about the true distribution through feedback connections. These attentional hypotheses have the mathematical form of an added prior, $p_a(\cdot)$, and $q(\cdot)$ approximates the normalised *product* of the true posterior and these extra parameterised 'priors' (compare Equation 6.5 to Equation 6.7)[2]. The normalisation constant Z_a can now be thought of as measuring the similarity between the true posterior implicit in the brain's learned recognition machine, and the attentional hypothesis:

$$\prod_{k=1}^{K} q_a\left(m^k\left(\mathbf{x},\theta^k\right)\right) \sim \tag{6.7}$$
$$\frac{1}{Z_a}\, p\left(\mathbf{s}|\left\{m^k\left(\mathbf{x},\theta^k\right)\right\}_{k=1}^{K}\right) p\left(\left\{m^k\left(\mathbf{x},\theta^k\right)\right\}_{k=1}^{K}\right) p_a\left(\left\{m^k\left(\mathbf{x},\theta^k\right)\right\}_{k=1}^{K}; \mathbf{r}_a\right)$$

The attentional hypothesis $p_a(\cdot)$ is parameterised by a vector $\mathbf{r}_a$, which reflects the internal state (embodied in some aspect of neuronal activity such as firing rates) of the attention-directing systems of the brain. The approximate distribution is again achieved by minimising, within algorithmic constraints, a KL divergence, although now with respect to the true

[2]Note that $p_a(\cdot)$ is referred to both as an 'attentional hypothesis' and as an 'attentional prior'.

posterior *informed by the attentional hypothesis* (compare Equation 6.6 to Equation 6.8):

$$\underset{q(\cdot)}{\operatorname{argmin}} \mathbf{KL}\left[\frac{1}{Z_a}\, p\left(\mathbf{s}|\left\{m^k\left(\mathbf{x},\theta^k\right)\right\}_{k=1}^{K}\right)\; p\left(\left\{m^k\left(\mathbf{x},\theta^k\right)\right\}_{k=1}^{K}\right)\ldots\right.$$
$$\left.\ldots\; p_a\left(\left\{m^k\left(\mathbf{x},\theta^k\right)\right\}_{k=1}^{K};\mathbf{r}_a\right)\middle\|\prod_{k=1}^{K} q_a\left(m^k\left(\mathbf{x},\theta^k\right)\right)\right] \quad (6.8)$$

We could express the effect of $p_a(\cdot)$ as simply *modulating* the true prior – i.e. multiply the last two terms in Equation 6.7, but we leave the objects separate for clarity, and to indicate that the attentional signal is represented separately from the latent prior implicit in the brain's learned recognition machine.

As the attentional hypothesis is represented in the brain, it must of course be subject to the general resource limitation proposed above. Thus the attentional hypothesis cannot represent proposals about complex conjunctive relationships across the visual scene. Instead we suggest it consists of one, or possibly a small number, of local modes. This concurs with behavioural observations that attention can only encompass a limited region of space at a time, observations that have in the past contributed to the discrete 'bottleneck' and 'spotlight' metaphors discussed in the literature review (see Itti et al., 2005, and Section 2.4.1). A smeared local mode is also suited to transmission by the relatively coarse specificity of feedback connections (see Friston, 2003). The attentional hypothesis is assumed to be represented in fronto-parietal regions, by neurons with flexible representational capacity (Kastner and Ungerleider, 2000; Huddleston and DeYoe, 2008), though there is ongoing debate about the exact role of the constituent regions in driving attentional allocation (Green and McDonald, 2008; Sommer et al., 2008).

Behavioural (e.g. Rossi and Paradiso, 1995), electrophysiological (e.g. Treue and Martinez Trujillo, 1999), and neuroimaging (e.g. Saenz et al., 2002) data also support the existence of distinct modes of spatial and feature-based attention (see for a review Maunsell and Treue, 2006). In our framework, spatial and featural components of the attentional hypothesis would each be defined over 'feature-map' multiplicity functions of both space and feature value, but maximal uncertainty over the other dimension would effectively restrict them to the dimension of interest. The architecture of the visual system suggests that these signals would be processed in different ways; how different components of $p_a(\cdot)$ are combined and transmitted throughout the cortical hierarchy is an open experimental, as well as theoretical, question (see Section 6.4 for further discussion)

The attentional hypothesis – the location of the mode or 'spotlight' – can reflect genuine prior information, top-down instructions (including task demands), bottom-up cues, and salience computations. For example, a spatial cue that indicates the location of an upcoming stimulus would be reflected by a mode centered on the cue location in the spatial component of the attentional hypothesis. Depending on the source of the attentional hypothesis, it can be described as saying very different things – for example, if driven by salience computations it says 'this region is the most important or unusual', whereas if dictated by a search instruction it says 'I propose a stimulus is here'. However, the underlying mathematical form is the same, which is a particular strength of this framework – different forces act to bias the ongoing allocation of distributed and various instantiations of a unified, computational resource limitation.

In the literature review, we discussed the distinction between selection for action, and selection for representation, and suggested that there is not a clear dividing line between the two. Our framework does not, in its general form, implement limits on action, but $p_a(\cdot)$ can carry information that is relevant to acting in the world – it improves representation, but can do so in a task-relevant manner. This reflects the rough parallel that is often drawn between the idea that you have to select a single action, and the idea that you have to select a focus of attention (see e.g. Taylor et al., 2008). However, this does not mean that representations are geared only towards driving motor plans, and attentional 'selection' via $p_a(\cdot)$ can improve many more features of the representation than are required to describe the target of action.

In the absence of direct biasing signals, the attentional prior continuously evolves towards a better match between itself and the true posterior, locally improving the brain's representation of the world. As explained above, this match is measured by the size of the normalising constant Z_a of their product (i.e. the normalising constant of the distribution on the right hand side of Equation 6.7). This is a simple consequence of the fact that the normalising constant is the sum of the probabilities given to all possible values of the variable. The more similar the true posterior and attentional prior, the more likely it is that they will award high probabilities to the same values, increasing the sum of their product.

The normalisation constant is often referred to as the 'model evidence' when Bayes rule is used for model comparison (see Equation 1.2 and page 15) – corresponding here to the marginal probability of the evoked sensory firing; $p(\mathbf{s}|\mathbf{r}_a)$. The evolution of the attentional prior can thus be conceptualised as a process of optimisation, of continuous model comparison or hypothesis testing that moves $p_a(\cdot)$ in the direction of a better model of the world described by the true posterior. As laid out in Equation 6.8, the approximate posterior, $q(\cdot)$, is found by minimising the KL divergence between the product of the prior, likelihood, and attentional hypothesis, and the approximating distribution. Therefore, as the

attentional hypothesis evolves, the KL divergence will also evolve, and as the approximate posterior continuously works to minimise the KL it will reflect the attentional hypothesis and whatever influences on it are currently dominant. The simulations we present here do not explicitly implement the dynamic evolution of the attentional hypothesis, and important questions for future work include how Z_a can guide the evolution of $p_a(\cdot)$ whilst constraining its form, and can encourage implement sequential testing of local 'models' rather than broad-brush approximation.

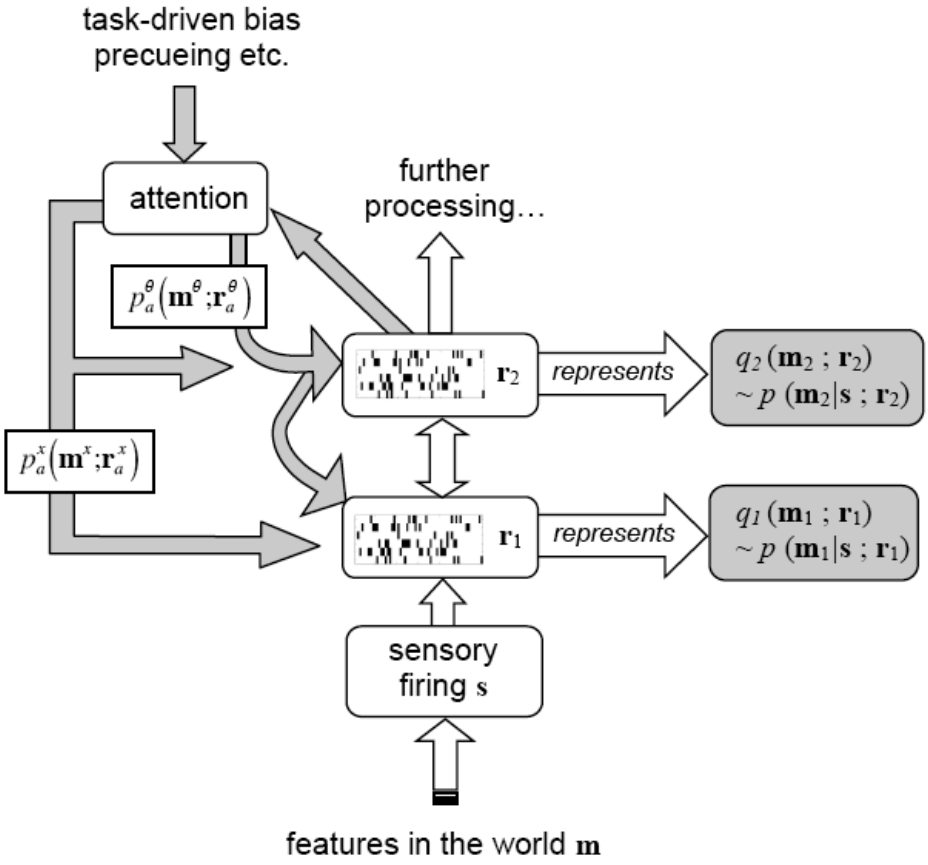

Figure 6.1: ***Simple schematic of attentional framework.*** *Simple schematic of a probabilistic framework for top-down attentional selection, illustrating a hierarchy of approximate representations in visual cortex, which are biased or informed by attentional hypotheses passed down from fronto-parietal regions via feedback connections. See text for details.*

Figure 6.1 presents a simple schematic of the framework we propose here – objects in the world $\mathbf{m}$ evoke firing at the sensory epithelia, $\mathbf{s}$. This is transmitted through a hierarchy of regions, each of which represents some subset of features at some scale of analysis; $\mathbf{m}^i$, in its firing rates; $\mathbf{r}^i$, but in approximate form; q_i. Attentional hypotheses; p_a, which are driven by a variety of sources, are represented in fronto-parietal regions of the brain and passed down via feedback connections to bias the approximation process – spatial signals

are transmitted more directly to each layer, whilst featural hypotheses are translated as they pass back down the hierarchy.

6.2.3 Formalising the Effects of Attention

So how does introducing the attentional 'hypothesis' improve perception? There are two main effects – the first extends previous models of attention as a Bayesian prior (Dayan and Zemel, 1999; Rao, 2005), but without the restricting semantics of having to carry prior information. This acts on probabilistic uncertainty whether or not the posterior is approximate, but the second, novel role we propose for attention is to selectively reveal correlations neglected in an approximation. The framework within which we define attention also extends the classic Bayesian setting of belief distributions over single features to distributions over multiplicity functions, enabling us to deal with complex, multi-object scenes (see Section 2.2.2). This allows us to consider cue-combination and 'binding' tasks without invoking separate distributions for each object as in SDT models. Below we will discuss the two roles of attention, and briefly consider how they relate to existing approaches. We will then demonstrate both using simulations in a simple model.

In classic precueing tasks, a flash of light or symbolic cue indicates the location of an upcoming stimulus, resulting in improved judgements about stimulus features (see e.g. Luck et al., 1996; Cameron et al., 2002). In such scenarios, the attentional signal can be thought of as carrying prior information about the upcoming scene, correctly indicating the high probability of a particular stimulus location. Dayan and Zemel (1999) modelled this effect, considering computation in two cortical layers. In their model, a lower layer with smaller RFs (corresponding roughly to $V1$) represents a joint distribution over space and orientation. A higher layer with larger RFs (corresponding roughly to $V4$) then integrates out space from this joint distribution to yield a marginal distribution over orientation;

$$p\left(\theta|\mathbf{r}\right) = \int dx\; p(\theta, x|\mathbf{r}) \tag{6.9}$$

Equation 6.10 breaks down the joint posterior according to Bayes rule, revealing the prior over space that is then altered by attention, indicated by $p^*(x)$. This allows more of the probability mass to be allocated to the region of interest, and if $p^*(x)$ sets the probability of any location outside the 'focus of attention' to *zero*, this effectively reduces the limits of

integration (see Equation 6.11):

$$p_a(\theta|\mathbf{r}) \sim \int_{-\infty}^{\infty} dx \, \frac{p(\mathbf{r}|\theta, x)\, p(\theta)\, p^*(x)}{p(\mathbf{r})} \tag{6.10}$$

$$\sim \int_{\alpha}^{\beta} dx \; p(\theta, x|\mathbf{r}) \tag{6.11}$$

Attention thus yields an improved marginal distribution over orientation, on the basis of which improved judgements can be made. In the work of Dayan and Zemel (1999), attention affects the transfer of information from one cortical layer to another, modulating inference in the layer above. In our framework, it has such an effect within a single posterior representation. Integration to yield marginal distributions can then be thought of as occuring in the same layer, at a subsequent stage of visual processing, or even in decision-making regions of the brain – the principle is the same. We also extend the effect to distributions over multiplicity functions, where we can encompass improved judgements about objects within complex scenes, rather than simply excluding 'noise' with regard to a single object.

In response-cueing tasks two or more stimuli always appear, and the cue indicates which one is relevant (i.e. to which a response will be required; see e.g. Pestilli et al., 2007), rather than indicating where a single relevant stimulus will appear. In this case the cue cannot strictly be called a prior over the visual image, as it carries no information about its spatial structure (c.f. Rao, 2005). However, it seems intuitively clear that a similar mechanism might be responsible. In our framework both effects are due to an attentional hypothesis that has the mathematical *form* of a prior without its attendant, restrictive, semantics.

This picture is closely related to SDT approaches (see page 21), in which attention reduces uncertainty by ruling out irrelevant distractors or empty locations (e.g. Gould et al., 2007; Dosher and Lu, 2000). Proponents of signal enhancement have argued that this downplays the capabilities of selective attention, which can directly enhance the 'signal' as well as reducing 'noise' (e.g. Cameron et al., 2002). Our approach is clearly sympathetic to the noise reduction hypothesis, and has no explicit notion of isolated signal enhancement. As demonstrated in the Dayan and Zemel (1999) model described above, when space and feature value are *jointly represented*, spatial uncertainty reduction leads to improved feature judgements. However, this is basically a further, indirect form of uncertainty reduction – if the attentional hypothesis is a good one, neglecting information outside its spatial focus will also down-weight irrelevant *feature* information.

So what of the evidence for 'pure' signal enhancement? The behavioural evidence largely consists of showing that attentional benefits are observed even when there is no apparent

uncertainty – e.g. when there are no distractors, or where spatial uncertainty is minimal (see Cameron et al., 2002). However, the experimental debate still rages (see for example Gould et al., 2007; Dosher and Lu, 2000), and from a probabilistic perspective, it is hard to conclude that *no* uncertainty is present. For example, Cameron et al. (2002) showed that when a spatial precue indicated the upcoming location of a high contrast stimulus, observers could make close-to-perfect *localisation* judgements even without attention, whilst their imperfect *orientation* judgements showed a benefit of the cue. But if 'perfect' localisation judgements are discretised, probabilistic location uncertainty could still be present, and its reduction sufficient to improve a marginal distribution over orientation. Thus, when spatial uncertainty exists but is not evident in behaviour, its silent reduction could support improved feature judgements that appear to represent pure signal enhancement.

Another line of argument for signal enhancement consists of looking for its signature in the effects of attention on behavioural (Ling and Carrasco, 2006) and electrophysiological (Williford and Maunsell, 2006; Li et al., 2008) contrast-response (C-R) functions. This debate has again reached somewhat of a stalemate, perhaps in part due to the absence of algorithmic models that convincingly link behavioural and neural data in the context of well-defined notions of noise and signal enhancement. Lu and Dosher (1998) developed an SDT model that explicitly separated different sources of noise, and defined signal enhancement in mechanistic terms. Interestingly, they found that reducing additive internal noise had the same effect as boosting the signal, which highlights the importance of carefully defining the terms of the debate. The limitations of this model come from the nature of an SDT representation, in which each object is represented by a separate distribution.

In our framework, the locations and feature values of multiple objects are represented in the same multiplicity function, which provides a richer substrate for future investigations into this question. This richer representation also makes the effects of attention less intuitive though – $p_a(\cdot)$ weights a region of a feature map, rather than weighting particular values a single object might take. The basic effects of attention as prior are still observed – when $p_a(\cdot)$ is a good match to the true posterior, spatial and featural judgements should be improved within the focus of attention. How this could be extended to shed further light on the potentially complex interrelationship between uncertainty reduction and enhanced representations in a probabilistic representation is an intriguing question for future work.

As well as extending the idea of attention as a Bayesian prior, we propose a novel role for the attentional hypothesis – to selectively reveal correlations neglected in an approximate (perhaps factored) distribution. This occurs as the attentional hypothesis evolves towards a closer match to the true posterior, and corresponds to the sequential allocation of attention during viewing of natural scenes or artificial visual search tasks. This evolution is guided by the size of the normalisation constant of the product of the true posterior and attentional

hypothesis, Z_a, and can thus be thought of as a process of continuous model comparison – by sequentially testing possible explanations against the true posterior those favoured by the correlational structure are more likely to be selected and then reflected in the approximation.

Negative correlations are often induced by 'explaining away' effects in the posterior – where more than one explanation could underlie the observed data (see page 149 above). For example, the apparent size of an object is determined by both its actual size and its distance from the viewer – to have the same apparent size, a smaller object must be closer and a larger object further away. The range of likely sizes and distances is set jointly by our prior expectations, information that would be lost if the perceptual system represented distance and size separately. However, the moment we know the size we can derive the distance, and vice versa – conditioned on the object being close, there is no longer any anti-correlation to resolve. As the attentional hypothesis explores the possibilities of close and far, it implicitly satisfies the anti-correlation without having to explicitly represent it[3]. Furthermore, for each attentional setting, the system can evaluate the evidence Z_a for the hypothesis $p_a(\cdot)$, which can be used to help $p_a(\cdot)$ converge on a good explanation of the visual image (see Section 6.4 for a discussion of how this might be implemented).

An analogous situation applies when features are *positively* correlated. Features in the world tend to be colocated, and spatially extended – in other words, they tend to make up objects – and may also express generic associations between different feature values. These correlations are expressed in the true prior (see page 149), and might be neglected in the approximation. For example, in a visual search task observers must locate a conjunction of two different features. If those features are represented separately in the perceptual system, information about their colocation would be lost. The attentional hypothesis travels the scene, and when it settles on a stimulus location the observer is more likely to report features belonging to that stimulus as 'bound', and less likely to report illusory conjunctions (see Treisman and Schmidt, 1982, Section 2.4.3, and Section 6.3.3). This corresponds to the binding role for attention described in Feature Integration Theory (FIT) (see Treisman and Gelade, 1980), but in a continuous form that naturally allows various degrees of preattentive 'bundling' (see Wolfe et al., 1989). If the observer is looking for a particular stimulus, the evolution of the attentional hypothesis can be thresholded by a relevant quantity – for example, if an observer is searching for a red vertical bar, $p_a(\cdot)$ might continue evolving until the joint probability of 'red' and 'vertical' exceeds some threshold. It is also important to note that the noise reduction effects discussed above operate alongside the dynamic

[3]This approach of setting one variable within a complicated joint distribution to a series of known values and recomputing the distribution over the other variables is similar to a probabilistic inference algorithm called 'cutset conditioning' (Pearl, 1988b).

evolution of the attentional hypothesis. When $p_a(\cdot)$ is a good match to the true posterior, this can further improve performance on location and feature judgements.

Another important feature of our framework is that it elucidates the relationship between the resource limitation in visual cortex, and the resource limitation in the attentional hypothesis. Both arise from the same generic representational constraint, but play very different roles. Visual cortex is prevented from representing the full joint posterior over the state of the world, and attempts to represent as much of the posterior as possible in its approximation – $\mathbf{KL}\left[p(\cdot)|q(\cdot)\right]$. The attentional hypothesis is also unable to represent large, complex posteriors, but doesn't attempt to do so – it consists of simple, local hypotheses that serve to bias the ongoing approximation of the full posterior (see page 155). Combining a broad approximation with a narrow, focused modulation reflects the principle that the brain should process as much as possible to a rough level, as this can govern rapid responses and also guide the allocation of fine-grained further processing. This 'rough processing' has been described as a parallel sweep, as a salience computation, and in our framework provides a comprehensive basis from which the dynamic evolution of the attentional hypothesis can reflect the true posterior as well as possible. Locating this concept on the computational level allows us to encompass the evidence for a wide variety of loci of both parallel processing and fine-grained attentional effects. A probabilistic characterisation admits of degrees rather than distinctions – approximations are better or worse, improved or degraded, rather than being discretely parallel or serial, bound or unbound.

6.3 Simulating Key Attentional Phenomena

In the next section we will demonstrate these two key effects of attention in a simple, abstract domain. We will first look at scenarios in which attention acts as a prior, performing noise reduction and improving stimulus judgements. This effect will be demonstrated for simple analogues of a precueing task, and a response-cueing task, illustrating the generality of an attentional hypothesis with the mathematical form but not the attendant semantics of a Bayesian prior. We will then model a simple illusory conjunction task in which attention improves binding judgements, simulating the effects of the dynamic 'hypothesis-testing' process by revealing correlations neglected in the approximate posterior (see page 160).

6.3.1 Setting up the Model

The simulations take place in a simple 'grid world' consisting of a single, discretised spatial dimension ($\mathbf{x}$) and two discretised feature dimensions – corresponding to colour (o) and orientation (c). Figure 6.2 illustrates a single state of the grid world, and what the two multiplicity functions or 'feature maps' that describe the grid world would look like for this state. Four spatial locations, labelled x_1, x_2, x_3, and x_4 each take one value of colour and orientation, indicated by the grey entries in the orientation feature map; $m^o(\mathbf{x}, o)$, and colour feature map; $m^c(\mathbf{x}, c)$. Greyscale values indicate the strength of the feature, so for orientation this corresponds to contrast and for colour this corresponds to luminance. In Figure 6.2 all contrasts and luminances are equal, as indicated by the shared grey level of the 'on' entries in the feature maps.

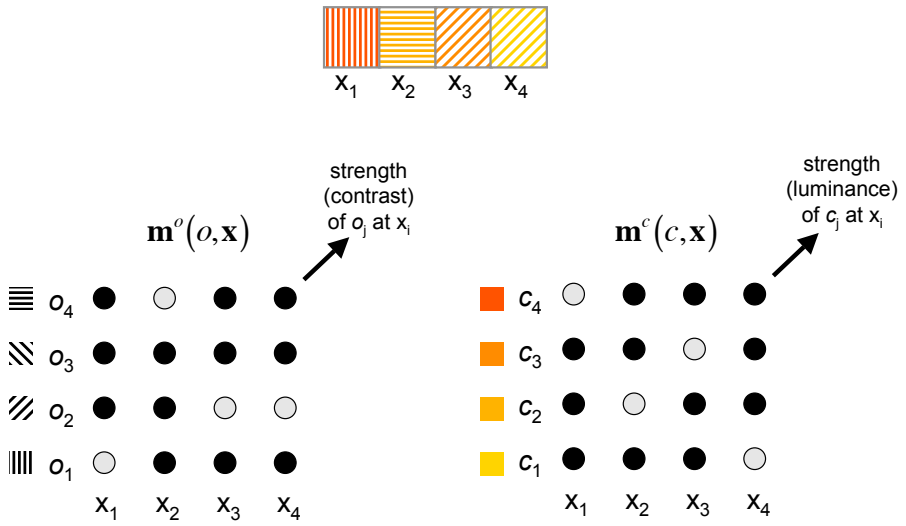

Figure 6.2: ***'Grid world' setting for simulations.*** *Illustrates the simple 'grid world' in which the simulations take place – each discretised spatial location can take values on a discretised orientation and colour dimension. Two multiplicity functions or feature maps describe the exemplar state of the world shown, and the magnitude of the entries in the maps (indicated by the greyscale values) indicate the 'strength' of the corresponding features along an appropriate dimension.*

Below we describe a simple mathematical model for how states of the grid world produce noisy observations (corresponding to neural firing at the sensory epithelia), and how inverse inference works back from these observations to an approximate posterior belief about the state of the world that caused them. We then show how introducing attentional hypotheses corresponding to particular paradigms affects the approximate posterior, and how simulated judgements made from this modulated posterior mimic the behavioural effects observed in these paradigms.

Figure 6.3 illustrates the generation of noisy observations in the model. The prior over objects is sparse, so that only a small number of objects – and therefore features – will be present at any one time. The locations, orientations, and colours of the objects present are encoded in a binary vector $\mathbf{u}$. Each '1' element in this vector corresponds to an object with a particular location and pair of feature values, as depicted in the 3D grid at the left of the figure – the vector is formed by a scan-rasterised version of this 3D grid that essentially 'unwraps' it. Two rectangular projection matrcies, P^c and P^o, can then be used to obtain two 2D representations, one for each feature type, as shown in the next panel of the figure. The corresponding rasterised vectors indicate the spatial locations of non-zero colour values ($\mathbf{u}^c$) and orientation values ($\mathbf{u}^o$):

$$\begin{aligned} \mathbf{u}^c &= \mathrm{P}^c\mathbf{u} \\ \mathbf{u}^o &= \mathrm{P}^o\mathbf{u} \end{aligned} \tag{6.12}$$

The binary vectors $\mathbf{u}^c$ and $\mathbf{u}^o$ thus indicate the *location* of the peaks in the multiplicity functions. The *amplitudes* of these peaks – that is, the strength of the corresponding features (see page 43) – are each drawn independently from a zero-mean Gaussian distribution. Equivalently, we can view the multiplicity functions themselves[4] as drawn from zero-mean multivariate normal distributions with diagonal covariances U^c and U^o, whose diagonal elements are given by the vectors $\mathbf{u}^c$ and $\mathbf{u}^o$:

$$\begin{aligned} \mathbf{m}^c &\sim \mathcal{N}[\mathbf{0}; \mathrm{U}^c] \quad \textit{where} \quad \mathrm{U}^c = \mathrm{diag}[\mathbf{u}^c] \\ \mathbf{m}^o &\sim \mathcal{N}[\mathbf{0}; \mathrm{U}^o] \quad \textit{where} \quad \mathrm{U}^o = \mathrm{diag}[\mathbf{u}^o] \end{aligned} \tag{6.13}$$

[4] In this discretised model, multiplicity functions have of course become multiplicity vectors, but we continue to use the same terminology as the effects of interest are the same.

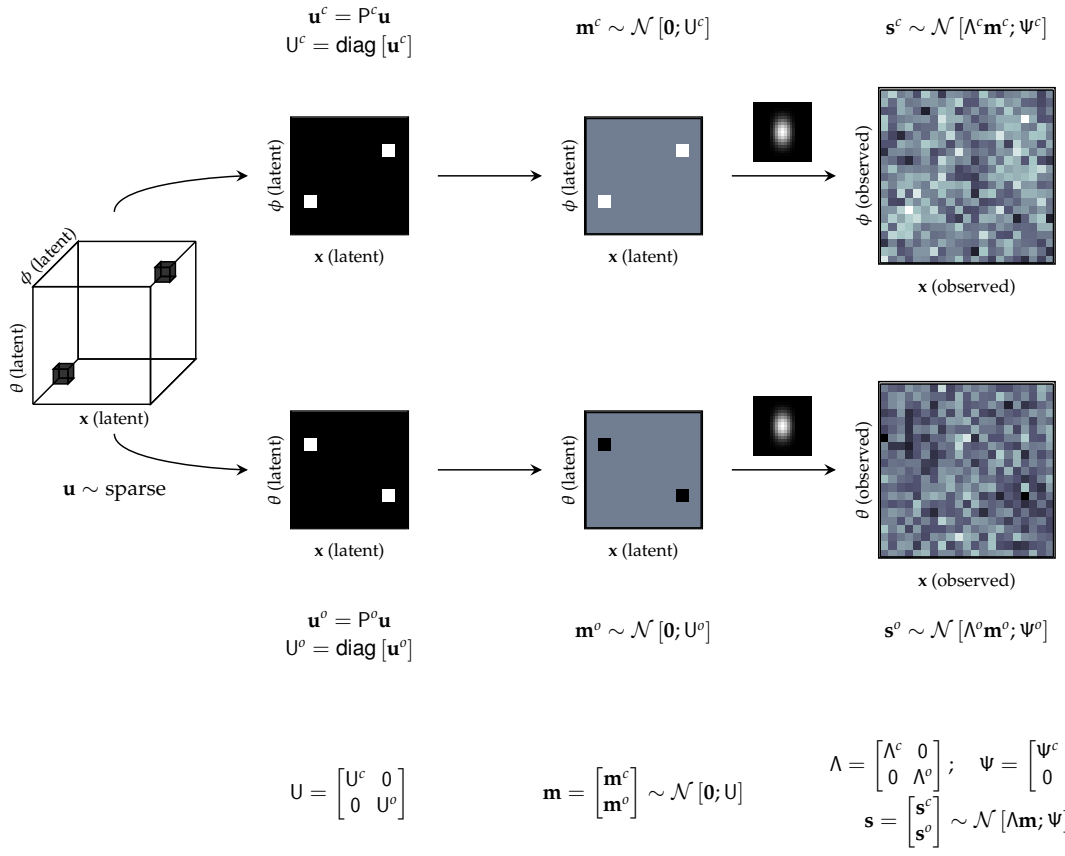

Figure 6.3: ***The generative model of 'grid world'*** *This schematic illustrates the generation of noisy observations,* $\mathbf{s}$*, from an underlying state of the world,* $\mathbf{u}$*, in the simple grid world in which simulations take place. The white squares in* $\mathbf{u}^c$ *and* $\mathbf{u}^o$ *indicate which feature-location pairs are 'on', and the strength of these features (for example, contrast for orientation and luminance for colour) is generated from a zero-mean Gaussian distribution, producing feature maps (multiplicity functions)* $\mathbf{m}^c$ *and* $\mathbf{m}^o$*. In these feature maps, an intermediate, grey colour indicates zero strength, and black and white indicate the extremes of an opponent axis. To generate noisy observations,* $\mathbf{m}^c$ *and* $\mathbf{m}^o$ *are passed through a Gaussian weight matrix,* Λ*, which smears out the feature map entries. One component of* Λ *is illustrated above the arrows that join* $\mathbf{m}^k$ *to* $\mathbf{s}^k$*, and the matrix consists of one such component centered on each location-feature pair. Independent Gaussian noise,* Ψ*, then corrupts the observations.*

Thus, where there is no peak in the feature map – i.e. a '0' entry in $\mathbf{u}^c$ or $\mathbf{u}^o$ – the value of the multiplicity function is zero, as we are generating from a zero mean, zero covariance Gaussian. However, when there is a peak in the feature map – i.e. a '1' entry in $\mathbf{u}^c$ or $\mathbf{u}^o$ – we generate from a zero mean Gaussian with a covariance of 1 and therefore generate a range of feature strengths. Note that because the Gaussian is zero mean, high feature strengths can be represented by high positive *or* high negative values. This concurs with neurally inspired representations of features in terms of a pair of opposing axes – for example a blue-yellow axis for colour or a positive-negative polarity axis for orientation contrast.

The third panel in Figure 6.3 shows an example of feature maps drawn from this distribution, $\mathbf{m}^o$ and $\mathbf{m}^c$, where grey corresponds to the absence of a feature. Finally, noisy observations $\mathbf{s}$ are generated by passing the multiplicity functions though a gaussian weight matrix Λ, which smears out the discrete entries. Λ consists of the combination of receptive-field-like weight matrices centered on each feature-location pair – a single component is shown above the arrows that join $\mathbf{m}^k$ to $\mathbf{s}^k$ in Figure 6.3. Gaussian noise with diagonal covariance, Ψ, is then added to the smeared observations:

$$\begin{aligned} \mathbf{s}^c &\sim \mathcal{N}\left[\Lambda^c \mathbf{m}^c; \Psi^c\right] \\ \mathbf{s}^o &\sim \mathcal{N}\left[\Lambda^o \mathbf{m}^o; \Psi^o\right] \end{aligned} \tag{6.14}$$

The observation space is taken to be higher dimensional (here represented by a finer discretisation) than the feature-map space, and an example of observations generated by this process is shown in the rightmost panel of Figure 6.3. This transformation from a state of the world, $\mathbf{m}$, to noisy observations, $\mathbf{s}$, can be thought of as representing all the sources of stochasticity that render perceptual inference necessarily probabilistic – including noise in the external world, unreliable neural firing, and coarse response properties (see Section 2.4).

The model can be written more compactly by concatenating the two feature dimensions. First, multiplicity functions are generated from binary feature-location vectors through zero mean Gaussians:

$$\mathbf{m} = \begin{bmatrix} \mathbf{m}^c \\ \mathbf{m}^o \end{bmatrix} \sim \mathcal{N}\left[\mathbf{0}; \mathrm{U}\right] \quad \textit{where} \quad \mathrm{U} = \begin{bmatrix} \mathrm{U}^c & 0 \\ 0 & \mathrm{U}^o \end{bmatrix} \tag{6.15}$$

Second, observations are generated from the multiplicity functions, which are passed through a weight matrix to form the mean of a Gaussian with diagonal noise:

$$\mathbf{s} = \begin{bmatrix} \mathbf{s}^c \\ \mathbf{s}^o \end{bmatrix} \sim \mathcal{N}\left[\Lambda\mathbf{m}; \Psi\right] \quad \textit{where} \quad \Lambda = \begin{bmatrix} \Lambda^c & 0 \\ 0 & \Lambda^o \end{bmatrix} \quad , \quad \Psi = \begin{bmatrix} \Psi^c & 0 \\ 0 & \Psi^o \end{bmatrix} \tag{6.16}$$

These equations represent a generative model for noisy feature observations, expressed as a prior on feature maps; $\int d\mathbf{u}\, p\left(\mathbf{m}|\mathbf{u}\right)\, p_0\left(\mathbf{u}\right)$, where p_0 is the sparse prior; and a likelihood $p\left(\mathbf{s}|\mathbf{m}\right)$. Perceptual inference involves inverting the generative model to compute a posterior belief about the true feature map given the noisy observations it generated:

$$\begin{aligned} p(\mathbf{m}|\mathbf{s}) &\propto p(\mathbf{s}|\mathbf{m})\, p(\mathbf{m}) & (6.17) \\ &\propto p(\mathbf{s}|\mathbf{m}) \int d\mathbf{u}\, p\left(\mathbf{m}|\mathbf{u}\right)\, p_0\left(\mathbf{u}\right) & (6.18) \end{aligned}$$

Here we come to the problem that lies at the heart of our framework. We need to integrate over $\mathbf{u}$ (or sum over its values in a discretised settings), but the prior embodies knowledge about the sparse distribution of objects made up of spatially colocated features, and is thus highly correlated. In the terms of the model, the prior $p_0\left(\mathbf{u}\right)$ consists of something like a mixture of sparse distributions for different numbers of objects, and therefore has a complex correlation matrix, U. If the binary vector $\mathbf{u}$ is n entries long it has 2^n possible settings – summing over each of these is intractable. Of course, if each element in $\mathbf{u}$ were independent, the sum in Equation 6.18 would involve only n operations.

Below we derive an approximation to the posterior that ignores correlations in the prior, exploiting the Gaussian properties of the generative model and approximation. We start by noting that conditioned on $\mathbf{u}$, $\mathbf{m}$ and $\mathbf{s}$ are jointly Gaussian with zero mean:

$$\begin{aligned} p\left(\begin{bmatrix} \mathbf{m} \\ \mathbf{s} \end{bmatrix} \middle| \mathbf{u}\right) &= p\left(\mathbf{s}|\mathbf{m}\right)\; p\left(\mathbf{m}|\mathbf{u}\right) \\ &= \mathcal{N}\left[\Lambda\mathbf{m}; \Psi\right] \;\times\; \mathcal{N}\left[\mathbf{0}; \mathrm{U}\right] \\ &= \mathcal{N}\left[\mathbf{0}; \begin{bmatrix} \mathrm{U} & \mathrm{U}\Lambda' \\ \Lambda\mathrm{U} & \Lambda\mathrm{U}\Lambda' + \Psi \end{bmatrix}\right] \end{aligned} \tag{6.19}$$

To move from here to the posterior we next need to integrate out $\mathbf{u}$, which yields a mixture of zero-mean Gaussians, one for each possible setting:

$$p\left(\begin{bmatrix}\mathbf{m}\\ \mathbf{s}\end{bmatrix}\right) = \sum_{\mathbf{u}} p_0(\mathbf{u})\, p\left(\begin{bmatrix}\mathbf{m}\\ \mathbf{s}\end{bmatrix}\Big|\mathbf{u}\right) \tag{6.20}$$

However, as explained above, this step is intractable. We therefore approximate the joint posterior by minimising the KL divergence between the true joint and a Gaussian approximation, $q(\cdot)$. This optimal approximating distribution is also zero mean, and simply consists of replacing the covariance matrix U in the conditional distribution (Equation 6.19) with its average under the prior, $\overline{\mathrm{U}}_0$:

$$\begin{aligned} q\left(\begin{bmatrix}\mathbf{m}\\ \mathbf{s}\end{bmatrix}\right) &= \underset{q}{\operatorname{argmin}}(\cdot)\ \mathbf{KL}\left[\sum_{\mathbf{u}} p_0(\mathbf{u})\ \mathcal{N}\left[\mathbf{0};\begin{bmatrix}\mathrm{U} & \mathrm{U}\Lambda'\\ \Lambda\mathrm{U} & \Lambda\mathrm{U}\Lambda'+\Psi\end{bmatrix}\right]\ \Big\|\ q\left(\begin{bmatrix}\mathbf{m}\\ \mathbf{s}\end{bmatrix}\right)\sim\mathcal{N}\right] \\ &= \mathcal{N}\left[\mathbf{0};\begin{bmatrix}\overline{\mathrm{U}}_0 & \overline{\mathrm{U}}_0\Lambda'\\ \Lambda\overline{\mathrm{U}}_0 & \Lambda\overline{\mathrm{U}}_0\Lambda'+\Psi\end{bmatrix}\right] \end{aligned} \tag{6.21}$$

The covariance matrix U is diagonal, with '1' entries on the diagonal indicating the presence of a particular feature-location combination. The *average* of this quantity under the prior will be a number lying between 0 and 1, and is essentially the marginal prior probability of getting that feature-location combination (i.e. the probability of the relevant entry on the diagonal of $\overline{\mathrm{U}}_0$ being 'on'). From Equation 6.21, it is then straightforward to derive the two quantities we want for perceptual inference – the approximate posterior belief distribution (Equation 6.22), and the normalising constant (Equation 6.23):

$$q_0\left(\mathbf{m}|\mathbf{s}\right) = \mathcal{N}\left[\overline{\mathrm{U}}_0\Lambda'\left(\Lambda\overline{\mathrm{U}}_0\Lambda'+\Psi\right)^{-1}\mathbf{s};\overline{\mathrm{U}}_0-\overline{\mathrm{U}}_0\Lambda'\left(\Lambda\overline{\mathrm{U}}_0\Lambda'+\Psi\right)^{-1}\Lambda\overline{\mathrm{U}}_0\right] \tag{6.22}$$

$$\log Z_0 = -\frac{1}{2}\left[\log\left|2\pi\left(\Lambda\overline{\mathrm{U}}_0\Lambda'+\Psi\right)\right|+\mathbf{s}'\left(\Lambda\overline{\mathrm{U}}_0\Lambda'+\Psi\right)^{-1}\mathbf{s}\right] \tag{6.23}$$

In the equations used to lay out the framework in Section 6.2, the true posterior was approximated by a fully factored distribution (see Equation 6.4). In the model described here, we do not explicitly ask for $q(\cdot)$ to be factored (and indeed, many other approximations are possible) – we rather stipulate that it be Gaussian (see Equation 6.21). This yields a simple analytic solution to the KL minimisation (replacing U with $\overline{\mathrm{U}}_0$), and the resulting approximate posterior is still 'factored' in that it ignores correlations in the true prior.

The true prior carries both positive and negative correlations, expressing knowledge about the sparse distribution of objects in the world, about typical associations between feature values, and about the colocation of features in spatially extended objects. Ignoring these correlations results in the approximate posterior being completely factored over the two multiplicity functions, $\mathbf{m}^c$ and $\mathbf{m}^o$, which renders it unable to express knowledge about the colocation of the two features. Correlations between the elements of each multiplicity functions that come from the prior are also ignored, but those that come from the generative process (i.e. from Λ) are still present. This is a natural form of approximation for this model, and is sufficient to demonstrate the two effects of attention described above. Along with the effect of forcing the approximation to be Gaussian, treating the prior as independent serves as a proxy for the neural capacity limit we discussed above (see page 66).

The final component of the model is the attentional hypothesis, which acts to locally refine the impoverished representation of the true posterior. The attentional hypothesis has the form of a prior over $\mathbf{u}$, $p_a(\mathbf{u})$. In this model approximating the product of the true distribution and the attentional hypothesis is equivalent to replacing the average covariance of $\mathbf{u}$ under the prior; $\overline{\mathrm{U}}_0$, with a modified version that expresses the attentional hypothesis; $\overline{\mathrm{U}}_a$ (compare Equations 6.22 and 6.23 to Equations 6.24 and 6.25)[5]:

$$q_a(\mathbf{m}|\mathbf{s}) = \mathcal{N}\left[\overline{\mathrm{U}}_a\Lambda'\left(\Lambda\overline{\mathrm{U}}_a\Lambda'+\Psi\right)^{-1}\mathbf{s};\overline{\mathrm{U}}_a-\overline{\mathrm{U}}_a\Lambda'\left(\Lambda\overline{\mathrm{U}}_a\Lambda'+\Psi\right)^{-1}\Lambda\overline{\mathrm{U}}_a\right] \tag{6.24}$$

$$\log Z_a = -\frac{1}{2}\left[\log\left|2\pi\left(\Lambda\overline{\mathrm{U}}_a\Lambda'+\Psi\right)\right|+\mathbf{s}'\left(\Lambda\overline{\mathrm{U}}_a\Lambda'+\Psi\right)^{-1}\mathbf{s}\right] \tag{6.25}$$

6.3.2 Precueing and Response-Cueing

As discussed in Section 6.2.3, there are two main effects of the attentional hypothesis. The first extends existing notions of attention as Bayesian prior (see Dayan and Zemel, 1999; Rao, 2005), and enriches signal detection theory approaches to attention as uncertainty reduction (e.g. Palmer et al., 1993; Gould et al., 2007; Dosher and Lu, 2000). Below we reproduce the qualitative pattern of results in a simple analogue of a precueing task, where the attentional hypothesis carries genuine 'prior' information (after Cameron et al., 2002; Luck et al., 1996), and in a simple analogue of a task in which the attentional hypothesis instead represents an instruction about relevance (after Pestilli et al., 2007). By configuring

[5]Here, the true prior is modulated by the attentional hypothesis rather than explicitly multiplying two separate objects – this is for notational convenience, due to the particular way in which the approximate posterior is derived.

attention as an extra hypothesis with the mathematical form, but without the restrictive semantics, of a prior, we can replicate both scenarios in the same model.

A classic precueing task is illustrated in Figure 6.4, in which a brief spatial cue precedes an oriented Gabor patch (reproduced from Cameron et al. (2002)). If the cue is valid, i.e. if it indicates the true location of the upcoming stimulus, judgements about stimulus orientation are improved relative to those following a neutral cue. If, on the other hand, the cue is invalid and misleads the observer as to the location of the upcoming stimulus, performance is impaired. The behavioural data shown in Figure 6.5c illustrates this trend, and is reproduced from Luck et al. (1996)[6]. We also reproduce behavioural data from a study by Cameron et al. (2002), who compared only valid cues to a neutral condition, but found a beneficial effect of attention to the cued location at a range of different contrast levels, giving a C-R function (see Figure 6.5a).

Figure 6.6 illustrates how we modeled perceptual inference in this task, and the effect of attention on the inference. The world is represented by a single multiplicity function over space and orientation; $\mathbf{m}^o$, which generates noisy observations; $\mathbf{s}^o$, according to Equations 6.15 and 6.16. An approximate posterior is then computed according to Equation 6.22, representing perceptual inference without attention. The attentional hypothesis consists of a local mode at the cued location – which for a valid cue is the location of the upcoming stimulus, and for an invalid cue is on the opposite side of the array. This local mode is implemented by increasing the value of each diagonal element of $\overline{\mathrm{U}}_a$ that corresponds to the cued location – these correspond to a column in the rasterised 2D represention of the vector $\mathbf{u}^c$ (see Figure 6.3). With attention, the posterior is therefore computed with $\overline{\mathrm{U}}_0$ replaced by $\overline{\mathrm{U}}_a$ (see Equation 6.24). Figure 6.6 illustrates inference with a valid cue, and with neutral, distributed attention (corresponding to a uniform attentional prior).

We then use the approximate posterior distributions computed in the different attentional conditions to simulate an orientation judgement, by selecting the mean of the posterior; $\hat{\mathbf{m}}_0^o$, and then integrating out space to produce a marginal orientation distribution whose peak constitutes the model's 'judgement'. In the example shown in Figure 6.6 the posterior computed with a valid precue produces a more accurate decision than that computed with neutral, distributed attention – the peak of the marginal orientation distribution occurs in the correct location. By simulating multiple such 'trials', we can plot the percentage correct on the orientation judgement task for different cue, and contrast, conditions. In Figure 6.5b we reproduce the C-R function for valid vs. neutral cues shown in Figure 6.5a

[6]Luck et al. (1996) use two different masking conditions, indicated by the two curves in Figure 6.5a. We don't model masking effects, and so reproduce the qualitatively consistent pattern of results shown in both curves.

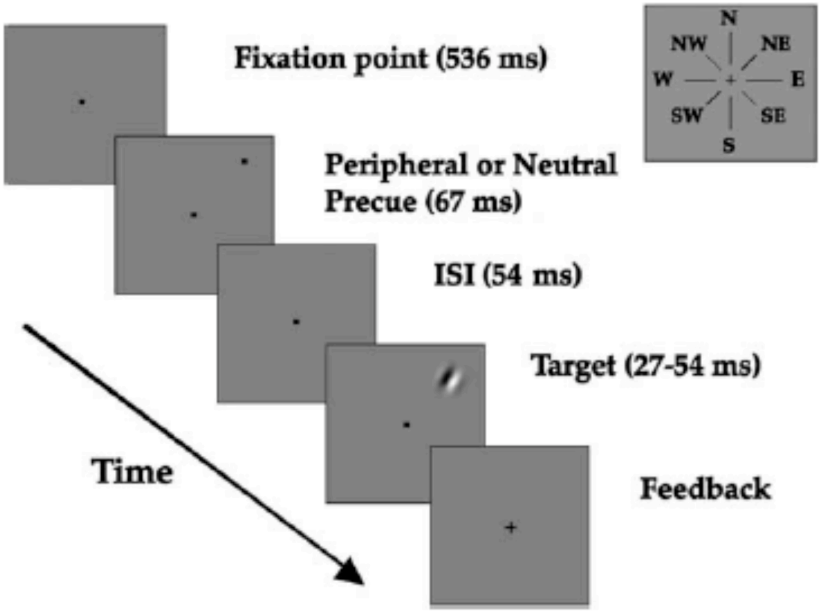

Figure 6.4: ***Precueing: Task schematic*** *This schematic illustrates a typical precueing task, adapted from Cameron et al. (2002). A spatial precue precedes a stimulus to which an observer must then make a featural judgement – here, indicating whether a Gabor patch is oriented to the left or right of vertical. The schematic illustrates the case in which the cue is valid, correctly indicating which of the eight compass locations the stimulus will subsequently appear in, which supports improved judgements relative to a neutral condition (see Figure 6.5a and c for behavioural and model results). Cues can also be invalid, in which case featural judgements are impaired relative to neutral.*

(reproduced from Cameron et al. (2002)), and in Figure 6.5d we reproduce the pattern of results for valid, neutral, and invalid cues at a single contrast level shown in Figure 6.5c (reproduced from Luck et al. (1996)). In both cases there is a good qualitative match.

As in previous models of attention as prior, excluding an irrelevant spatial region takes with it irrelevant featural information, improving feature judgements in the focus of attention[7] Here, we extend this effect to distributions over multiplicity functions, rather than over the location and feature value of a single object. Returning to the noise reduction vs. signal enhancement debate (see page 160), it is clear how a discretised location judgement with a *coarser grain* than that of $p_a(\cdot)$ could be perfect whilst still leaving enough spatial uncertainty to produce an improvement in feature judgements following a valid precue. The tasks we model here and for response-cueing involve exogenous cues that involuntarily

[7] A similar effect is exploited in Automatic Relevance Determination methods in machine learning (see e.g. Sahani and Nagarajan, 2004; Beal, 2003).

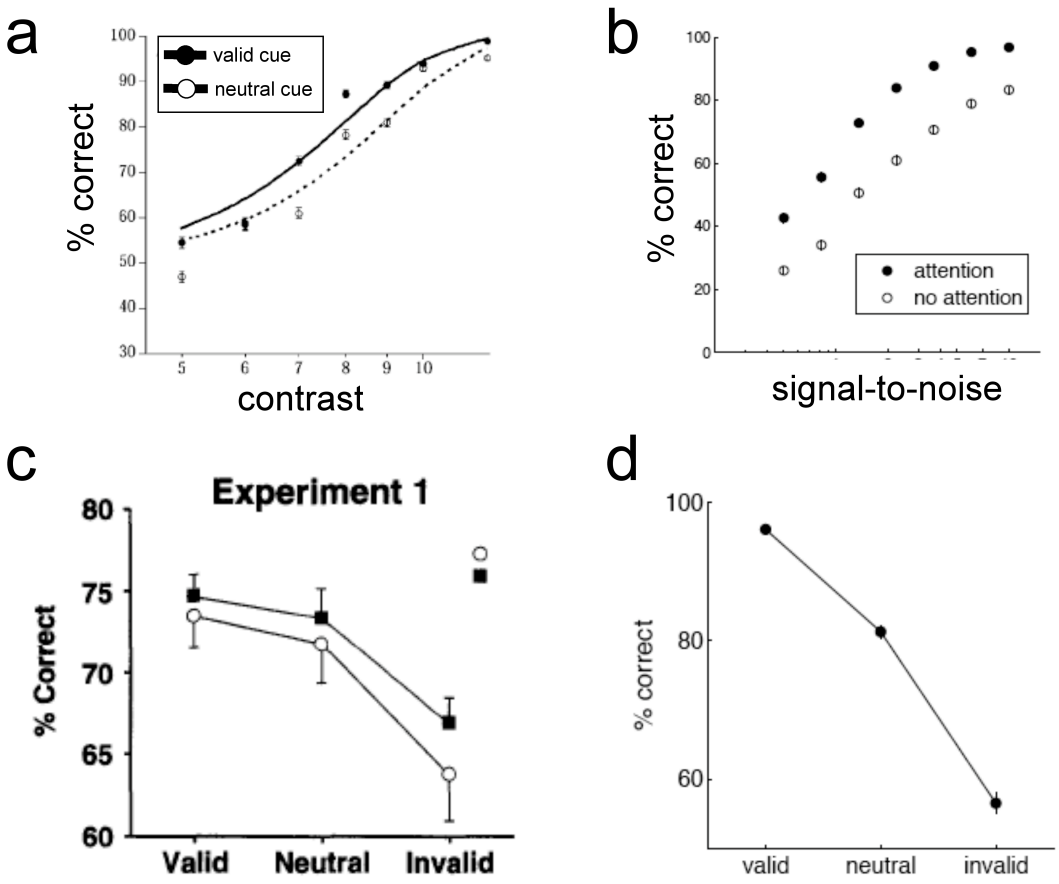

Figure 6.5: ***Precueing: Behavioural and model results a,*** *Shows the effect of a valid cue relative to neutral at different stimulus contrasts, for the task illustrated in Figure 6.4, resulting in a C-R function (reproduced from Cameron et al. (2002)).* ***b,*** *Data from the model simulation, replicating the qualitative pattern of results.* ***c,*** *Behavioural data showing for a single stimulus contrast the effect of valid and invalid cues relative to a neutral cue (reproduced from Luck et al. (1996)).* ***d,*** *Data from the model simulation again reproducing the qualitative pattern of results, though note that we don't address the effect of masking illustrated by the two curves in* ***c****. See text for details.*

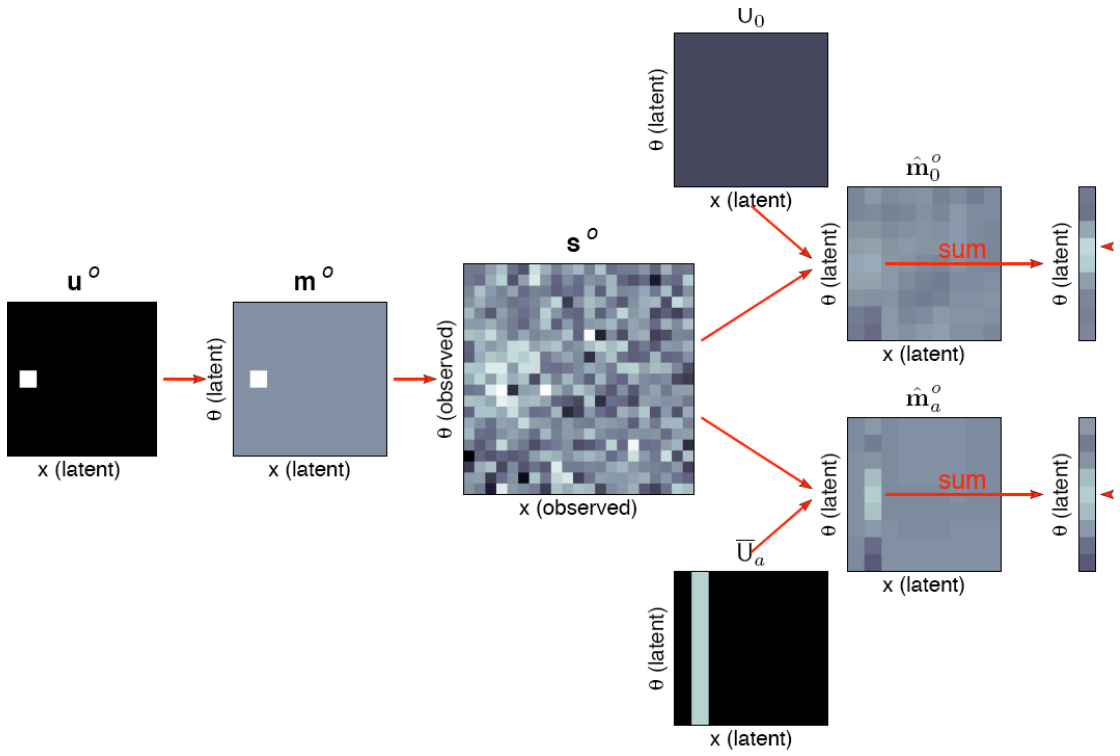

Figure 6.6: ***Precueing: Exemplar observation and inference.*** *This schematic depicts the generative model that produces noisy observations,* $\mathbf{s}$*, from a state of the world,* $\mathbf{m}$*, and illustrates how an approximate posterior is constructed on the basis of these observations with and without an attentional precue. The panels at the far right of the schematic indicate how a behavioural decision that mimics that made by observers in the real paradigm (Cameron et al., 2002; Luck et al., 1996) could be made on the basis of these approximate posteriors. See text for details.*

attract attention. However, since our model does not say where $p_a(\cdot)$ originates from, the same effect could apply to endogenously driven or symbolic cues.

As discussed above, previous models have configured attention as a prior over locations both for precueing tasks in which it genuinely carries prior information (Dayan and Zemel, 1999), and for response-cueing tasks in which the cue instead indicates which of a number of upcoming stimuli a response will subsequently be required (Rao, 2005). Next we simulate the latter scenario, reproducing behavioural data from a response-cueing task by Pestilli et al. (2007) (see Figure 6.7a for a task schematic). The parameters of the model were very similar to those used for the precueing task, but here two objects were always present at two fixed locations, one in each half of the environment. Their orientations were chosen randomly, and their contrast levels were set to be equal to each other and were varied parametrically. The resulting feature map was projected into the observed space by the same

weight matrix as before, and the observations again corrupted by uncorrelated Gaussian noise of fixed variance.

Inference was simulated under three different conditions, in which the attentional hypothesis corresponded to the valid, neutral, and invalid cues used by Pestilli et al. (2007). In the *valid* condition, attention was directed to the location that the response cue would subsequently appear in, by the brief presentation of a peripheral cue shortly before the oriented stimuli. The physical cue used in the experiment was broad, and so we modelled the attentional prior as concentrated on three locations, centred on the correct one (illustrated at the bottom right of Figure 6.8). In the *neutral* condition the attentional cue appeared near the centre of the screen. As the allocation of attention in the experiment was designed to be involuntary, and was therefore presumably invoked as much by this central cue as by the peripheral ones, we modelled the attentional prior in this case as directed towards three locations in the centre of the field (illustrated at the top right of Figure 6.8). Finally, in *invalid* cue trials, the attention cue was on the side opposite to the response cue. We thus modelled the attentional prior as directed to the opposite side (not shown).

Decisions about the orientation of the stimulus at the response-cued location were then made on the basis of the inferred posteriors for each attentional condition. In the precuing experiment discussed above, where only one object was present, we modelled the response on an integral over all space. Here, we assumed that the response cue biased the integration to one hemifield, with contributions from locations in the opposite hemifield being progressively down-weighted with distance from the cued location. The integrated feature map yielded an estimate of total contrast at each possible orientation, and the highest contrast orientation was taken as the model's response. The results of the simulation are shown in Figure 6.7c. The qualitative agreement with the results of Pestilli et al. (2007) is strong and in particular, the model reproduces the vertical shift of the C-R function under the different attentional conditions. Above we argued that linking changes such as shifts or multiplications of C-R functions to mechanisms of noise reduction vs. signal enhancement is poorly constrained without an explicit model of the underlying mechanisms (see e.g. Lu and Dosher, 1998). The framework presented here provides a basis on which to build such models, one which extends the SDT approach to continuous, feature-map representations, modelling the baseline effects of noisy or approximate representations whilst also allowing a role for attentional selection.

6.3.3 Binding, Illusory Conjunctions and Visual Search

The second effect of attention we described in Section 6.2.3 was to implicitly reveal correlations neglected in the approximate distribution, as the attentional hypothesis dy-

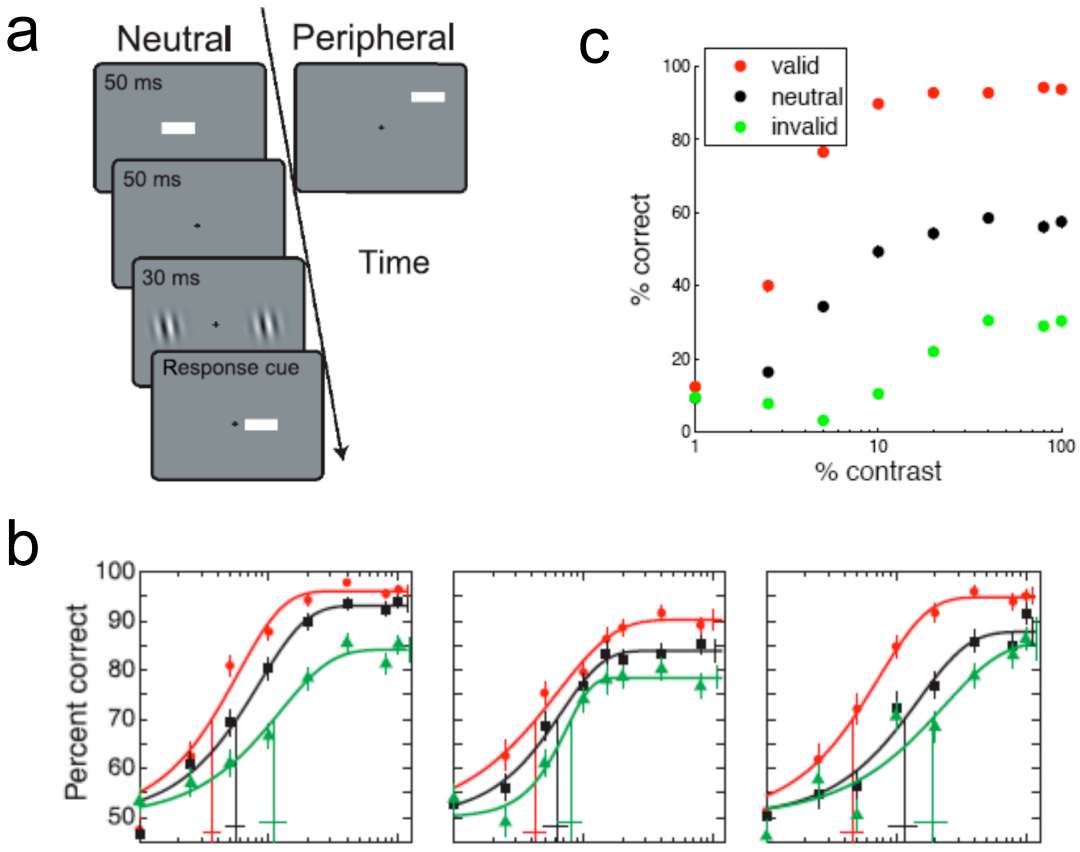

Figure 6.7: ***Response-cueing: Task schematic and results.*** ***a,*** *A schematic of a typical attentional paradigm where the attentional signal is not well characterised as a data driven prior – i.e. where the attentional signal is due to an instruction about relevance, not to a change in the physical stimulus information. A precue either validly, neutrally, or invalidly indicates the location of the stimulus that observers will subsequently be prompted to respond to. As in the precueing task, a valid cue improves performance, and an invalid cue impairs performance, relative to the neutral cue.* ***b,*** *gives the results from three observers on this task at different contrasts (both* ***a*** *and* ***b*** *adapted from Pestilli et al. (2007)).* ***c,*** *shows the results from the model, which reproduce the beneficial effect of a valid cue relative to neutral, and the detrimental effect of an invalid cue. See text for details.*

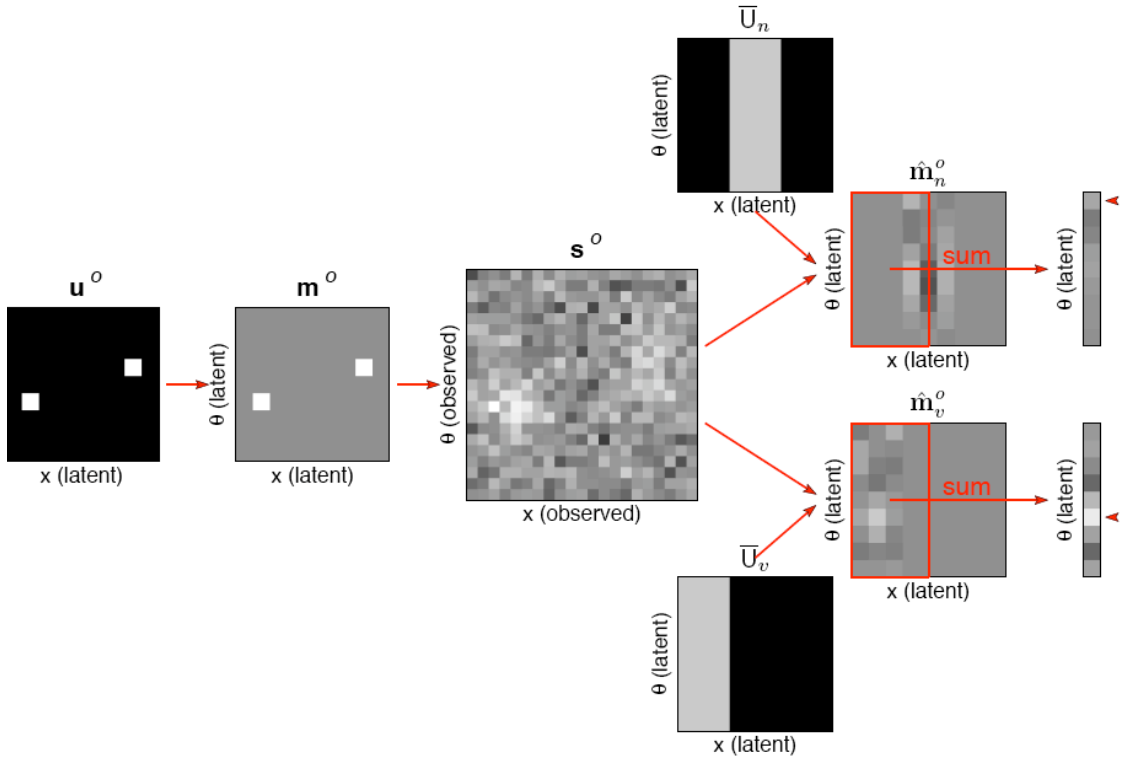

Figure 6.8: ***Response-cueing: Exemplar observation and inference.*** *This schematic depicts the generative model, which produces noisy observations in the same manner as for the precueing scenario described above. The attentional hypothesis modulates the prior in the same way as for precueing, but the behavioural decision is made on the basis of a slightly different computation. See text for details.*

namically evolves towards a better match to the true posterior. This can also be thought of as a process of model selection, in which the attentional hypothesis tests possible explanations for the data, thus locally removing the need for negative 'explaining away' correlations or mimicking the effect of positive correlations such as those due to co-location. This is closely related to the binding problem discussed in Section 2.4.3 of the literature review, and to the idea that attention assists with binding features into objects (Robertson, 2005; Treisman, 1998). Here, we model an illusory conjunction paradigm, used to argue for the role of attention in binding colocated features.

Figure 6.9a schematises the classic illusory conjunction task reported by (Treisman and Schmidt, 1982). Observers were asked to monitor two streams of black digits for a target, and also to report the colour and orientation of the central bars. The primary monitoring task distracted attention from the central bars, and observers often misbound colour and orientation – they might, for example, report seeing a red vertical or green horizontal bar

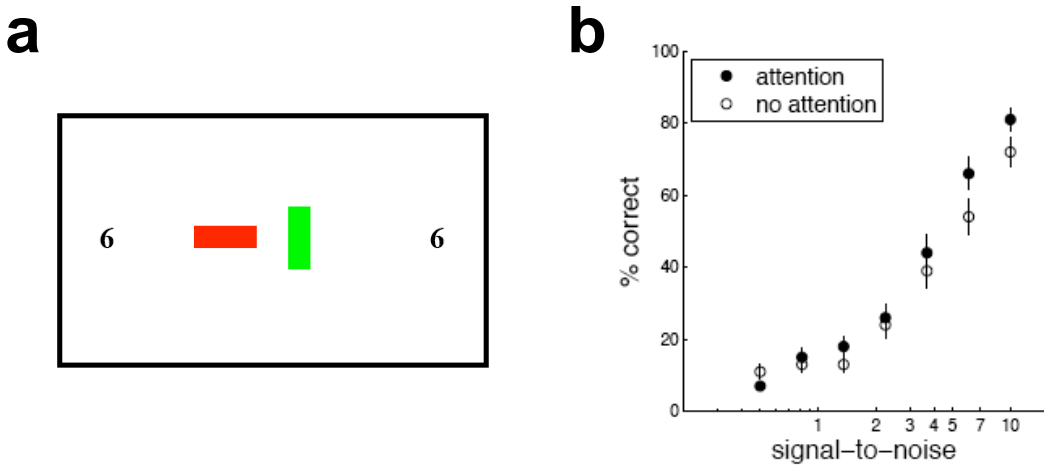

Figure 6.9: ***Illusory conjunctions: Task schematic and results.*** ***a,*** *schematises a classic illusory-conjunction paradigm, in which observers were asked to monitor two peripheral streams of black digits for a target, and also to report the colour and orientation of the central bars – with attention distracted by the monitoring task, observers often misbound colour and orientation.* ***b,*** *Results from the simulation, giving % correct binding judgements with and without attention, and at different SNR. See text for details.*

rather than one of the two correct pairings. There are many different examples of the illusory conjunction paradigm, and many debates about their interpretation (see Section 2.4.3 and Ashby et al., 1996). Following Treisman and Schmidt (1982), many experimenters abandoned the secondary, attention-distracting task due to potential memory confounds, and instead used short presentation times, assuming that attention would be prevented from moving around the objects in the scene and that *with* full attention no illusory conjunctions would occur – similar in spirit to visual search paradigms (see page 71).

Here, we modelled a setting similar to that studied experimentally by Hazeltine et al. (1997), which avoided a secondary task, instead showing observers a brief array of coloured letters followed by a bright mask. There are multiple different ways of eliciting binding judgements (see e.g. Cohen and Ivry, 1989; Ashby et al., 1996; Hazeltine et al., 1997; Prinzmetal et al., 1986), and the "report everything you see" method of Treisman and Schmidt (1982) is both difficult to model, and makes the interpretation of behavioural results particularly hard. We follow the reporting method used by Hazeltine et al. (1997), in which observers were asked to report the location of the green letter by pointing at the screen, and then say whether that green letter was an 'O'. In half the trials, the green letter was indeed an O, in a quarter an O of another colour appeared somewhere else in the display,

and in the remaining quarter no Os were present. Hazeltine et al. (1997) were interested in trials where the green letter was misidentified as an O *whilst an O of another colour was present*, and in whether the reported location of the green letter was displaced towards the location of the actual O, suggestive of an illusory conjunction (see Figure 6.11a). A second experiment yielded similar results when the roles of letter identity and colour were reversed – subjects reporting the location of the letter O and then whether or not it was green – showing that the role of the two features was symmetric.

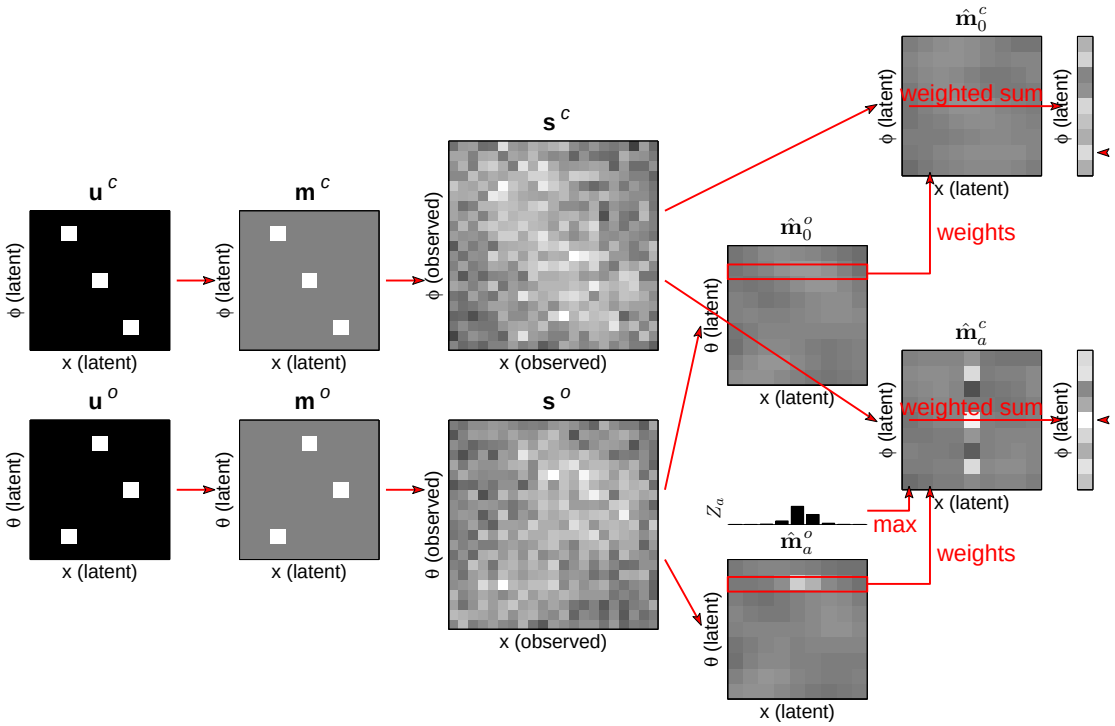

Figure 6.10: ***Illusory conjunctions: Exemplar observation and inference.*** *Depicts the generative model, which produces noisy observations in the same manner as for the precueing scenario described above. The model is asked to report the location of an object defined in the θ feature dimension, and then to report whether that object also has a particular value on the other feature dimension, ϕ. With attention, an attentional hypothesis evolves to select the most likely location of the target value of θ, rather than the model finding the centre of mass along the location dimension for the target value of θ (i.e. in the restricted region of the feature map indicated by the red horizontal box). See text for details.*

We used the same 'grid world' as for the precueing and task-relevance simulations, this time placing 3 objects close together in the centre of the field[8] and using a slightly extended

[8]Hazeltine et al. used an array of 5 letters, but we were able to obtain similar results with the smaller, three object array.

spatial spread in the generative weights, Λ. The simulation is illustrated in Figure 6.10, and the 3 different feature conjunctions are represented by the three white entries in each of the two scan-rasterised $\mathbf{u}$ matrices at the far left. Although Hazeltine et al. (1997) did not explicitly manipulate attention, previous studies (Treisman and Schmidt, 1982; Prinzmetal et al., 1986) have shown that the rate of illusory conjunctions is higher when attention is engaged in a simultaneous task. We therefore modelled responses both with and without a dynamically adapted attentional hypothesis.

In the absence of attention (upper half of the right-hand side of Figure 6.10), posterior distributions over both feature maps, $\mathbf{m}_0^o$ and $\mathbf{m}_o^c$, were inferred according to Equation 6.22, and the means of these posteriors (the MAP estimates; $\hat{\mathbf{m}}_0^o$ and $\hat{\mathbf{m}}_0^c$), were used to make judgements. A particular discretised value of θ (corresponding to "the letter is green" in the experiment) was taken to identify the target. We modelled the reported location of this target as the centre of mass along the location dimension, x, for that target value (i.e. in a restricted region of the feature map, as indicated by the red horizontal box in $\hat{\mathbf{m}}_0^o$). Each location within the ϕ feature map (each column of $\hat{\mathbf{m}}_0^c$) was then weighted accorded to its value in the restricted region of the θ map. We then summed the weighted map over space to obtain a marginal distribution over ϕ, from which the highest mean feature strength could be selected. This perceived feature value was compared to the target value of ϕ (corresponding to "is the green letter an O?") to yield a binary response.

The mechanism with attention was similar (lower half of the right-hand side of Figure 6.10), except that the posteriors were obtained with an attentional prior, $p_a(\cdot)$, which was allowed to evolve to focus on the location most likely to contain the target θ value. We modelled the effect of this evolution by comparing the 'evidence' values Z_a for attentional hypotheses centred at each possible location (according to Equation 6.25), where each hypothesis stated with certainty that an object was present at the given location with the target value of θ and an unknown value of ϕ (with other objects potentially being present with the same background rate as in the no-attention case). The relative values of Z_a for each hypothesis in the example in Figure 6.10 are shown in the bar graph above the posterior mean of the orientation feature map; $\hat{\mathbf{m}}_0^o$. Attention was assumed to settle on the hypothesis with the maximal value of Z_a, yielding an attentionally-modulated posterior according to Equation 6.24. Inference and decision making then proceeded exactly as above, but with the associated attentional prior $\overline{\mathrm{U}}_a$ replacing $\overline{\mathrm{U}}_0$ (see Equation 6.24).

We found that the introduction of attention did indeed reduce the rate of binding errors (including illusory conjunctions – where the O is incorrectly reported to be green whilst another green letter is present) as reported in previous studies. This was true across a range of different signal-to-noise values Figure 6.9b). We also found the same displacement effect as reported by Hazeltine et al. (1997), illustrated in Figure 6.11a, both with and

without attention, illustrated in Figure 6.11c and b respectively. All plots in Figure 6.11 include data from trials where the target value of θ is *not* co-located with the target value of ϕ, and plot localisation of the target for both correct rejections (where the observer correctly says that the green letter was *not* an O; white circles), and false positives (where the observer incorrectly reports that the green letter was an O; black squares). In all cases, the 'distractor' letter was located to the right of the target on the plot, and there is an attraction towards the location of the distractor when the observer incorrectly judges that both target feature values came from the same object.

In the literature review, we discussed psychophysical evidence that attention does not impact on the ventriloquism effect, which can be thought of as a kind of 'binding' between cues from different modalities that each carry information about the location of a common cause (see Vroomen et al., 2001; Bertelson et al., 2000, and page 73). We suggested that, rather than attention having no role to play in improving representations of the co-location of features, in this task the underlying representation was sufficiently good that attention had little role to play. Here, we make this more explicit, in a model where the context-dependent effects of attention arise from the interaction between the fidelity of the judgement required, the capability of the approximate posterior, and the form of the attentional hypothesis. Future experimental work trying to isolate an affect of attention in cue combination scenarios would strengthen this argument.

In future work, this model could be extended to a visual search paradigm with different numbers of objects, though this will require implementing the dynamics of $p_a(\cdot)$ and introducing thresholding mechanisms into the decision process (see Section 6.4 below). As discussed in Section 2.4.3, the distinction between pre-attentive 'pop-out' search for single features and serial, attentive search for feature conjunctions has been repeatedly challenged. In our framework, attention is not an either/or effect of a serially allocated discrete spotlight, rather it is a continuum of effects of a continuously evolving, probabilistic hypothesis. As such, we would expect search behaviour to sometimes look like parallel search (when the baseline approximate posterior is sufficient for the judgement required), sometimes to look like serial search (where the approximate posterior is very impoverished relative to the judgement), and anywhere in between – incorporating the evidence for a continuum of search effects that caused such trouble for FIT. Following challenges to the FIT picture of free-floating features discretely and uniquely bound by attention, Wolfe and Cave (1999) proposed that we should see the pre-attentive world as "populated by unrecognized bundles of features loosely held together by virtue of their shared location", and that higher resolution knowledge of the relationship between features then requires spatial attention (see page 71). This idea is echoed in our framework, but we extend it to *degrees* of bundling without attention, and *degrees* of improvement with attention.

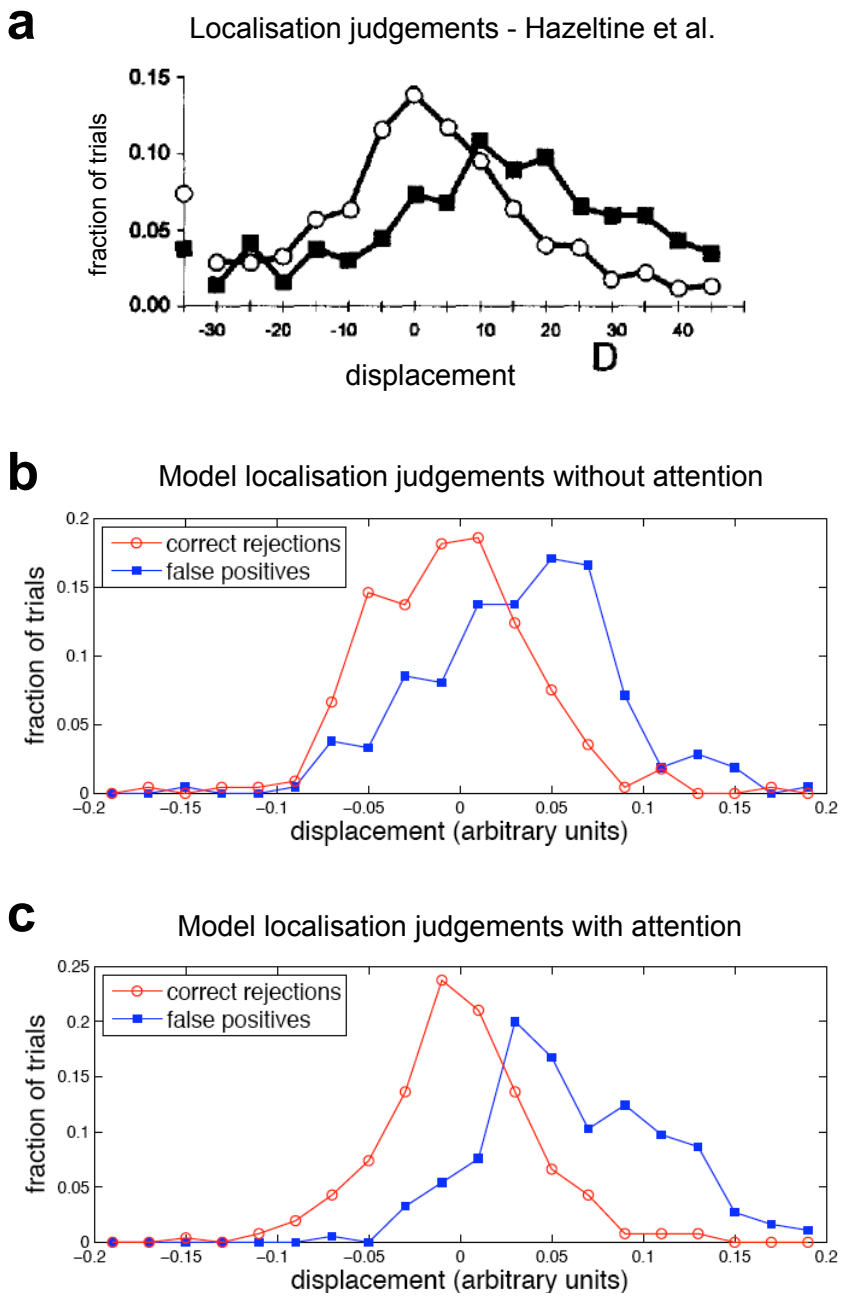

Figure 6.11: ***Localisation judgements with and without attention.*** ***a,*** *Results from Hazeltine et al. (Figure 1) – observers were asked to locate the green letter, then say if it was the letter O – trials plotted here are those in which the green letter was not an O, but an O of another colour was present in the display. White circles indicate localisation when the observer correctly rejects the hypothesis that the green letter was an O, and this is centered on the true location, 0. Black squares indicate localisation when the observer incorrectly reports that green letter was an O, and is biased towards the location of the 'distractor' O (located at 'D').* ***b, c,*** *Our model produces the same bias, with and without the attentional hypothesis. See text and Figure 6.10 for details.*

6.4 Challenges for Modelling Under the Framework

We hope that the framework developed here will support new, probabilistic perspectives on attentional selection. Above we presented simulations of key attentional paradigms in a simple, abstract model, illustrating how the framework might help to resolve some of the apparent dichotomies and disagreements in selective attention research. In the future, building more detailed models of specific behavioural tasks under the framework should allow for concrete predictions about the dynamics and consequences of attentional selection, and a richer investigation of the role of different sources of uncertainty and noise. Below we discuss some of the implementational challenges this will raise.

Electrophysiological data on the size of RFs, the response properties of neurons in different cortical layers, and the anatomical connections between regions should be used to guide the form of the approximation for a particular task. For example, in the factored approximation approach, a computation thought to be supported by cells with very small receptive fields would factor quite finely over space, whilst a computation thought to be supported by cells that are responsive to combinations of features might not factorise over those particular features. How the approximation is computed will depend on its form – above we used a simple, fully-factored Gaussian approximation that could be analytically computed, but in most cases minimising the KL divergence will be intractable. Sampling methods can be used to approximate intractable integrals, but are in general less amenable to biologically plausible implementations. Another option is to use a form of approximate BP or EP (see Minka, 2001, 2005; Mackay, 2004, and page 153). These techniques involve using local 'messages' to iteratively update tractable portions of the distribution, and their applicability in situations such as those described in the attention framework is an active area of research. When modelling inference in interconnected areas, this approximate inference will take place in a hierarchical structure (Friston, 2003, 2005; Hinton et al., 1995, 2006; Deneve, 2008a,b).

In the simulations above we avoided implementing the dynamics of the attentional hypothesis, aiming to demonstrate the key effects in as simple a setting as possible. In order to implement the dynamics of $p_a(\cdot)$, it is necessary to measure Z_a, and compute a gradient measure such that the parameters of $p_a(\cdot)$ can, under constraints on its form, evolve towards higher Z_a and thus a better match to the true posterior. As discussed in Section 6.2.1, the generative model is encoded in the connections set during learning of the approximate recognition model. It is not directly accessible – and indeed, had better not be, since representing the true likelihood and prior is exactly the step we claim is intractable. This makes it hard to compute Z_a directly, but an estimate of the normalising constant of the explicitly represented *approximation* could be used as a proxy that would reflect the relative increase

or decrease of Z_a over time. In most neural coding schemes, a simple and easily accessible quantity such as mean firing rate reflects the certainty in the distribution and thus its normalising constant. It might also be possible to use the approximate posterior to obtain a proxy for the gradient, or else a sampling method could be used that repeatedly draws 'proposals' from the stochastic diffusion of $\mathbf{r}_a$, which are accepted if they lead to an increase in Z_a. Another option might be to propagate gradients in a message-passing algorithm, though this raises similar issues of biological plausibility to those levelled at backpropogation (see Stork, 1989). And finally, one could explore the use of a variational free-energy bound, although in its canonical form this involves optimising $\mathbf{KL}\left[q(\cdot)\|p(\cdot)\right]$ rather than $\mathbf{KL}\left[p(\cdot)\|q(\cdot)\right]$ as in our framework.

During the evolution of the attentional hypothesis, its mode will pass through various local minima, and dynamic implementations will therefore require a mechanism for escaping these diversions. A threshold on the value of Z_a or its gradient could be used to trigger stochastic diffusion of $\mathbf{r}_a$, perhaps biased away from previously visited regions to reflect 'inhibition of return' phenomena (see Klein and Ivanoff, 2005; Itti and Koch, 2001). The relationship between the evolution of spatial and featural components of the attentional hypothesis should be constrained by experimental data, and might require independent or linked thresholding mechanisms. A recent study by Egner et al. (2008) suggested that featural and spatial attention could be observed as a common, combined signal in IPS. Modelling any behavioural paradigm requires a mapping between the approximate posterior and a perceptual judgement, which should be chosen to reproduce basic behavioural effects and to reflect existing knowledge about Bayesian decision-making. The project we have argued for in Chapter 5 should help in this endeavour. Relatedly, the framework does not consider the role of mechanisms external to the posterior representation, for example, salience, arousal or memory. Extensions could be made to incorporate these influences, much as Peters and Itti (2007) 'bolted on' top-down signals to an existing model of bottom-up salience computations.

These are challenges for implementing a model on the probabilistic level, inspired by the properties of the neural populations thought to represent the relevant quantities. To take it one step further, neural coding models that implement these probabilistic models could be built, providing a substrate for linking behavioural and electrophysiological data as schematised in Figure 1.3 (see also Section 2.2.4). This would require extensions of the existing PPC codes that consider distributions over features and locations of single objects, extending the DDPC approach proposed by Sahani and Dayan (2003) to implement distributions over multiplicity functions of the type we use here. Developing neural coding models might also contribute to an improved understanding of the neurophysiological data on attention (see for example Treue, 2001; Williford and Maunsell, 2006). For example, we

might be able to identify particular parameters that lead to particular attentional changes in the response profiles of the modelled neurons, and thus isolate the statistical characteristic of attention that relates to those changes.

Implementing a full hierarchical model, which could be applied to any given paradigm, is far beyond the current capabilities of the machine learning techniques that would be required. In addition, current knowledge about coding properties of neurons in different cortical areas, and their interconnections, is not specific enough to inform such a model (or in the words of Roskies (1999), we don't currently have enough anatomical knowledge to properly constrain the binding problem). However, it is still important to situate detailed models of particular computations in a bigger picture of how the hierarchical, recurrent structure of the brain might perform Bayesian inference. This raises important questions about how approximate inference on one level impacts on inference at the next – the simple semantics of Bayes' rule are clearly muddied by this process, and by the introduction of the attentional hypothesis. As we discussed in the introduction (see page 25) this scenario is common in machine-learning approaches to intractably complex problems, and we likewise think of approximate inference as preserving the principles and machinery of Bayes' rule as far as possible under constraints. Recordings from a particular cortical area, whilst an animal performs a simple task, are often thought of as reflecting the full posterior belief about that particular object, and as such don't invoke these issues. When extending to complex, real world scenarios, linking models of approximate, hierarchical inference to neural responses, and considering the relation of such models to existing accounts of Bayesian inference in neural hierarchies (see Rao and Ballard, 1999; Lee and Mumford, 2003; Friston, 2005) is an important and exciting challenge.

There are two specific questions that must be answered for any hierarchical implementation. First, how the neural code for representing the attentional hypothesis is 'read' by the cortical regions that receive it. All cortical neurons have an RF, and so represent space to some degree, suggesting that the spatial component could be passed directly to various layers in the hierarchy. However, preferred features become successively more complex and invariant, and the featural component of attention might therefore have to be cascaded back down through the cortical hierarchy, piggybacking on the code translation that occurs between adjacent layers. Our framework does not currently account for 'object-based attention', but this could be expressed either in terms of high-level features, or in terms of a further layer in the model where latent objects correspond to distributions over features – experimentally, the neural distinction between features and objects is still somewhat unclear (see Cavanagh and Alvarez, 2005). The second, related question is how approximations on one 'level' of feature-analysis relates to approximation across 'levels' of increasingly abstract feature representations. It is sometimes claimed that the specialised, hierarchical nature of

cortical representation means that there isn't in fact a problem of combinatorial explosion in representing combinations of multiple, fine-scaled features. We argued against this position in Section 2.4.3, but another way of thinking about it might be that the fractured structure is part of the brain's *solution* to the combinatorial explosion that *would* exist were the brain to attempt to represent the full posterior. And perhaps, as we have argued for the preservation of probabilistic information for flexibility in further computations, preserving different scales of analysis yields similar flexibility.

6.5 Discussion

The concept of selection is central to attention research. However, the majority of studies have focused on how and when selection occurs, rather than on characterising why selection is necessary in the first place. One reason for this is that the rubric of attention covers a large range of effects which have heterogenous anatomical, functional, and neurophysiological bases. Recognition of this fact has led to a recent move away from the concepts of a generalised limited capacity process and universal attentional mechanism (see Driver, 2001; Zelinksy, 2005). In this chapter we have proposed a new framework for selective attention, which provides a unifying, normative *computational* account of the nature of the limited resource, why it is limited, and how attention makes tractable the computationally inaccessible operations that result. In the context of the BCH, we propose that representing the 'computationally intractable' (Papadimitriou, 1994) correlational structure of large joint distributions constitutes a general computational resource limitation. The brain therefore approximates the true posterior with a (perhaps factored) distribution that neglects some of its correlational structure, perhaps corresponding on the implementational level to cortically specialised regions and neurons with limited spatial and featural receptive fields (see also Dayan, 2008a; Yu et al., 2008).

Attention acts to bias this approximation by imposing a local 'hypothesis' about the state of the world, which takes the mathematical form of an extra prior. This attentional hypothesis is driven by a variety of bottom-up and top-down signals, and existing knowledge and expectations. As in previous accounts of attention as a Bayesian prior (Dayan and Zemel, 1999; Rao, 2005), and in SDT approaches to attention as uncertainty reduction (Palmer et al., 1993; Gould et al., 2007; Dosher and Lu, 2000), attention improves stimulus judgements in the region of its local mode by neglecting information outside it, but we extend this case to probabilistic representations of complex, multi-object scenarios. This effect can locally improve any posterior, whether or not it is approximate, but here we also propose a new role for attention in revealing correlations neglected in an approximate

representation. This occurs via the continuous evolution of the attentional hypothesis towards a better explanation of the world – a better match to the true posterior. These two classes of effect were simulated in a simple, abstract model, the former demonstrated via simulations of a precueing and response-cueing task, and the latter via simulations of an illusory conjunction paradigm.

There are a number of opportunities for assuming a capability in the attentional mechanism that is lacking in the impoverished representations it was invoked to refine. This would clearly render the solution trivial, and we have covered several ways in which this is avoided. First, the attentional hypothesis is subject to the same basic representational capacity limitation, but rather than stretching its resources to represent as much of the true posterior as possible, its job is to choose a region for this coarser process to focus on. It therefore consists of local modes in space and feature dimensions, reflecting the evidence for a limited focus of attention that inspired the languishing bottleneck and spotlight metaphors of early accounts of attentional selection (see Driver, 2001). It is also key that attention does not *create* information, rather it locally refines the processing of the information available. This avoids the uneasy feeling in the signal enhancement account that the brain is reaching out a little too far into the world, but is still able to account for improved stimulus representations. Manipulating inference rather than reallocating representational semantics means that the neural code is preserved under attention – each neural population in the complex and recurrent hierarchy of the visual system still talks about the same quantities, but changes what it says about them (see Ghose and Maunsell, 2008). Finally, in computing an approximation to the true posterior we clearly can't invoke it – we therefore propose that the brain learns over time an approximate recognition model that implicitly embodies the true likelihood and prior, following algorithms such as the wake-sleep algorithm for learning in a Helmholtz machine (Hinton et al., 1995; Dayan et al., 1995).

A key feature of our approach is that it makes explicit what the processing limitation is that necessitates selection, and why it is instantiated so variously through the lens of different paradigms. A probabilistic representation admits of degrees rather than distinctions, and predicts that the properties of the underlying approximation might sometimes be well enough matched to the current task requirements that the attentional hypothesis has no observable effect. This helps to unify apparently disparate results concerning when attention is needed to bind features together - rather than proposing a non-bound, bound, and perhaps intermediate 'bundled' stage (see page 71 and Wolfe et al., 1989), we can look at degrees of accuracy in judgements about multiple, spatially colocated features. In the wake of the early vs. late selection debate (see Section 2.4.1), Allport (1989) suggested that a selection stage is only a relevant concept when there is a conflict between the receptive field properties of a cell and the properties of the current state of the world – our framework

expresses a related notion but on the computational level rather than in terms of a specific implementational property.

Signal detection theorists have argued that apparent behavioural 'signatures' of the serial allocation of attention can be explained in terms of the addition of noise (see e.g. Morgan et al., 1998; Palmer, 1994). Our probabilistic framework extends this notion, consistent with other Bayesian models showing that apparent attentional effects could arise from inference in a probabilistic, and approximate, model (see Dayan, 2008a; Yu et al., 2008; Li, 2002; Morgan et al., 1998). But our framework also articulates a role for attention within this 'noise reduction' view. Proponents of signal enhancement have used a variety of arguments to suggest that attention can directly enhance the stimulus, even when there is apparently no 'noise' to be reduced (e.g. Cameron et al., 2002). We argued that uncertainty sufficient to produce concurrent improvements in the stimulus representation might be present, even if not measurable in behavioural judgements, and suggested that developing models which make explicit sources of noise and clearly define signal enhancement are critical to disentangling this somewhat knotty debate (see Lu and Dosher, 1998, for an example of this approach in the SDT framework).

The 'biased competition' model we discussed in Section 2.4.4 described attention as biasing an ongoing competition for stimulus representation (Desimone and Duncan, 1995; Desimone, 1998), but left open questions about the source and nature of the limitation. Here, we make explicit, on the computational level, why there is a competition in the first place – the brain is trying to approximate a complex posterior with limited representational resources. The approximation process induces a competition for probability mass that ignores information that should bias this competition in favour of the more likely of a number of competing explanations – in the factored example used here, it treats them as independent. Attention coarsely mimics the neglected information within a local region, biasing the competition for probability towards explanations within its local mode, and via its evolution, towards those that reflect neglected correlations. In the biased competition model, attention is viewed as a collection of distributed and emergent effects, which raises questions about how a coherent focus of attention is constructed. In our probabilistic framework there is a discrete attentional signal acted upon by different influences, ameliorating this problem and concurring with anatomical data that suggests a front-parietal attentional control signal. Thus far there has been limited work on Bayesian characterisations of attention. Most of the existing work focuses on implementing models of visual search or predicting fixation paths (Navalpakkam and Itti, 2007; Mozer and Baldwin, 2008; Najemnik and Geisler, 2005, 2008), which in our framework corresponds more to the setting of the attentional hypothesis than to its raison d'être or effects. Approaches to attention as a Bayesian prior (Dayan and Zemel, 1999; Rao, 2005) are extended in the present work,

by making the attentional hypothesis an object with the mathematical properties but not the limiting semantics of a prior, and allowing inference to be modulated rather than gated by this object. The lack of work on attention comes in part from the focus on *evidence* for the BCH, which takes the form of optimality proofs with regard to single objects in the focus of attention (see Knill and Pouget, 2004; Doya et al., 2007, and Section 2.1 for full references). This has been paralleled by theoretical treatments of pdfs over features that are effectively conditioned on belonging to a particular object. Part of the *job* of the brain in perceptual inference is to decide which feature values belong to which objects, and part of the job of attention to assist with this – i.e. to solve the 'binding problem' it is often invoked to resolve (Roskies, 1999; Wolfe and Cave, 1999). We therefore consider probability distributions over multiplicity functions, corresponding to neurally inspired 'feature maps' and allowing us to deal with complex, multi-object scenes.

One of the aims of developing this framework was to motivate a reconsideration of the notion of a limited processing resource, and of the level of analysis at which selection and capacity limits might be usefully described. In the future, implementing detailed models of particular top-down attentional paradigms, which represent tractable portions of the picture presented in Figure 6.1, should motivate more specific reconsideration of how limitations in particular representations relate to the properties of the relevant neurons. Such models face a number of implementational challenges, and will rely on developments in machine learning techniques (see Section 6.4). Extending the doubly-distributional population code of Sahani and Dayan (2003) to multi-object displays would then allow neural network models to be built that embody the probabilistic equations, and open up bridges to link attentional effects in the probabilistic model to the changes in neural firing observed in contentiously interpreted neurophysiological experiments. To conclude, a computational perspective on the problem of attentional selection gives shape to the pervasive feeling that there is something common to its disparate behavioural, anatomical, and physiological substrates. Locating this commonality in a probabilistic framework makes central the previously problematic evidence that attentional effects admit of degrees rather than discretely localised distinctions, and opens up new avenues for interpreting experimental data in Bayesian terms.

7

General Conclusions

The Bayesian Coding Hypothesis represents the coming together of ancient approaches to the perceiving, reasoning mind with formal mathematical descriptions that have shown their utility in a remarkable range of domains. Bayes' rule stands as a structure, a prescription for reason whose crank can be turned whatever ingredients you put in. Proponents of the BCH believe that for the analysis of perception and action, and for understanding their neural substrates, the ingredients are well constrained enough that the normative models that come out are both informative and useful. To ensure this is the case, we need to make strong links between anatomical, algorithmic, and physiological implementations of the functions that Bayesian theorists seek to describe. In order to make the BCH more widely applicable as a theory of perception, we also need to move away from the picture of an optimal inference machine in simple, lab-based scenarios. In this thesis, we focused on exploring the limits of optimality, and on integrating knowledge about the neuroanatomy of decision-making and the psychology of attentional selection with theoretical, Bayesian approaches. Below we will discuss contributions, limitations, and future work for each of the three sections of the thesis – psychophysical investigations of Bayesian optimality with regard to stimuli of different complexity, a neuroimaging experiment asking where perceptual uncertainty and value are integrated in the brain, and a theoretical study attempting to broaden the remit of the BCH to complex scenes and the attentional selection invoked to deal with them. The thesis concludes with some questions and speculations about the potential of the BCH as an overarching explanatory framework for the perceiving, reasoning brain.

7.1 The Limits of Behavioural Optimality

7.1.1 Contributions

Experiments showing that people can behave Bayes-optimally, performing in a way that demands they take probabilistic uncertainties into account, is a key source of support for the BCH. In some of these studies, people must optimally combine different cues to a common underlying quantity, and in others observed illusions and biases are explained via the influence of Bayesian priors (see e.g. Knill and Pouget, 2004). As we discussed in Section 2.1 of the literature review, in order for a paradigm to demonstrate the use of uncertainty, optimality in that paradigm usually demands the integration of multiple distributions – be it two posteriors due to different cues, or a biased prior and likelihood – appropriately weighted by their uncertainties. We were interested in whether we could demonstrate optimality for single, unimodal visual posteriors, but making decisions on the basis of a single distribution usually requires just the mean, and so is uninfluenced by uncertainty.

In the motor domain, participants have been asked to combine knowledge of their motor uncertainty in a single posterior over a movement endpoint with *external loss functions* in order to maximise gain (see page 35 and e.g. Trommershauser et al., 2003b), as dictated by Bayesian decision theory (see Equation 2.9). In Chapters 3 and 4 we designed a perceptual version of this paradigm (see also Landy et al., 2007), in order to probe whether posterior uncertainty in even single visual quantities is taken into account in decision-making processes. The other consequence of a focus on the combination of distributions seems to have been a preponderance of studies on intermediate visual quantities such as motion direction (e.g. Weiss et al., 2002), depth cues (e.g. Jacobs, 1999), surface orientation (e.g. Saunders and Knill, 2001), or macroscopic size (e.g. Ernst and Banks, 2002). We wanted to add to our understanding of the limits of Bayesian optimality in perception, by asking whether uncertainty about very simple and very complex visual stimuli is used to guide probabilistic inference.

In Chapter 3 we used a simple Vernier offset task (Westheimer, 1979), asking observers to categorise the offset as left or right in the face of asymmetric penalties for answering "left" vs. "right" incorrectly. A Bayesian decision theory analysis was used to quantify the optimal shift of the psychometric function for different loss functions, given the observers' uncertainty and the external loss function. We found evidence for qualitative optimality in the curve shifting strategy, and for quantitative optimality in the obtained score as curves were shifted relative to a biased centre. This raises important questions about the most

meaningful way to measure functional optimality, and the constraints or biases against which optimality is defined, which will be discussed in more detail below. The contribution of this work to a clearer picture of when Bayesian inference takes place was strengthened by careful control analyses, and by the attempt to rule out the use of adaptive feedback-driven strategies. This kind of approach is critical to the utility of what is ultimately indirect evidence that populations of neurons are *doing* Bayesian inference. Observers' uncertainty was also found to be different in the two experimental sessions, and for most their strategy remained close to optimal, reinforcing the claim that online representations of uncertainty are used to guide perceptual decisions.

In Chapter 4 we applied the same paradigm to a complex, semantically rich stimulus axis, consisting of face-house mixtures running from 100% face to 100% house. Observers were asked to categorise stimuli as faces or houses under asymmetric penalties, and debriefing suggested that they were unaware of the stimulus continuum. Here we found again that observers shifted their psychometric curves in the right direction, and by an amount consistent with the slope of the psychometric function. However, performance was not quantitatively optimal. In the optimality analysis used for the Vernier data, observers should retain the same psychometric function slope as the loss function changes, allowing the slope to serve a proxy for constant sensory uncertainty. In the face-house categorisation task, the data was not well modelled by a single slope parameter for all value conditions, and we were therefore unable to predict and measure optimal relative shifts. In the next section we will discuss the limitations of the conclusions that can be drawn from such studies, but our results suggest that Bayesian optimality in perceptual inference might be more prevalent (or perhaps more evident) in the processing of simpler stimuli.

In this work we have contributed to the development of a range of behavioural tools for probing optimality at different levels of visual processing. We have also suggested that flexible, online representations of uncertainty extend down to the earliest visual processing, yet are available to higher decision-making regions, and have questioned the optimality of processing with regard to complex objects. A more distant aim was to start thinking about paradigms that are amenable to combination with electrophysiological data via intervening neural coding models. Simple tasks whose substrate is thought to lie in relatively well understood regions of early visual cortex seem like good candidates. As discussed in Section 2.2.4 of the literature review (and illustrated in Figure 1.3), the smoking gun for the BCH would consist of experiments that triangulate behavioural, electrophysiological, and modelling data. However, the links between these apices are under-constrained, and having unimodal, early visual paradigms might help to tighten them up.

7.1.2 Limitations and Future Work

This work adds one more piece to the puzzle of where and when perception is Bayesian – there is much more to be done. In Chapter 3, the evidence for representation of uncertainty was strengthened by control analyses and attempts to show that the process is intrinsic to perceptual inference rather than adaptively learnt. Further support would come from demonstrations that performance can be optimal in the face of rapid changes in perceptual uncertainty – such that any kind of adaptive strategy becomes intractable. Manipulations such as using random-dot-kinematograms, or noisy texture cues, allow you to explicitly control uncertainty, but we wanted uncertainty to arise largely from internal sensory noise over a fixed stimulus axis and so avoided this manipulation. In related work, Landy et al. (2007) did manipulate uncertainty in this way, and found a variety of suboptimal strategies. The problem with *sub*-optimality is that it is much harder to interpret – was this due to extra cognitive load or interference from memory strategies? From our failure to properly model an experimentally-irrelevant but ecologically-valid prior? Or is the combination of sensory uncertainty with information about external loss functions not as flexible and rapid as we suggest? In general, the logic of the optimality study suffers from classic issues with interpreting the null hypothesis, which are critical if we want to move away from narrowly constrained experimental domains in which optimal performance is possible.

In Chapter 4, we found suboptimal performance on a face-house categorisation task. The data suggested that observers were biased towards seeing faces, and that the loss function altered sensitivity as well as criterion – in Section 7.2 we will discuss possible sources of a value-related effect on sensitivity. A further possibility is that a Gaussian noise model is not the best representation of uncertainty in the face-house categorisation task – unlike for offset, the face-house axis does not correspond to a single, continuous physical dimension. A challenge for future work is to develop complex stimuli more naturally characterised along a single dimension, or to directly identify noise models for multi-dimensional stimuli, and then to derive optimality analyses for the resulting (potentially intractable) distributions. The loss function was changed every two trials, rather than block-by-block, which could have added additional cognitive load that interfered with the automatic processing of uncertainty. In future work, it would be interesting to investigate the factors of stimulus complexity, lability of uncertainty, and lability of value in a fully factorial design, to pull apart the circumstances in which optimality can be observed. This highlights the utility of what, for observers, are likely the more boring experiments – those involving very simple stimuli and very static task demands. Although recent work has started to configure higher-level cognitive tasks and reasoning as optimal under constraints (e.g. Griffiths and Tenenbaum, 2006; Gigerenzer, 2002), rather than as demonstrating irrational heuristics and biases (e.g. Kahneman and Tversky, 1979), it is still unclear where the boundary should

be drawn, what the contents of an ecological prior should be, and where constraints might be found. These issues are harder to avoid with more complex stimuli and tasks (see also Stocker and Simoncelli, 2008, and discussion on page 33).

This discussion highlights two contradictory drives for future work on behavioural evidence for the BCH. We have argued for developing super-simple paradigms in which a positive result has a clear implication, and where links to well-understood neural substrates and biologically-inspired neural coding models are clear. But we have also argued that we need a richer picture of where inference is Bayesian, and where it might approximate optimality. For the latter, there are a number of serious practical and interpretive barriers to be overcome. Future work looking into the influence of task design and conscious strategies would be helpful, as would the design of tasks in which the observer's task is tangential to the judgement meant to reveal uncertainty. For example, Carrasco et al. (2004) wanted to know if attention changed perceived contrast, but rather than asking observers explicitly if the attended stimulus in the display was of higher contrast, they asked them to make a tangential judgement *about the higher contrast stimulus*. Such a design minimises interference from any expectation that the attended stimulus should be of higher contrast, and an analogue could be developed to probe knowledge of uncertainty. The use of an external loss function, in which observers are characterised as using decision-theory to maximise gains, raises important questions about how the Bayesian perceptual inference paradigm can be situated in a larger picture of the decision-making brain. In general, connecting probabilistic inference to functional anatomy and cognitive theory, as well as to electrophysiological substrates of particular distributions, is an important future challenge.

7.2 Searching for Bayesian Decision Making in the Brain

7.2.1 Contributions

In the behavioural paradigm discussed above, an observer must combine knowledge of their internal uncertainty with knowledge of an externally determined loss function in order to make a decision that maximises gain (or minimises loss). This represents the convergence of economic, value-based decision making with probabilistic perceptual inference, as described by Bayesian decision theory and used for many years to discriminate perceptual sensitivity from decision criteria in signal detection theory (see page 79 and Green and Swets, 1966). However, we lack an understanding of the neural basis of this process, which

has traditionally been researched under the separate banners of value-based, and perceptual decision making (see Rangel et al., 2008; Heekeren et al., 2008, respectively).

Work in the neuroscience of value-based decision making has mapped out the various components of evaluating decision or action outcomes in cortico-striatal circuitry, and has looked at how these evaluations are made over different timescales in Pavlovian, habitual, goal-directed, and episodic learning (see Rangel et al., 2008). Work in the neuroscience of perceptual decision-making has focused on the accumulation of sensory evidence in fronto-parietal cortex, and on the role of the basal ganglia in implementing threshold crossing for diffusion-to-bound models of this accumulation process (see Gold and Shadlen, 2007). Recently, human fMRI studies have found activity in fronto-parietal regions that seems to correspond to such decision variables (e.g. Heekeren et al., 2008; Summerfield et al., 2006a), and regions of visual cortex that support the representation of stimulus categories such as faces and houses are well known (e.g. Haxby et al., 1994). However, the two have not yet come together. Sensory uncertainty, trial-by-trial task difficulty, and external value should all contribute to computation of an expected utility signal, but the underlying functional anatomy of this convergence is unknown.

The study reported in Chapter 5 involved 15 of the participants from the psychophysics study reported in Chapter 4 repeating a smaller number of trials in the scanner, with the out-of-scanner psychophysics data used to parameterise the imaging analysis. Using a face-house continuum allowed us to identify separate regions of visual cortex that responded to each stimulus category, and to ask whether changes in the external loss function affected sensory processing, or were evident only later in the decision making process. Another way of thinking about this is in terms of questioning the separability and seriality of the Bayesian decision – does the brain compute a posterior which is then combined with a value signal in a decision-making computation, for example in changing a threshold, or can the value signal also impact directly on the posterior?

Our results indicated that the effect of external value is associated with regions in a basal ganglia-prefrontal loop that has been previously implicated in the computation of EU, and is not observed in face- and house-selective regions. The difficulty of the perceptual decision was reflected in the ACC, and cumulative feedback in ventral striatum and medial PFC, again consistent with analogous components of value-based decision making. Changes in sensory uncertainty were however reflected in FFA, suggestive of a posterior representation in sensory cortex that is transmitted to an action selection mechanism alongside value signals, rather than being modified by them. However, whether these changes in sensitivity were directly mediated by value, as opposed to being an indirect effect of increased attention in asymmetric value conditions, is unclear (see Pleger et al., 2008; Simoncini and Baldassi, 2008, for suggestions that a value-based effect is separable from an attentional one). An

architecture in which criterion shifts are implemented at a later decision-making stage might be advantageous in a world where particular sensory qualities need to be flexibly associated with different values – flexible action selection requires control mechanisms that do not solely depend on evidence accumulation (Stafford and Gurney, 2007; Maloney, 2002). Previous studies have observed effects of attention, task set, or motivational state in sensory cortex (see page 123), and the extent of such top-down influences is not fully understood. It seems plausible that improvements in sensitivity, such as that we observed here (see also Pleger et al., 2008; Simoncini and Baldassi, 2008), would be advantageous whatever the source of the signal, but that a biasing effect of value is better implemented at a later stage. In general we found a high degree of overlap between the correlates of external value in our study, and those seen in value-based decision making studies, though there was a hint that fronto-parietal decision-variables are more readily reflected in perceptual tasks.

7.2.2 Limitations and Future Work

There are two caveats to our conclusion that direct effects of value are reflected in fronto-striatal decision circuitry, whilst indirect effects via changes in sensitivity might be observed in sensory regions. The first is the spatio-temporal resolution of the BOLD signal – there might be subtle effects of value in sensory cortex, for example if the effect is not reflected in increases or decreases in average activity. Future work should approach this problem in two ways – first, by supplementing human imaging with electrophysiological recordings from different hypothesised components of the decision making circuitry, as in the preliminary results reported by Ding and Gold (2008). Second, we could theoretically extend a neural coding model to map population codes that implement the probabilistic inference observed in behaviour to the BOLD signal, though whether this would have the fidelity to identify optimality and tell between different competing coding models is an open question.

The second major limitation of this study as evidence for the BCH is clearly the lack of behavioural optimality – we are not looking for the anatomical correlates of Bayes *optimal* behaviour, rather we are starting to lay a functional anatomy for the processes involved in Bayesian perceptual decision-making. If behaviour had been optimal, the conclusions would still be limited by the methodological caveats described in the previous paragraph, but we would have been able to explicitly separate the effects of value and uncertainty in the psychophysical parameters. Unless the assumptions of the optimality analysis are met, the observer's sensitivity or psychometric slope does not serve as a direct proxy for sensory uncertainty, and therefore distinguishing direct effects of value on criterion and indirect effects on sensitivity is less straightforward. In the future, obtaining optimal performance would strengthen these conclusions, and comparing participants who do and don't perform

optimally might be interesting in terms of untangling exactly underlies suboptimal inference – a problem we raised above in Section 7.1.2. In addition, explicitly manipulating uncertainty would enable us to construct an uncertainty regressor that could be orthogonalised against value. To further explore the functional homology of the Bayesian decision theory formalism with neuroanatomy, it would also be interesting to manipulate a biased prior and ask whether its effects are restricted, unlike those of external value, to modulating sensory posterior representations (see Summerfield and Koechlin, 2008).

There are a multitude of questions still remaining in the neuroscience of decision-making, concerning how different valuation systems compete for control of action, how action outcomes are selected for representation and comparison, and how the machinery of evidence accumulation is impacted by various valuation components (see Rangel et al., 2008; Heekeren et al., 2008, for discussion). As has been demonstrated with RL models of how animals acquire action-value contingencies (see Sutton and Barto, 1998), comparing computational models of probabilistic inference to electrophysiological recordings (Montague et al., 1996) and functional imaging data (e.g. Hare et al., 2008) can be a powerful way to delineate correlates of component processes. However, these models have parameters that change over time as contingencies change, whereas we hypothesise that the integration of value with uncertainty occurs within a single trial and the parameters of this process are static throughout the session. Using learning methodologies might be another way (alongside the external manipulations of uncertainty discussed above) of improving the fidelity with which we could identify correlates of posterior uncertainties, external value, and biased priors.

Predictive coding, like inverse inference, is a concept that can be applied to the brain on many levels of analysis – as discussed above (see page 46), the idea that perception can be thought of as the comparison between what was expected and what was observed date back to the 1950s (see MacKay, 1956; Lee and Mumford, 2003). Prediction of outcomes, and the comparison of anticipated states to what actually unfolds, has clear benefits for an organism in terms of flexibility, learning, and preparedness for action (see e.g. Schutz-Bosbach and Prinz, 2007; Mehta, 2001). The question of how this is reflected on various levels of neural encoding and inference, and how this can be reconciled with representational codes, is very important. Looking at how the anatomical correlates of perceptual decision-making maps onto those of economic, value-based choice, is one way to approach this question, and Bayesian models of the underlying computations could lend important specificity to this process.

7.3 Bringing Bayes to Attention

7.3.1 Contributions

The theoretical work reported in Chapter 6 had two main aims – the first was to consider more complex scenarios than those typically used to provide evidence for the BCH, which tend to involve at most a small handful of objects in the focus of attention. This necessitates a consideration of approximate probabilistic representation and inference, in circumstances where the brain cannot be fully optimal, and therefore evokes some of the practical issues with interpreting behavioural suboptimality discussed in Section 7.1.2. It also evokes issues of selective attention, and improvements in processing due to top-down signals, which have not been much considered in the context of the BCH (though see Dayan and Zemel, 1999). The foundation for a more principled approach to richer inferential settings is to develop a probabilistic notation for approximate perceptual inference, and for the role of attention in locally improving these approximations. We develop such a notation, for a framework in which the brain is forced to represent approximations to highly correlated, joint posteriors over the multitude of spatially distributed features that make up real world scenes. This picture evokes problems of approximate inference common in machine learning, and with the factored approximation we use to demonstrate the framework, resonates with notions of receptive fields and cortical specialisation. The approximation process is then embedded in a framework in which attention acts as an extra hypothesis to locally reduce noise, and selectively reveal some of the neglected correlations, extending ideas of attention as a Bayesian prior (see Dayan and Zemel, 1999; Rao, 2005).

The second aim of our attentional framework was to give shape to the idea that there is a common limited capacity resource underlying the diverse range of limitations evidenced in behaviour (see e.g. Itti et al., 2005; Driver, 2001). By giving a computational level description that admits of degrees rather than distinctions, and which supports diverse implementations of a representational capacity limit and of attentional improvement, we hope to unify apparently disparate results. The potential of this approach was demonstrated via simulations of inference in an abstract 'grid world', corresponding to three key attentional paradigms – pre-cueing, response-cueing, and illusory conjunctions.

In the pre-cueing and response-cueing simulations, the attentional hypothesis acted as a prior, performing uncertainty reduction by ruling out regions of the posterior outside of its mode. When directed towards *good* explanations by valid instructions or precues, this improves judgements, and when directed towards explanations not supported by the posterior this damages performance – concordant with behavioural evidence (e.g. Luck et al.,

1996) and with models of competitive interactions biased by attention (see e.g. Desimone and Duncan, 1995). Importantly, our approach to attention as a hypothesis with the mathematical form but not the semantics of a prior makes it appropriate for modelling scenarios in which the attentional signal does not strictly carry prior information (Pestilli et al., 2007; Rao, 2005). In the illusory conjunction paradigm, the attentional hypothesis acted to selectively reveal positive correlations between colocated features neglected in the factored approximation. Here, the attentional hypothesis represents a continuous, probabilistic analogue of the FIT spotlight of attention that was so troublesomely all-or-none, and configures binding as a behaviourally defined notion supported by a continuous underlying process.

Using a probabilistic framework also helps to extend existing analyses of the consequences of 'noise' for attentional selection. The Bayesian observer is, as discussed throughout the thesis, the probabilistic inferential analogue of the SDT observer (see page 21). The SDT approach has been instrumental in challenging the interpretation of visual search tasks as revealing the serial allocation of a limited capacity resource – showing where set-size effects can be explained simply in terms of increasing uncertainty (e.g. Palmer et al., 1993; Eckstein et al., 2000). By moving to a probabilistic representation we preserve this insight, whilst also describing a role for attention in the performance decrements left unexplained by SDT models (see e.g. Palmer, 1994; Shaw, 1984). A related debate has been between characterisations of the role of attention in terms of reducing uncertainty, and in terms of directly enhancing the signal even in the absence of uncertainty (see e.g. Cameron et al., 2002). In our framework, attention reduces uncertainty, resulting in improved judgements about both location and feature value. We suggested that this might be the case even when spatial uncertainty is not apparent in behaviour, producing what could appear to be 'pure' signal enhancement.

7.3.2 Limitations and Future Work

Future work on the simple model presented in Chapter 6 will apply it to a wider range of attentional paradigms, considering cases with multiple, spatially extended objects. Introducing dynamics into the allocation of the attentional hypothesis will also enable us to explore visual search tasks and natural viewing, informed by neural and behavioural data on the parameters of this process. Future work will also continue the investigation into the relationship between noise reduction and signal enhancement, delineating the different contributions to uncertainty, and how dealing with uncertainty might in fact result in signal enhancement under certain definitions (see Lu and Dosher, 1998, for a similar approach in the SDT framework). Through probabilistic models in which signal enhancement is explicitly defined, we also hope to address the relationship between behavioural contrast-response

functions and the underlying mechanism of attention (see Cameron et al., 2002; Ling and Carrasco, 2006). By building neural coding models that implement the probabilistic representations (see below), *neurophysiological* C-R functions might also be better understood (e.g. Williford and Maunsell, 2006; Li et al., 2008).

The overarching aim of this work was to present a novel framework for thinking about capacity limitations and the role of attention in resolving them, translating restricted cognitive notions into a multiply instantiated computational description. As such, the project is at a relatively early stage, and lays the groundwork for future work building detailed, biologically-inspired models of particular attentional paradigms which yield more tightly constrained, testable hypotheses. In Section 6.4 we considered the challenges for such modelling, which were both technical and theoretical. Most immediately, moving to posteriors of real-world complexity will require machine learning methods for approximate inference and learning still under development.

On the theoretical side, great efforts will have to be made to tether simulations to a web of constraints. As for many other computational models of attention, we have to choose a mapping from a representation computed with or without attention to a behavioural decision. This mapping should be constrained by behavioural evidence and decision-making models, and wherever possible the impact of different mappings should be compared. Relatedly, the framework does not consider the role of mechanisms external to the posterior representation, for example, salience, arousal or memory. We discussed the possibility of 'bolting on' components such as bottom-up salience, much as salience approaches have 'bolted on' top-down signals (see e.g. Peters and Itti, 2007). It might instead be interesting to try and translate salience computations into the probabilistic framework presented here, to present a more unified picture. Within the machinery of the framework itself, the true prior, the approximate posterior, and the attentional hypothesis present a further group of free parameters that must be reined in by empirical constraints. This is very difficult for the kind of abstract tasks we model in Chapter 6, where the ability to qualitatively match behavioural data should be interpreted as a demonstration of the principles of the framework rather than a direct challenge to other, more detailed models of those particular data sets. In the future, detailed biologically-inspired models will provide stronger constraints – for example, the RF size and tuning curve of neurons thought to represent the relevant feature should provide constraints on the form of the approximate posterior.

Building biologically-constrained probabilistic models of particular tasks with well understood neural substrates is the first step – the next will be to build population coding models of the probabilistic inference that can be directly compared to relevant electrophysiological data. This will require further development of population codes such as the DDPC approach of Sahani and Dayan (2003), and the challenges for methodological tri-

angulation discussed in the literature review will apply to attempts to link these codes back into electrophysiological and behavioural data (see Figure 1.3 and Section 2.2.4). We also made some tentative speculations about how our framework might map onto the functional neuroanatomy of the brain, including the passage of approximate posteriors through a hierarchy of representations – considering how our framework might relate to models of empirical Bayesian inference in hierarchical networks (e.g. Friston, 2005; Lee and Mumford, 2003; Rao and Ballard, 1999), and to predictive coding schemes (see e.g. Friston, 2005), is another important avenue for future work.

7.4 Final Thoughts

The Bayesian Coding Hypothesis states that perception and action are not only well-described by Bayes' rule, but that the brain actually implements the probabilistic representation and computation implied by these descriptions (see Knill and Pouget, 2004; Doya et al., 2007). It is an excitingly general hypothesis that can be applied on many levels of explanation, ranging from encoding models to describe spike trains in the retina to descriptions of the embodied brain in action. In this thesis we have considered Bayes' rule as a description of perceptual inference, coupled with the assertion that neural coding models of such inference can be usefully mapped to electrophysiological data. We have focused on expanding our understanding of the circumstances in which perceptual inference is Bayes-optimal, on linking Bayesian approaches to existing literatures on decision-making and selective attention, and on descriptions of approximate inference in more realistic domains.

Above, we discussed future challenges for each of these endeavours, which can be summarised in terms of constraining each link in the integrative methodology illustrated in Figure 1.3 (see page 24). Many of these challenges are practical, and progress is likely to be cumulative and piecemeal. But as the remit of the BCH is expanded to more complex domains, invoking all the elements of cognition along the way, some more fundamental questions may also demand attention. One we have touched on in different ways throughout this thesis is where the boundary lies between sub-optimal inference, and inference that is optimal under constraints or ecologically valid priors. This debate appears in various guises in a wide range of domains – in terms of the informational capacity of neural representation, low-level perception and action, high level cognitive and economic reasoning, and even evolution (Marcus, 2008). This question is particularly acute for behavioural optimality paradigms, which provide important evidence for the BCH. However, with an increasing focus on approximate and hierarchical inference the question of when a brain is optimal under strict constraints as opposed to failing to utilise (or failing to evolve) resources in

what we define as an optimal way might come to seem a strangely anthropomorphic one. Perhaps the ability of the probabilistic approach to challenge hard-to-define dichotomies in the attentional literature might also manifest itself on this more general level. It remains to be seen how distinct an *approximate* Bayesian neural code would be from alternative, non-probabilistic representations.

The larger the scale of the object of study, the harder it is to pin down the components of a probabilistic model. Turning the crank of the Bayesian machine is easy, but as philosophers of science found when they tried to use Bayesian inference to define scientific reasoning, without principled methods for defining the ingredients it can be a dangerously flexible descriptive tool rather than a normative or predictive one (see Chalmers, 1999). For the amorphous 'beliefs' behind any particular scientific enterprise, this problem appears insurmountable, but for perceptual inference there is enough empirical evidence to provide convincing constraints. However, whilst there is something profoundly appropriate about Bayes rule as a description of perceptual reasoning, there is also something profoundly difficult about using perceptual behaviour as evidence for its use. Humans co-opt a vast array of computation in the service of single decisions and actions, and automatic or unconscious processing interacts with conscious biases and strategies. Developing new behavioural tools for probing probabilistic inference – be it restricted and optimal, or broad and approximate – is crucial. And tying behavioural evidence into an integrative trinity with neural data and neural coding models pushes such models away from the pleasingly descriptive towards the powerfully explanatory. Bayes' rule provides us with a multifocal lens through which neural operations on many scales can be better understood. Discovering how far neurons actually speak a Bayesian tongue is a fascinating and fundamental question.

Notations and Abbreviations

Probabilistic Abbreviations

BBH *Bayesian Brain Hypothesis* 16

BCH *Bayesian Coding Hypothesis* 16

BMC *Bayesian model comparison* 85

BP *Belief propagation* 151

EP *Expectation propagation* 151

MAP *Maximum a Posteriori estimate* 14

ML *Maximum Likelihood estimate* 14

pdf *Probability density function* 22

POMDP *Partially observable Markov decision process* 51

PPC *Probabilistic population code* 36

DDPC *Doubly distributional population code* 43

SPRT *Sequential probability ratio test* 57

Theoretical Terminology

2D *Two Dimensional* 17

3D *Three Dimensional* 17

c *Signal detection criterion* 129

d' *Signal detection sensitivity* 129

EU *Expected Utility* 18

PT *Prospect Theory* 55

FIT *Feature integration theory* 69

GLM *General linear model* 125

KL *Kullback-Leibler divergence* 150

RL *Reinforcement learning* 61

SDT *Signal Detection Theory* 20

Methodological Terminology

AR(1) *First-order auto-regressive moving average model* 125

BOLD *Blood-oxygen level dependent* 56

C-R *Contrast-response function* 157

CP *Choice probability function* 125

D *Difficulty function* 125

D_{stim} *Stimulus difficulty* 56

EPI *Echo-planar imaging sequence* 124

FIR *Finite impulse response function* 126

fMRI *Functional magnetic resonance imaging* 26

FWE *Family-wise error correction* 126

HRF *Haemodynamic response function* 124

MNI *Montreal Neurological Institute template* 125

FV *Face value condition* 107

HV *House value condition* 107

NV *Neutral value condition* 107

RDK *Random dot kinematogram* 37

RT *Reaction time* 57

ROI *Region of Interest* 126

SNR *Signal-to-noise ratio* 36

SOA *Stimulus onset asynchrony* 107

SPM *Statistical Parametric Mapping software* 125

T1 *The time constant of recovery of longitudinal magnetization* 124

TE *Echo time in an EPI sequence* 124

TR *Repetition time in an EPI sequence* 124

Brain Areas and Neural Properties

ACC *Anterior cingulate cortex* 54

dACC *Dorsal anterior cingulate cortex, or paracingulate gyrus* 136

BG *Basal ganglia* 55

FEF *Frontal eye fields* 44

FFA *Fusiform face area* 56

GP *Globus pallidus* 59

IFG *Inferior frontal gyrus* 136

IFS *Inferior frontal sulcus* 133

IOG *Intra-occipital gyrus* 130

IPS *Intraparietal sulcus* 62

LIP *Lateral intraparietal area* 37

M1 *Primary motor cortex* 122

MT *Medial temporal visual area* 48

OFC *Orbitofrontal Cortex* 54

PFC *Prefrontal cortex* 54

dlPFC *Dorsolateral prefrontal cortex* 54

mPFC *Medial prefrontal cortex* 140

PMd *Premotor cortex* 133

PPA *Parahippocampal place are* 56

PSE *Point of subjective equality* 106

RF *Receptive field* 37

ROC *Receiver operating characteristic* 77

SC *Superior colliculus* 59

SEF	*Supplementary eye fields*	62
STN	*Substantia nigra*	59
V1	*Striate cortex*	39

Bibliography

Afraz, S. R., Kiani, R., and Esteky, H. (2006). Microstimulation of inferotemporal cortex influences face categorization. *Nature*, 442(7103):692–695.

Ahrens, M. B., Linden, J. F., and Sahani, M. (2008). Nonlinearities and contextual influences in auditory cortical responses modeled with multilinear spectrotemporal methods. *Journal of Neuroscience*, 28(8):1929–1942.

Alais, D. and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3):257–262.

Allport, A. (1989). Visual attention. In Posner, M. I., editor, *Foundations of Cognitive Science*. MIT Press, Cambridge, Mass.

Anastasio, T. J., Patton, P. E., and Belkacem-Boussaid, K. (2000). Using Bayes' rule to model multisensory enhancement in the superior colliculus. *Neural Computation*, 12(5):1165–1187.

Anderson, C. (1994). Neurobiological computational systems. In Marks, R. J., Zurada, J. M., and Robinson, C. J., editors, *Computational intelligence imitating life*, pages 213–222. IEEE Press, New York, NY.

Anderson, C. and Abrahams, E. (1987). The Bayes connection. In *Proceedings of the IEEE First International Conference on Neural Networks*, volume 3, pages 105–112, San Diego, CA. SOS Print.

Aron, A. R., Behrens, T. E., Smith, S., Frank, M. J., and Poldrack, R. A. (2007). Triangulating a cognitive control network using diffusion-weighted magnetic resonance imaging (MRI) and functional MRI. *Journal of Neuroscience*, 27(14):3743–3752.

Ashby, F. G., Prinzmetal, W., Ivry, R., and Maddox, W. T. (1996). A formal theory of feature binding in object perception. *Psychological Review*, 103(1):165–192.

Atkins, J. E., Fiser, J., and Jacobs, R. A. (2001). Experience-dependent visual cue integration based on consistencies between visual and haptic percepts. *Vision Research*, 41(4):449–461.

Atkins, J. E., Jacobs, R. A., and Knill, D. C. (2003). Experience-dependent visual cue recalibration based on discrepancies between visual and haptic percepts. *Vision Research*, 43(25):2603–2613.

Augustine, J. R. (1996). Circuitry and functional aspects of the insular lobe in primates including humans. *Brain Research Reviews*, 22(3):229–244.

Averbeck, B. B., Latham, P. E., and Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7(5):358–366.

Balan, P. F., Oristaglio, J., Schneider, D. M., and Gottlieb, J. (2008). Neuronal correlates of the set-size effect in monkey lateral intraparietal area. *PLoS Biology*, 6(7):e158.

Baldassi, S. and Burr, D. C. (2000). Feature-based integration of orientation signals in visual search. *Vision Research*, 40(10-12):1293–1300.

Baldassi, S. and Burr, D. C. (2004). "Pop-out" of targets modulated in luminance or colour: the effect of intrinsic and extrinsic uncertainty. *Vision Research*, 44(12):1227–1233.

Balleine, B. W. (2005). Neural bases of food-seeking: affect, arousal and reward in cortico-striato-limbic circuits. *Physiology & Behaviour*, 86(5):717–730.

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

Balleine, B. W., Doya, K., O'Doherty, J., and Sakagami, M. (2008). *Reward and decision making in corticobasal ganglia networks*. Wiley Blackwell, New York, NY.

Barber, M. J., Clark, J. W., and Anderson, C. H. (2003). Neural representation of probabilistic information. *Neural Computation*, 15(8):1843–1864.

Barlow, H. B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception*, 1(4):371–394.

Barlow, H. B. (1990). *Vision: Coding and efficiency*. Cambridge University Press, Cambridge, UK.

Barraclough, D. J., Conroy, M. L., and Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience*, 7(4):404–410.

Battaglia, P. W., Jacobs, R. A., and Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, 20(7):1391–1397.

Bayes, T. (1764). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*.

Bays, P. M. and Wolpert, D. M. (2007). Computational principles of sensorimotor control that minimize uncertainty and variability. *Journal of Physiology*, 578(2):387–396.

Beal, M. (2003). *Variational algorithms for approximate Bayesian inference*. PhD thesis, University College London.

Beck, D. M. and Kastner, S. (2008). Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision Research*, In press.

Behrens, T. E., Johansen-Berg, H., Woolrich, M. W., Smith, S. M., Wheeler-Kingshott, C. A., Boulby, P. A., Barker, G. J., Sillery, E. L., Sheehan, K., Ciccarelli, O., Thompson, A. J., Brady, J. M., and Matthews, P. M. (2003). Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. *Nature Neuroscience*, 6(7):750–757.

Berger, J. O. (1985). *Statistical decision theory and bayesian analysis*. Springer Verlag, New York, NY.

Bertelson, P. and Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception & Psychophysics*, 29(6):578–584.

Bertelson, P., Vroomen, J., de Gelder, B., and Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception & Psychophysics*, 62(2):321–332.

Bestmann, S., Harrison, L. M., Blankenburg, F., Mars, R. B., Haggard, P., Friston, K. J., and Rothwell, J. C. (2008). Influence of uncertainty and surprise on human corticospinal excitability during preparation for action. *Current Biology*, 18(10):775–780.

Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R., and Warland, D. (1991). Reading a neural code. *Science*, 252(5014):1854–1857.

Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A., and Ward, B. D. (2004). Neural correlates of sensory and decision processes in auditory object identification. *Nature Neuroscience*, 7(3):295–301.

Birn, R. M., Diamond, J. B., Smith, M. A., and Bandettini, P. A. (2006). Separating respiratory-variation-related fluctuations from neuronal-activity-related fluctuations in fMRI. *Neuroimage*, 31(4):1536–1548.

Bogacz, R. (2007). Optimal decision-making theories: Linking neurobiology with behaviour. *Trends in Cognitive Sciences*, 11(3):118–125.

Bogacz, R. and Gurney, K. (2007). The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural Computation*, 19(2):442–477.

Botvinick, M. M. (2007). Conflict monitoring and decision making: Reconciling two perspectives on anterior cingulate function. *Cognitive & Affective Behavioural Neuroscience*, 7(4):356–366.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, 10(4):433–436.

Brefczynski, J. A. and DeYoe, E. A. (1999). A physiological correlate of the 'spotlight' of visual attention. *Nature Neuroscience*, 2(4):370–374.

Breiter, H. C., Aharon, I., Kahneman, D., Dale, A., and Shizgal, P. (2001). Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron*, 30(2):619–639.

Bresciani, J. P., Dammeier, F., and Ernst, M. O. (2006). Vision and touch are automatically integrated for the perception of sequences of events. *Journal of Vision*, 6(5):554–564.

Brett, M., Anton, J. L., Valabregue, R., and Poline, J. B. (2002). Region of interest analysis using an SPM toolbox. *Neuroimage*, 16(2):abstract 497.

Broadbent, D. E. (1958). *Perception and communication*. Pergamon, Oxford, UK.

Brown, E. N., Frank, L. M., Tang, D., Quirk, M. C., and Wilson, M. A. (1998). A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18(18):7411–7425.

Bryson, A. E. and Ho, Y. C. (1975). *Applied optimal control*. Wiley, New York, NY.

Butler, B. E., Mewhort, D. J., and Browse, R. A. (1991). When do letter features migrate? A boundary condition for feature-integration theory. *Perception and Psychophysics*, 49(1):91–99.

Cameron, E. L., Tai, J. C., and Carrasco, M. (2002). Covert attention affects the psychometric function of contrast sensitivity. *Vision Research*, 42(8):949–967.

Cameron, E. L., Tai, J. C., Eckstein, M. P., and Carrasco, M. (2004). Signal detection theory applied to three visual search tasks - identification, yes/no detection and localization. *Spatial Vision*, 17(4-5):295–325.

Carpenter, R. H. S. (1988). *The movements of the eyes*. Pion, London, UK.

Carrasco, M. (2005). Transient covert attention increases contrast sensitivity and spatial resolution: Support for signal enhancement. In Itti, L., Rees, G., and Tsotsos, J. K., editors, *The neurobiology of attention*. Elsevier Academic Press, Oxford, UK.

Carrasco, M., Evert, D. L., Chang, I., and Katz, S. M. (1995). The eccentricity effect - target eccentricity affects performance on conjunction searches. *Perception & Psychophysics*, 57(8):1241–1261.

Carrasco, M., Ling, S., and Read, S. (2004). Attention alters appearance. *Nature Neuroscience*, 7(3):308–313.

Cavanagh, P. and Alvarez, G. (2005). Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences*, 9(7):349–354.

Chalmers, A. F. (1999). *What is this thing called science?* Open University Press, 3rd edition.

Cherry, E. C. (1953). Some experiments on the recognition of speech with one and with two ears. *Journal of the Acoustical Society of America*, 25:975–979.

Churchland, A. K., Kiani, R., and Shadlen, M. N. (2008). Decision-making with multiple alternatives. *Nature Neuroscience*, 11(6):693–702.

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(1485):1585–1599.

Clarke, J. J. and Yuille, A. L. (1990). *Data fusion for sensory information processing systems.* Kluwer Academic Publishers, Dordrecht, Netherlands.

Cohen, A. and Ivry, R. (1989). Illusory conjunctions inside and outside the focus of attention. *Journal of Experimental Psychology: Human Perception and Performance*, 15(4):650–663.

Cohen, J. D., McClure, S. M., and Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, 362(1481):933–942.

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

Cox, R. T. (1961). *The algebra of probable inference.* Johns Hopkins University Press, Baltimore, MD.

Critchley, H. D., Mathias, C. J., and Dolan, R. J. (2001). Neural activity in the human brain relating to uncertainty and arousal during anticipation. *Neuron*, 29(2):537–545.

Dale, A. I. (1982). Bayes or Laplace? an examination of the origin and early application of Bayes' theorem. *Archive for the History of the Exact Sciences*, 27(1):23–47.

Davis, K. D., Taylor, S. J., Crawley, A. P., Wood, M. L., and Mikulis, D. J. (1997). Functional MRI of pain- and attention-related activations in the human cingulate cortex. *Journal of Neurophysiology*, 77(6):3370–3380.

Davison, M. C. and Tustin, R. D. (1978). The relation between the generalized matching law and signal-detection theory. *Journal of the Experimental Analysis of Behaviour*, 29(2):331–336.

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

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

Dayan, P. (1994). Computational modelling. *Current Opinion in Neurobiology*, 4(2):212–217.

Dayan, P. (2008a). Loads of attentional Bayes. *Advances in Neural Information Processing Systems*, In press.

Dayan, P. (2008b). The role of value systems in decision-making. In Singer, W., editor, *Better than conscious? Implications for performance and institutional analysis*, pages 51–70. MIT Press, Ernst Strüngmann Forum, Cambridge, MA.

Dayan, P. and Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. MIT Press, Cambridge, MA.

Dayan, P. and Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive & Affective Behavioural Neuroscience*, in press.

Dayan, P., Hinton, G. E., Neal, R., and Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation*, 7(5):1022–1037.

Dayan, P., Niv, Y., Seymour, B., and Daw, N. D. (2006). The misbehavior of value and the discipline of the will. *Neural Networks*, 19(8):1153–1160.

Dayan, P. and Zemel, R. (1999). Statistical models and sensory attention. *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, pages 1017–1022.

de Araujo, I. E., Rolls, E. T., Velazco, M. I., Margot, C., and Cayeux, I. (2005). Cognitive modulation of olfactory processing. *Neuron*, 46(4):671–679.

De Martino, B., Kumaran, D., Seymour, B., and Dolan, R. J. (2006). Frames, biases, and rational decision-making in the human brain. *Science*, 313(5787):684–687.

deCharms, R. C. and Zador, A. (2000). Neural representation and the cortical code. *Annual Review of Neuroscience*, 23:613–647.

Deneve, S. (2008a). Bayesian spiking neurons I: Inference. *Neural Computation*, 20(1):91–117.

Deneve, S. (2008b). Bayesian spiking neurons II: Learning. *Neural Computation*, 20(1):118–145.

Deneve, S., Latham, P. E., and Pouget, A. (2001). Efficient computation and cue integration with noisy population codes. *Nature Neuroscience*, 4(8):826–831.

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, 353(1373):1245–1255.

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

Deutsch, J. A. and Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review*, 87:272–300.

Ding, L. and Gold, J. I. (2008). Caudate activity in a decision-making reaction time task. In *Computational and Systems Neuroscience (CoSyNe) Abstracts*, page 222, Salt Lake City, UT.

Ditterich, J., Mazurek, M. E., and Shadlen, M. N. (2003). Microstimulation of visual cortex affects the speed of perceptual decisions. *Nature Neuroscience*, 6(8):891–898.

Dolan, R. J., Fink, G. R., Rolls, E., Booth, M., Holmes, A., Frackowiak, R. S., and Friston, K. J. (1997). How the brain learns to see objects and faces in an impoverished context. *Nature*, 389(6651):596–599.

Dosher, B. A. and Lu, Z. L. (2000). Noise exclusion in spatial attention. *Psychological Science*, 11(2):139–146.

Doya, K., Ishii, S., Pouget, A., and Rao, R. P. N., editors (2007). *Bayesian brain: Probabilistic approaches to neural coding*. Computational Neuroscience. MIT Press, Cambridge, MA.

Dragoi, V. and Lockhead, G. (1999). Context-dependent changes in visual sensitivity induced by Müller-Lyer stimuli. *Vision Research*, 39(9):1657–1670.

Dreher, J. C., Kohn, P., and Berman, K. F. (2006). Neural coding of distinct statistical properties of reward information in humans. *Cerebral Cortex*, 16(4):561–573.

Driver, J. (2001). A selective review of selective attention research from the past century. *British Journal of Psychology*, 92(1):53–78.

Duncan, J. (1980). The locus of interference in the perception of simultaneous stimuli. *Psychological Review*, 87(3):272–300.

Duncan, J. and Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, 96(3):433–458.

Eckstein, M. P. (1998). The lower efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science*, 2(2):111–118.

Eckstein, M. P., Thomas, J. P., Palmer, J., and Shimozaki, S. S. (2000). A signal detection model predicts the effects of set size on visual search accuracy for feature, conjunction, triple conjunction, and disjunction displays. *Perception & Psychophysics*, 62(3):425–451.

Egner, T. and Hirsch, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nature Neuroscience*, 8(12):1784–1790.

Egner, T., Monti, J. M. P., Trittschuh, E. H., Wieneke, C. A., Hirsch, J., and Mesulam, M. M. (2008). Neural integration of top-down spatial and feature-based information in visual search. *Journal of Neuroscience*, 28(24):6141–6151.

Eliasmith, C. and Anderson, C. H. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. Computational Neuroscience. MIT, Cambridge, MA.

Engel, A. K., Fries, P., and Singer, W. (2001). Dynamic predictions: oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience*, 2(10):704–716.

Engel, A. K., Knig, P., Kreiter, A. K., Chillen, T. B., and Singer, W. (1992). Temporal coding in the visual cortex: New vistas on integration in the nervous system. *Trends in Neurosciences*, 15(6):218–225.

Epstein, R. and Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676):598–601.

Ermentrout, G. B., Galan, R. F., and Urban, N. N. (2008). Reliability, synchrony and noise. *Trends in Neurosciences*, 31(8):428–434.

Ernst, M. O. (2005). A Bayesian view on multimodal cue integration. In Knoblich, G., Thornton, I., Grosjean, M., and Shiffrar, M., editors, *Human body perception from the inside out*, pages 105–131. Oxford University Press, New York, NY.

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision*, 7(5):7 1–14.

Ernst, M. O. (2008). Multisensory integration: A late bloomer. *Current Biology*, 18(12):R519–521.

Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433.

Faisal, A. A., Selen, L. P., and Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, 9(4):292–303.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press, Stanford, CA.

Fienberg, C. E. (2006). When did Bayesian inference become "Bayesian"? *Bayesian Analysis*, 1(1):1–40.

Fine, I. and Jacobs, R. A. (1999). Modeling the combination of motion, stereo, and vergence angle cues to visual depth. *Neural Computation*, 11(6):1297–1330.

Foldiak, P. (1993). The ideal homunculus: Statistical inference from neural population responses. In Eeckman, F. and Bower, J., editors, *Computation and Neural Systems*, pages 55–60. Kluwer Academic Publishers, Dordrecht, Netherlands.

Frank, M. J. (2006). Hold your horses: A dynamic computational role for the subthalamic nucleus in decision making. *Neural Networks*, 19(8):1120–1136.

Fries, P., Reynolds, J. H., Rorie, A. E., and Desimone, R. (2001). Modulation of oscillatory neuronal synchronization by selective visual attention. *Science*, 291(5508):1560–1563.

Fries, P., Womelsdorf, T., Oostenveld, R., and Desimone, R. (2008). The effects of visual stimulation and selective visual attention on rhythmic neuronal synchronization in macaque area V4. *Journal of Neuroscience*, 28(18):4823–4835.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, 16(9):1325–1352.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society Series B: Biological Sciences*, 360(1456):815–836.

Friston, K. J., Jezzard, P., and Turner, R. (1994). Analysis of functional MRI time-series. *Human Brain Mapping*, 2(1-2):69–78.

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review*, 96(2):267–314.

Geisler, W. S. and Chou, K. (1995). Separation of low-level and high-level factors in complex tasks: Visual search. *Psychological Review*, 102(2):356–378.

Gepshtein, S., Burge, J., Ernst, M. O., and Banks, M. S. (2005). The combination of vision and touch depends on spatial proximity. *Journal of Vision*, 5(11):1013–1023.

Ghahramani, Z. and Wolpert, D. M. (1997). Modular decomposition in visuomotor learning. *Nature*, 386(6623):392–395.

Ghahramani, Z., Wolpert, D. M., and Jordan, M. I. (1995). Computational structure of coordinate transformations: A generalization study. *Advances in Neural Information Processing Systems*, 7.

Ghose, G. M. and Maunsell, J. (1999). Specialized representations in visual cortex: a role for binding? *Neuron*, 24(1):79–85.

Ghose, G. M. and Maunsell, J. H. (2008). Spatial summation can explain the attentional modulation of neuronal responses to multiple stimuli in area V4. *Journal of Neuroscience*, 28(19):5115–5126.

Gigerenzer, G. (2002). *Bounded rationality: The adaptive toolbox*. Dahlem Workshop Reports. MIT Press, Cambridge, MA.

Gilbert, C., Ito, M., Kapadia, M., and Westheimer, G. (2000). Interactions between attention, context and learning in primary visual cortex. *Vision Research*, 40(10-12):1217–1226.

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

Glimcher, P. W. (2005). Indeterminacy in brain and behavior. *Annual Review of Psychology*, 56:25–56.

Glimcher, P. W. and Rustichini, A. (2004). Neuroeconomics: The consilience of brain and decision. *Science*, 306(5695):447–452.

Gold, J. I. and Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, 5(3):134–134.

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

Gollisch, T. and Meister, M. (2008). Rapid neural coding in the retina with relative spike latencies. *Science*, 319(5866):1108–1111.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., and Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1):108–154.

Gori, M., Del Viva, M., Sandini, G., and Burr, D. C. (2008). Young children do not integrate visual and haptic form information. *Current Biology*, 18(9):694–698.

Gottfried, J. A., O'Doherty, J., and Dolan, R. J. (2002). Appetitive and aversive olfactory learning in humans studied using event-related functional magnetic resonance imaging. *Journal of Neuroscience*, 22(24):10829–10837.

Gould, I. C., Wolfgang, B. J., and Philip, P. L. (2007). Spatial uncertainty explains exogenous and endogenous attentional cuing effects in visual signal detection. *Journal of Vision*, 7(13):2, 1–17.

Gray, C. M. (1999). The temporal correlation hypothesis of visual feature integration: Still alive and well. *Neuron*, 24(1):31–47.

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. Peninsula Publishing, Los Altos, CA.

Green, J. J., Conder, J. A., and McDonald, J. J. (2008). Lateralized frontal activity elicited by attention-directing visual and auditory cues. *Psychophysiology*, 45(4):579–587.

Green, J. J. and McDonald, J. J. (2008). Electrical neuroimaging reveals timing of attentional control activity in human brain. *PLoS Biology*, 6(4):e81.

Griffiths, T. L. and Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, 17(9):767–773.

Griffiths, T. L. and Tenenbaum, J. B. (2007). From mere coincidences to meaningful discoveries. *Cognition*, 103(2):180–226.

Grinband, J., Hirsch, J., and Ferrera, V. P. (2006). A neural representation of categorization uncertainty in the human brain. *Neuron*, 49(5):757–763.

Guigon, E., Baraduc, P., and Desmurget, M. (2008). Optimality, stochasticity, and variability in motor behavior. *Journal of Computational Neuroscience*, 24(1):57–68.

Gull, S. (1988). Bayesian inductive inference and maximum entropy. In Erickson, G. and Smith, C., editors, *Foundations*, volume 1 of *Maximum entropy and Bayesian methods in science and engineering*. Kluwer Academic Press, Dordrecht, Netherlands.

Haber, S. N. (2003). The primate basal ganglia: Parallel and integrative networks. *Journal of Chemical Neuroanatomy*, 26(4):317–330.

Hansen, T., Olkkonen, M., Walter, S., and Gegenfurtner, K. R. (2006). Memory modulates color appearance. *Nature Neuroscience*, 9(11):1367–1368.

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., and Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Journal of Neuroscience*, 28(22):5623–5630.

Harris, C. M. and Wolpert, D. M. (1998). Signal-dependent noise determines motor planning. *Nature*, 394(6695):780–784.

Haxby, J. V., Horwitz, B., Ungerleider, L. G., Maisog, J. M., Pietrini, P., and Grady, C. L. (1994). The functional organization of human extrastriate cortex: a PET-rCBF study of selective attention to faces and locations. *Journal of Neuroscience*, 14(11):6336–6353.

Hazeltine, R. E., Prinzmetal, W., and Elliott, W. (1997). If it's not there, where is it? Locating illusory conjunctions. *Journal of Experimental Psychology: Human Perception and Performance*, 23(1):263–277.

Heekeren, H. R., Marrett, S., Bandettini, P. A., and Ungerleider, L. G. (2004). A general mechanism for perceptual decision-making in the human brain. *Nature*, 431(7010):859–862.

Heekeren, H. R., Marrett, S., Ruff, D. A., Bandettini, P. A., and Ungerleider, L. G. (2006). Involvement of human left dorsolateral prefrontal cortex in perceptual decision making is independent of response modality. *Proceedings on the National Academy of Sciences of the United States of America*, 103(26):10023–10028.

Heekeren, H. R., Marrett, S., and Ungerleider, L. G. (2008). The neural systems that mediate human perceptual decision making. *Nature Reviews Neuroscience*, 9(6):467–479.

Heinen, S. J., Rowland, J., Lee, B. T., and Wade, A. R. (2006). An oculomotor decision process revealed by functional magnetic resonance imaging. *Journal of Neuroscience*, 26(52):13515–13522.

Helbig, H. B. and Ernst, M. O. (2007a). Knowledge about a common source can promote visual- haptic integration. *Perception*, 36(10):1523–1533.

Helbig, H. B. and Ernst, M. O. (2007b). Optimal integration of shape information from vision and touch. *Experimental Brain Research*, 179(4):595–606.

Helmholtz, H. L. F. (1925). *Physiological optics, Vol. III: The perceptions of vision.* Optical Society of America, Rochester, NY.

Hernandez, A., Zainos, A., and Romo, R. (2002). Temporal evolution of a decision-making process in medial premotor cortex. *Neuron*, 33(6):959–972.

Hillis, J., Watt, S., Landy, M., and Banks, M. (2004). Slant from texture and disparity cues: Optimal cue combination. *Journal of Vision*, 4:967–992.

Hillis, J. M., Ernst, M. O., Banks, M. S., and Landy, M. S. (2002). Combining sensory information: Mandatory fusion within, but not between, senses. *Science*, 298(5598):1627–1630.

Hinton, G. E. (1999). Products of experts. In *Ninth International Conference on Artificial Neural Networks (ICANN 9)*, volume 1, pages 1–6. Piscataway, NJ.

Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161.

Hinton, G. E., Osindero, S., and Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554.

Hodgson, T. L. (2002). The location marker effect - saccadic latency increases with target eccentricity. *Experimental Brain Research*, 145(4):539–542.

Hsu, S. M. and Pessoa, L. (2007). Dissociable effects of bottom-up and top-down factors on the processing of unattended fearful faces. *Neuropsychologia*, 45(13):3075–3086.

Hubel, D. and Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160(1):106–154.

Huddleston, W. E. and DeYoe, E. A. (2008). The representation of spatial attention in human parietal cortex dynamically modulates with performance. *Cerebral Cortex*, 18(6):1272–1280.

Hulme, O. J. and Whiteley, L. (2007). The "mesh" as evidence – model comparison and alternative interpretations of feedback. *Behavioral & Brain Sciences*, 30(5-6):505–506.

Huys, Q. J., Zemel, R. S., Natarajan, R., and Dayan, P. (2007). Fast population coding. *Neural Computation*, 19(2):404–441.

Ito, M. and Gilbert, C. D. (1999). Attention modulates contextual influences in the primary visual cortex of alert monkeys. *Neuron*, 22(3):593–604.

Ito, S., Stuphorn, V., Brown, J. W., and Schall, J. D. (2003). Performance monitoring by the anterior cingulate cortex during saccade countermanding. *Science*, 302(5642):120–122.

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

Itti, L., Rees, G., and Tsotsos, J. K. (2005). *The neurobiology of attention*. Elsevier Academic Press, Oxford, UK.

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

Jacobs, R. A. (2002). What determines visual cue reliability? *Trends in Cognitive Sciences*, 6(8):345–350.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK.

Jazayeri, M. and Movshon, J. A. (2007). A new perceptual illusion reveals mechanisms of sensory decoding. *Nature*, 446(7138):912–915.

Jeffreys, H. (1939). *Theory of Probability*. Oxford University Press, Oxford, UK.

Johansson, R. S. and Birznieks, I. (2004). First spikes in ensembles of human tactile afferents code complex spatial fingertip events. *Nature Neuroscience*, 7(2):170–177.

Johnson, M. R., Mitchell, K. J., Raye, C. L., D'Esposito, M., and Johnson, M. K. (2007). A brief thought can modulate activity in extrastriate visual areas: Top-down effects of refreshing just-seen visual stimuli. *Neuroimage*, 37(1):290–299.

Johnston, J. C. and Pashler, H. (1990). Close binding of identity and location in visual feature perception. *Journal of Experimental Psychology: Human Perception and Performance*, 16(4):843–856.

Josephs, O., Turner, R., and Friston, K. J. (1997). Event-related fMRI. *Human Brain Mapping*, 5(4):243–248.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291.

Kaiser, J., Lennert, T., and Lutzenberger, W. (2007). Dynamics of oscillatory activity during auditory decision making. *Cerebral Cortex*, 17(10):2258–2267.

Kalman, R. E. (1960). A new approach to linear filtering and prediction theory. *Transactions of the ASME: Journal of Basic Engineering*, 82(Series D):35–45.

Kalman, R. E. and Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Transactions of the ASME: Journal of Basic Engineering*, 83(Series D):95–108.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11):4302–4311.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kastner, S., De Weerd, P., Desimone, R., and Ungerleider, L. G. (1998). Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. *Science*, 282(5386):108–111.

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., and Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, 22(4):751–761.

Kastner, S. and Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience*, 23:315–341.

Kepecs, A., Uchida, N., Zariwala, H., and Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, 455(7210):224–227.

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

Kilner, J. M., Friston, K. J., and Frith, C. D. (2007). The mirror-neuron system: A Bayesian perspective. *Neuroreport*, 18(6):619–623.

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

Kinchla, R. A. (1992). Attention. *Annual Review of Psychology*, 43:711–742.

Klein, R. M. and Ivanoff, J. (2005). Inhibition of return. In Itti, L., Rees, G., and Tsotsos, J. K., editors, *The neurobiology of attention*. Elsevier Academic Press, Oxford, UK.

Knill, D. C. (1998). Discrimination of planar surface slant from texture: human and ideal observers compared. *Vision Research*, 38(11):1683–1711.

Knill, D. C. (2003). Mixture models and the probabilistic structure of depth cues. *Vision Research*, 43(7):831–854.

Knill, D. C. (2007). Learning Bayesian priors for depth perception. *Journal of Vision*, 7(8):13, 11–20.

Knill, D. C. and Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–719.

Knill, D. C. and Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research*, 43(24):2539–2558.

Knill, E. C. and Richards, W., editors (1996). *Perception as Bayesian inference*. Cambridge University Press, Cambridge, UK.

Knudsen, E. I. (2007). Fundamental components of attention. *Annual Review of Neuroscience*, 30:57–78.

Knutson, B., Fong, G. W., Adams, C. M., Varner, J. L., and Hommer, D. (2001). Dissociation of reward anticipation and outcome with event-related fMRI. *Neuroreport*, 12(17):3683–3687.

Koch, C. (1994). *Large-scale neuronal theories of the brain*. MIT Press, Cambridge, MA.

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

Kording, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., and Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE*, 2(9):e943.

Kording, K. P., Fukunaga, I., Hovard, I. S., Ingram, J. N., and Wolpert, D. M. (2004). A neuroeconomics approach to inferring utility functions in sensorimotor control. *PLoS Biology*, 2(10):1652–1656.

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

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

Landy, M., Goutcher, R., Trommershauser, J., and Mamassian, P. (2007). Visual estimation under risk. *Journal of Vision*, 7(6):4, 1–15.

Laplace, P.-S. (1840). *A Philosophical Essay on Probabilities.* 6th edition.

Latham, P. E., Deneve, S., and Pouget, A. (2003). Optimal computation with attractor networks. *Journal of Physiology*, 97(4-6):683–694.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, 21:451–468.

Lavie, N. and Tsal, Y. (1994). Perceptual load as a major determinant of the locus of selection in visual attention. *Perception & Psychophysics*, 56(2):183–197.

Lee, T. S. and Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A: Optics Image Science and Vision*, 20(7):1434–1448.

Lengyel, M. and Dayan, P. (2008). Hippocampal contributions to control: The third way. *Advances in Neural Information Processing Systems*, 21:889–896.

Lengyel, M., Kwag, J., Paulsen, O., and Dayan, P. (2005). Matching storage and recall: hippocampal spike timing-dependent plasticity and phase response curves. *Nature Neuroscience*, 8(12):1677–1683.

Leon, M. I. and Shadlen, M. N. (1999). Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. *Neuron*, 24(2):415–425.

Lewald, J., Ehrenstein, W. H., and Guski, R. (2001). Spatio-temporal constraints for auditory–visual integration. *Behavioural Brain Research*, 121(1-2):69–79.

Lewald, J. and Guski, R. (2003). Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. *Cognitive Brain Research*, 16(3):468–478.

Li, X., Lu, Z. L., Tjan, B. S., Dosher, B. A., and Chu, W. (2008). Blood oxygenation level-dependent contrast response functions identify mechanisms of covert attention in early visual areas. *Proceedings of the National Academy of Sciences of the United States of America*, 105(16):6202–6207.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, 6(1):9–16.

Ling, S. and Carrasco, M. (2006). Sustained and transient covert attention enhance the signal via different contrast response functions. *Vision Research*, 46(8-9):1210–1220.

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

Lu, Z. L. and Dosher, B. A. (1998). External noise distinguishes attention mechanisms. *Vision Research*, 38(9):1183–1198.

Luce, R. (1986). *Response times*. Oxford University Press, New York, NY.

Luck, S. J., Hillyard, S. A., Mouloua, M., and Hawkins, H. L. (1996). Mechanisms of visual-spatial attention: Resource allocation or uncertainty reduction? *Journal of Experimental Psychology: Human Perception and Performance*, 22(3):725–737.

Lundqvist, D., Flykt, A., and Öhman, A. (1998). *The Karolinska directed emotional faces - KDEF*. Karolinska Instituet. Available at www.facialstimuli.com.

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

Mach, E. (1980). *Contributions to the analysis of the sensations*. Open Court Publishing Company, Chicago, Illinois.

Mackay, D. J. C. (2004). *Information theory, inference, and learning algorithms*. Cambridge University Press, Cambridge, UK.

MacKay, D. M. (1956). The epistemological problem for automata. In Shannon, C. E. and McCarthy, J., editors, *Automata studies*, pages 235–251. Princeton University Press, Princeton, NJ.

MacKay, D. M. and Wills, S. A. (2005). Distributed phase codes for associative memory, prediction, and latent variable discovery. In *UC Berkeley Colloquium*.

Macmillan, N. A. and Creelman, C. D. (2005). *Detection theory: A user's guide*. Lawrence Erlbaum Associates, Mahwah, New Jersey, 2nd edition.

Maddox, W. T. and Bohil, C. J. (2004). Probability matching, accuracy maximization, and a test of the optimal classifier's independence assumption in perceptual categorization. *Perception & Psychophysics*, 66(1):104–118.

Maloney, L. T. (1999). Physics-based approaches to modeling surface color perception. In Gegenfurtner, K. R. and Sharpe, L. T., editors, *In color vision: From genes to perception*. Cambridge University Press, Cambridge, UK.

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

Mamassian, P. and Landy, M. S. (2001). Interaction of visual prior constraints. *Vision Research*, 41(20):2653–2668.

Maquet, P. (2001). The role of sleep in learning and memory. *Science*, 294(5544):1048–1052.

Marcus, G. F. (2008). *Kluge: The Haphazard Construction of the Human Mind*. Houghton Mifflin Co, Boston, MA.

Marr, D. (1982). *Vision*. W. H. Freeman, San Francisco, CA.

Maunsell, J. H. and Newsome, W. T. (1987). Visual processing in monkey extrastriate cortex. *Annual Review of Neuroscience*, 10:363–401.

Maunsell, J. H. and Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neuroscience*, 29(6):317–322.

Maunsell, J. H. R. (2004). Neuronal representations of cognitive state: Reward or attention? *Trends in Cognitive Sciences*, 8(6):261–265.

Maunsell, J. H. R. and McAdams, C. J. (2000). Effects of attention on neuronal response properties in visual cerebral cortex. In Gazzaniga, M. S., editor, *The New Cognitive Neurosciences*, pages 315–324. MIT Press, Cambridge, MA.

McAdams, C. J. and Maunsell, J. H. R. (1999). Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *Journal of Neuroscience*, 19(1):431–441.

McAdams, C. J. and Reid, R. C. (2005). Attention modulates the responses of simple cells in monkey primary visual cortex. *Journal of Neuroscience*, 25(47):11023–11033.

McClure, S. M., Li, J., Tomlin, D., Cypert, K. S., Montague, L. M., and Montague, P. R. (2004). Neural correlates of behavioral preference for culturally familiar drinks. *Neuron*, 44(2):379–387.

McCullagh, P. and Nelder, J. (1989). *Generalised Linear Models*. Chapman Hall, London, UK.

McHaffie, J. G., Stanford, T. R., Stein, B. E., Coizet, V., and Redgrave, P. (2005). Subcortical loops through the basal ganglia. *Trends in Neurosciences*, 28(8):401–407.

Mehta, M. R. (2001). Neuronal dynamics of predictive coding. *Neuroscientist*, 7(6):490–495.

Michel, M. M. and Jacobs, R. A. (2008). Learning optimal integration of arbitrary features in a perceptual discrimination task. *Journal of Vision*, 8(2):3, 1–16.

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

Minka, T. (2001). Expectation propagation for approximate Bayesian inference. In Breese, J. A. and Koller, D., editors, *UAI*, volume 17, pages 362–369, Washington, USA. Morgan Kaufman.

Minka, T. (2005). Divergence measures and message passing. Technical report, Microsoft Research.

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

Montague, P. R., King-Casas, B., and Cohen, J. D. (2006). Imaging valuation models in human choice. *Annual Review of Neuroscience*, 29:417–448.

Moray, N. P. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, 11(1):56–60.

Morgan, M. J., Ward, R. M., and Castet, E. (1998). Visual search for a tilted target: Tests of spatial uncertainty models. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 51(2):347–370.

Morgan, M. L., Deangelis, G. C., and Angelaki, D. E. (2008). Multisensory integration in macaque visual cortex depends on cue reliability. *Neuron*, 59(4):662–673.

Motter, B. C. (1993). Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *Journal of Neurophysiology*, 70(3):909–919.

Mozer, M. C. and Baldwin, D. S. (2008). Experience guided search: A theory of attentional conrol. *Advances in Neural Information Processing Systems*, 21.

Müller, C. M., Brenner, E., and Smeets, J. B. (2007). Living up to optimal expectations. *Journal of Vision*, 7(3):2.

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics*, 66(3):241–251.

Najemnik, J. and Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391.

Najemnik, J. and Geisler, W. S. (2008). Eye movement statistics in humans are consistent with an optimal search strategy. *Journal of Vision*, 8(3):4 1–14.

Natarajan, R., Huys, Q. J., Dayan, P., and Zemel, R. S. (2008). Encoding and decoding spikes for dynamic stimuli. *Neural Computation*, 20(9):2325–2360.

Navalpakkam, V. and Itti, L. (2007). Search goal tunes visual features optimally. *Neuron*, 53(4):605–617.

Neill, W. T. (1977). Inhibitory and facilitatory processes in attention. *Journal of Experimental Psychology: Human Perception and Performance*, 3:444–450.

Neisser, U. (1967). *Cognitive Psychology*. Appleton Century Crofts, New York, NY.

Newsome, W. T., Britten, K. H., and Movshon, J. A. (1989). Neuronal correlates of a perceptual decision. *Nature*, 341(6237):52–54.

Nissen, M. J. (1985). Accessing features and objects: Is location special? In Posner, M. I. and Marin, O. S., editors, *Attention and performance XI*, pages 205–219. Erlbaum, Hillsdale, NJ.

Niv, Y., Daw, N. D., Joel, D., and Dayan, P. (2007). Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology*, 191(3):507–520.

Norman, D. A. (1968). Toward a theory of memory and attention. *Psychological Review*, 75(6):522–536.

O'Doherty, J., Rolls, E. T., Francis, S., Bowtell, R., McGlone, F., Kobal, G., Renner, B., and Ahne, G. (2000). Sensory-specific satiety-related olfactory activation of the human orbitofrontal cortex. *Neuroreport*, 11(2):399–403.

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Current Opinion in Neurobiology*, 14(6):769–776.

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

Orbán, G., Berkes, P., Lengyel, M., and Fiser, J. (2008). Looking for hallmarks of generative models in the visual cortex. In *Computational and Systems Neuroscience (CoSyNe) Abstracts*, pages III–1, Salt Lake City, Utah.

Padoa-Schioppa, C. and Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, 441(7090):223–226.

Palmer, J. (1994). Set-size effects in visual-search - The effect of attention is independent of the stimulus for simple tasks. *Vision Research*, 34(13):1703–1721.

Palmer, J. (1995). Attention in visual search: Distinguishing four causes of a set-size effect. *Current Directions in Psychological Science*, 4(4):118–123.

Palmer, J., Ames, C. T., and Lindsey, D. T. (1993). Measuring the effect of attention on simple visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 19(1):108–130.

Papadimitriou, C, H. (1994). *Computational Complexity*. Addison Wesley, Indianapolis, IN.

Pashler, H. (1987). Detecting conjunctions of color and form: Reassessing the serial search hypothesis. *Perception & Psychophysics*, 41(3):191–201.

Pashler, H. (1998). *The Psychology of Attention*. MIT Press, Cambridge, MA.

Paulus, M. P. and Frank, L. R. (2003). Ventromedial prefrontal cortex activation is critical for preference judgments. *Neuroreport*, 14(10):1311–1315.

Pearl, J. (1988a). Embracing causality in default reasoning. *Artificial Intelligence*, 35(2):259–271.

Pearl, J. (1988b). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann Publishers, San Mateo, CA.

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10(4):437–442.

Pestilli, F., Viera, G., and Carrasco, M. (2007). How do attention and adaptation affect contrast sensitivity? *Journal of Vision*, 7(7):9 1–12.

Peters, R. J. and Itti, L. (2007). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN.

Phelps, E. A. and LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: from animal models to human behavior. *Neuron*, 48(2):175–187.

Philiastides, M. G. and Sajda, P. (2007). EEG-informed fMRI reveals spatiotemporal characteristics of perceptual decision making. *Journal of Neuroscience*, 27(48):13082–13091.

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

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., and Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999.

Plassmann, H., O'Doherty, J., and Rangel, A. (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *Journal of Neuroscience*, 27(37):9984–9988.

Plassmann, H., O'Doherty, J., Shiv, B., and Rangel, A. (2008). Marketing actions can modulate neural representations of experienced pleasantness. *Proceedings of the National Academy of Sciences of the United States of America*, 105(3):1050–1054.

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

Pleger, B., Blankenburg, F., Ruff, C. C., Driver, J., and Dolan, R. J. (2008). Reward facilitates tactile judgments and modulates hemodynamic responses in human primary somatosensory cortex. *Journal of Neuroscience*, 28(33):8161–8168.

Ploran, E. J., Nelson, S. M., Velanova, K., Donaldson, D. I., Petersen, S. E., and Wheeler, M. E. (2007). Evidence accumulation and the moment of recognition: Dissociating perceptual recognition processes using fMRI. *Journal of Neuroscience*, 27(44):11912–11924.

Poghosyan, V. and Ioannides, A. A. (2008). Attention modulates earliest responses in the primary auditory and visual cortices. *Neuron*, 58(5):802–813.

Pouget, A., Dayan, P., and Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience*, 1(2):125–132.

Pouget, A., Dayan, P., and Zemel, R. S. (2003). Inference and computation with population codes. *Annual Review of Neuroscience*, 26:381–410.

Preuschoff, K., Bossaerts, P., and Quartz, S. R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. *Neuron*, 51(3):381–390.

Preuschoff, K., Quartz, S. R., and Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience*, 28(11):2745–2752.

Prinzmetal, W. (1981). Principles of feature integration in visual perception. *Perception & Psychophysics*, 30(4):330–340.

Prinzmetal, W., Presti, D. E., and Posner, M. I. (1986). Does attention affect visual feature integration? *Journal of Experimental Psychology: Human Perception and Performance*, 12(3):361–369.

Rangel, A., Camerer, C., and Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, 9(7):545–556.

Rao, R. P. N. (1999). An optimal estimation approach to visual perception and learning. *Vision Research*, 39(11):1963–1989.

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

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

Rao, R. P. N. (2005). Bayesian inference and attentional modulation in the visual cortex. *Neuroreport*, 16(16):1843–1848.

Rao, R. P. N. and Ballard, D. H. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation*, 9(4):721–763.

Rao, R. P. N. and Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87.

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

Ratcliff, R. and Rouder, J. (1998). Modeling response times for two-choice decisions. *Psychological Science*, 9(5):347–356.

Ratcliff, R. and Smith, P. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, 111:333–367.

Reddi, B. A. J., Asrress, K. N., and Carpenter, R. H. S. (2003). Accuracy, information, and response time in a saccadic decision task. *Journal of Neurophysiology*, 90(5):3538–3546.

Reddy, L., Moradi, F., and Koch, C. (2007). Top-down biases win against focal attention in the fusiform face area. *Neuroimage*, 38(4):730–739.

Redgrave, P., Prescott, T. J., and Gurney, K. (1999). The basal ganglia: A vertebrate solution to the selection problem? *Neuroscience*, 89(4):1009–1023.

Reinagel, P. and Reid, R. C. (2000). Temporal coding of visual information in the thalamus. *Journal of Neuroscience*, 20(14):5392–5400.

Ress, D., Backus, B. T., and Heeger, D. J. (2000). Activity in primary visual cortex predicts performance in a visual detection task. *Nature Neuroscience*, 3(9):940–945.

Reynolds, J. H. and Desimone, R. (1999). The role of neural mechanisms of attention in solving the binding problem. *Neuron*, 24(1):19–29, 111–125.

Reynolds, J. H., Pasternak, T., and Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron*, 26(3):703–714.

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., and Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science*, 306(5695):443–447.

Riesenhuber, M. and Poggio, T. (1999). Are cortical models really bound by the "binding problem"? *Neuron*, 24(1):87–93.

Rizzolatti, G., Riggio, L., Dascola, I., and Umilta, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia*, 25(1A):31–40.

Roach, N. W., Heron, J., and McGraw, P. V. (2006). Resolving multisensory conflict: A strategy for balancing the costs and benefits of audio-visual integration. *Proceedings of the Royal Society of London Series B: Biological Sciences*, 273(1598):2159–2168.

Robertson, L. (2005). Attention and binding. In Itti, L., Rees, G., and Tsotsos, J. K., editors, *The neurobiology of attention*. Elsevier Academic Press, Oxford, UK.

Roelfsema, P. R., Lamme, V. A., and Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, 395(6700):376–381.

Roesch, M. R. and Olson, C. R. (2004). Neuronal activity related to reward value and motivation in primate frontal cortex. *Science*, 304(5668):307–310.

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

Rolls, E. T., McCabe, C., and Redoute, J. (2008). Expected value, reward outcome, and temporal difference error representations in a probabilistic decision task. *Cerebral Cortex*, 18(3):652–663.

Romo, R., Hernandez, A., and Zainos, A. (2004). Neuronal correlates of a perceptual decision in ventral premotor cortex. *Neuron*, 41(1):165–173.

Romo, R. and Salinas, E. (2003). Flutter discrimination: Neural codes, perception, memory and decision making. *Nature Reviews Neuroscience*, 4(3):203–218.

Roskies, A. L. (1999). The binding problem. *Neuron*, 24(1):7–9, 111–125.

Rossi, A. F. and Paradiso, M. A. (1995). Feature-specific effects of selective visual attention. *Vision Research*, 35(5):621–634.

Rowland, B., Stanford, T., and Stein, B. (2007). A bayesian model unifies multisensory spatial localization with the physiological properties of the superior colliculus. *Experimental Brain Research*, 180(1):153–161.

Rust, N. C. and Movshon, J. A. (2005). In praise of artifice. *Nature Neuroscience*, 8(12):1647–1650.

Saarinen, J. (1996a). Localization and discrimination of "pop-out" targets. *Vision Research*, 36(2):313–316.

Saarinen, J. (1996b). Target localisation and identification in rapid visual search. *Perception*, 25(3):305–311.

Saenz, M., Buracas, G. T., and Boynton, G. M. (2002). Global effects of feature-based attention in human visual cortex. *Nature Neuroscience*, 5(7):631–632.

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

Sahani, M. and Nagarajan, S. S. (2004). Reconstructing meg sources with unknown correlations. *Advances in Neural Information Processing Systems*, 16:1–8.

Salinas, E. and Sejnowski, T. J. (2001). Correlated neuronal activity and the flow of neural information. *Nature Reviews Neuroscience*, 2(8):539–550.

Sanger, T. D. (1996). Probability density estimation for the interpretation of neural population codes. *Journal of Neurophysiology*, 76(4):2790–2793.

Sato, Y., Toyoizumi, T., and Aihara, K. (2007). Bayesian inference explains perception of unity and ventriloquism aftereffect: Identification of common sources of audiovisual stimuli. *Neural Computation*, 19(12):3335–3355.

Saunders, J. A. and Knill, D. C. (2001). Perception of 3D surface orientation from skew symmetry. *Vision Research*, 41(24):3163–3183.

Saunders, J. A. and Knill, D. C. (2003). Humans use continuous visual feedback from the hand to control fast reaching movements. *Experimental Brain Research*, 152(3):341–352.

Schall, J. D., Stuphorn, V., and Brown, J. W. (2002). Monitoring and control of action by the frontal lobes. *Neuron*, 36(2):309–322.

Schoenbaum, G. and Roesch, M. (2005). Orbitofrontal cortex, associative learning, and expectancies. *Neuron*, 47(5):633–636.

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

Schutz-Bosbach, S. and Prinz, W. (2007). Prospective coding in event representation. *Cognitive Processing*, 8(2):93–102.

Schwartz, A. B. (1994). Direct cortical representation of drawing. *Science*, 265(5171):540–542.

Schwartz, O., Sejnowski, T. J., and Dayan, P. (2006). A Bayesian framework for tilt perception and confidence. *Advances in Neural Information Processing Systems*, 18:1201–1208.

Seo, H. and Lee, D. (2007). Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. *Journal of Neuroscience*, 27(31):8366–8377.

Sereno, M. I., Pitzalis, S., and Martinez, A. (2001). Mapping of contralateral space in retinotopic coordinates by a parietal cortical area in humans. *Science*, 294(5545):1350–1354.

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

Seymour, B., O'Doherty, J. P., Koltzenburg, M., Wiech, K., Frackowiak, R., Friston, K., and Dolan, R. (2005). Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nature Neuroscience*, 8(9):1234–1240.

Shadlen, M. N., Hanks, T. D., Churchland, A. K., Kiani, R., and Yang, T. (2007). The speed and accuracy of a simple perceptual decision: A mathematical primer. In Doya, K., Ishii, S., Pouget, A., and Rao, R. P., editors, *Bayesian Brain: Probabilistic Approaches to Neural Coding*, page 209238. MIT Press, Cambridge, MA.

Shadlen, M. N. and Movshon, J. A. (1999). Synchrony unbound: A critical evaluation of the temporal binding hypothesis. *Neuron*, 24(1):67–77.

Shadlen, M. N. and Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, 86(4):1916–1936.

Shaw, M. L. (1984). Division of attention among spatial locations: A fundamental difference between letters and detection of luminance increments. In Bouma, H. and Bouwhais, D. G., editors, *Attention and Performance X*. Earlbaum, Hillsdale, NJ.

Shipp, S. (2004). The brain circuitry of attention. *Trends in Cognitive Sciences*, 8(5):223–230.

Shuler, M. G. and Bear, M. F. (2006). Reward timing in the primary visual cortex. *Science*, 311(5767):1606–1609.

Simen, P., Cohen, J. D., and Holmes, P. (2006). Rapid decision threshold modulation by reward rate in a neural network. *Neural Networks*, 19(8):1013–1026.

Simoncini, C. and Baldassi, S. (2008). Reward tunes up the representation of an oriented target, and it is not attention! *Abstract for Cue Combination - Unifying Perceptual Theory Workshop*.

Singer, W. and Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience*, 18:555–586.

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

Somers, D. C., Dale, A. M., Seiffert, A. E., and Tootell, R. B. (1999). Functional MRI reveals spatially specific attentional modulation in human primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 96(4):1663–1668.

Sommer, W. H., Kraft, A., Schmidt, S., Olma, M. C., and Brandt, S. A. (2008). Dynamic spatial coding within the dorsal frontoparietal network during a visual search task. *PLoS ONE*, 3(9):e3167.

Spitzer, H., Desimone, R., and Moran, J. (1988). Increased attention enhances both behavioral and neuronal performance. *Science*, 240(4850):338–340.

Stafford, T. and Gurney, K. N. (2007). Biologically constrained action selection improves cognitive control in a model of the stroop task. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, 362(1485):1671–1684.

Stein, R. B., Gossen, E. R., and Jones, K. E. (2005). Neuronal variability: Noise or part of the signal? *Nature Reviews Neuroscience*, 6(5):389–397.

Steinman, S. B. (1987). Serial and parallel search in pattern vision. *Perception*, 16(3):389–398.

Stocker, A. and Simoncelli, E. (2008). A model of self-consistent perception. In *Computational and Systems Neuroscience (CoSyNe) Abstracts*, pages II–71, Salt Lake City, UT.

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

Stocker, A. A. and Simoncelli, E. P. (2006b). Sensory adaptation within a Bayesian framework for perception. *Advances in Neural Information Processing Systems*, 18:1291–1298.

Stork, D. G. (1989). Is backpropagation biologically plausible? *IEEE Transactions in Neural Networks*, 2:241–246.

Stuphorn, V., Taylor, T. L., and Schall, J. D. (2000). Performance monitoring by the supplementary eye field. *Nature*, 408(6814):857–860.

Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., and Hirsch, J. (2006a). Predictive codes for forthcoming perception in the frontal cortex. *Science*, 314(5803):1311–1314.

Summerfield, C., Egner, T., Mangels, J., and Hirsch, J. (2006b). Mistaking a house for a face: Neural correlates of misperception in healthy humans. *Cerebral Cortex*, 16(4):500–508.

Summerfield, C. and Koechlin, E. (2008). A neural representation of prior information during perceptual inference. *Neuron*, 59(2):336–347.

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

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press, Cambridge, MA.

Tassinari, H., Hudson, T., and Landy, M. (2006). Combining priors and noisy visual cues in a rapid pointing task. *Journal of Neuroscience*, 26(40):10154–10163.

Taylor, P. C., Rushworth, M. F., and Nobre, A. C. (2008). Choosing where to attend and the medial frontal cortex: An fMRI study. *Journal of Neurophysiology*, 100(3):1397–1406.

Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7):309–318.

Thielscher, A. and Pessoa, L. (2007). Neural correlates of perceptual choice and decision making during fear-disgust discrimination. *Journal of Neuroscience*, 27(11):2908–2917.

Thoenissen, D., Zilles, K., and Toni, I. (2002). Differential involvement of parietal and precentral regions in movement preparation and motor intention. *Journal of Neuroscience*, 22(20):9024–9034.

Thornton, T. L. and Gilden, D. L. (2007). Parallel and serial processes in visual search. *Psychological Review*, 134(1):73–103.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review*, 34:273–286.

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., and Schultz, W. (2007). Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. *Journal of Neurophysiology*, 97(2):1621–1632.

Todorov, E. (2005). Stochastic optimal control and estimation methods adapted to the noise characteristics of the sensorimotor system. *Neural Computation*, 17(5):1084–1108.

Todorov, E. and Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, 5(11):1226–1235.

Tolhurst, D. J., Movshon, J. A., and Dean, A. F. (1983). The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Research*, 23(8):775–785.

Tom, S. M., Fox, C. R., Trepel, C., and Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811):515–518.

Tong, F., Nakayama, K., Vaughan, J. T., and Kanwisher, N. (1998). Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron*, 21(4):753–759.

Torralba, A., Oliva, A., Castelhano, M. S., and Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 113(4):766–786.

Tranchina, D., Gordon, J., and Shapley, R. M. (1984). Retinal light adaptation–evidence for a feedback mechanism. *Nature*, 310(5975):314–316.

Treisman, A. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology*, 12(4):242–248.

Treisman, A. (1969). Strategies and models of selective attention. *Psychological Review*, 76(3):282–299.

Treisman, A. (1977). Focused attention in the perception and retrieval of multidimensional stimuli. *Perception & Psychophysics*, 22(1):1–11.

Treisman, A. (1982). Perceptual grouping and attention in visual search for features and for objects. *Journal of Experimental Psychology: Human Perception and Performance*, 8(2):194–214.

Treisman, A. (1988). Features and objects: The fourteenth bartlett memorial lecture. *Quarterly Journal of Experimental Psychology*, 40(2):201–237.

Treisman, A. (1995). Modularity and attention: Is the binding problem real? *Visual Cognition*, 2(2 & 3):303–311.

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, 353(1373):1295–1306.

Treisman, A. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136.

Treisman, A. and Sato, S. (1990). Conjunction search revisited. *Journal of Experimental Psychology-Human Perception and Performance*, 16(3):459–478.

Treisman, A. and Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology*, 14(1):107–141.

Treue, S. (2001). Neural correlates of attention in primate visual cortex. *Trends in Neurosciences*, 24(5):295–300.

Treue, S. and Martinez Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399(6736):575–579.

Trommershauser, J., Gepshtein, S., Maloney, L. T., Landy, M. S., and Banks, M. S. (2005). Optimal compensation for changes in task-relevant movement variability. *Journal of Neuroscience*, 25(31):7169–7178.

Trommershauser, J., Maloney, L. T., and Landy, M. (2003a). Statistical decision theory and trade-offs in the control of motor response. *Spatial Vision*, 16(3-4):255–275.

Trommershauser, J., Maloney, L. T., and Landy, M. S. (2003b). 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(7):1419–1433.

Trommershauser, J., Mattis, J., Maloney, L. T., and Landy, M. S. (2006). Limits to human movement planning with delayed and unpredictable onset of needed information. *Experimental Brain Research*, 175(2):276–284.

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

van Beers, R. J., Baraduc, P., and Wolpert, D. M. (2002). Role of uncertainty in sensorimotor control. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, 357(1424):1137–1145.

van Beers, R. J., Sittig, A. C., and Gon, J. J. (1999). Integration of proprioceptive and visual position-information: An experimentally supported model. *Journal of Neurophysiology*, 81(3):1355–1364.

van Ee, R., Adams, W. J., and Mamassian, P. (2003). Bayesian modeling of cue interaction: Bistability in stereoscopic slant perception. *Journal of the Optical Society of America A: Optics, Image Science and Vision*, 20(7):1398–1406.

Van Essen, D. C., Anderson, C. H., and Felleman, D. J. (1992). Information-processing in the primate visual-system - an integrated systems perspective. *Science*, 255(5043):419–423.

VanRullen, R. and Thorpe, S. J. (2002). Surfing a spike wave down the ventral stream. *Vision Research*, 42(23):2593–2615.

Verghese, P. and Nakayama, K. (1994). Stimulus discriminability in visual search. *Vision Research*, 34(18):2453–2467.

Verschure, P. F. M. J. and Althaus, P. (2003). A real-world rational agent: Unifying old and new ai. *Cognitive Science*, 27(4):561–590.

von der Malsburg, C. (1995). Binding in models of perception and brain function. *Current Opinion in Neurobiology*, 5(4):520–526.

Vroomen, J., Bertelson, P., and de Gelder, B. (2001). The ventriloquist effect does not depend on the direction of automatic visual attention. *Perception & Psychophysics*, 63(4):651–659.

Vuilleumier, P. (2005). How brains beware: Neural mechanisms of emotional attention. *Trends in Cognitive Sciences*, 9(12):585–594.

Vuilleumier, P., Armony, J. L., Driver, J., and Dolan, R. J. (2001). Effects of attention and emotion on face processing in the human brain: An event-related fMRI study. *Neuron*, 30(3):829–841.

Vuilleumier, P. and Driver, J. (2007). Modulation of visual processing by attention and emotion: Windows on causal interactions between human brain regions. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, 362(1481):837–855.

Vuilleumier, P., Richardson, M. P., Armony, J. L., Driver, J., and Dolan, R. J. (2004). Distant influences of amygdala lesion on visual cortical activation during emotional face processing. *Nature Neuroscience*, 7(11):1271–1278.

Wade, N. J. and Bruce, V. (2001). Surveying the seen: 100 years of British vision. *British Journal of Psychology*, 92(1):79–112.

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

Wallis, J. D. (2007). Orbitofrontal cortex and its contribution to decision-making. *Annual Review of Neuroscience*, 30:31–56.

Wallis, J. D. and Miller, E. K. (2003). Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *European Journal of Neuroscience*, 18(7):2069–2081.

Watanabe, M. and Sakagami, M. (2007). Integration of cognitive and motivational context information in the primate prefrontal cortex. *Cerebral Cortex*, 17 (Suppl. 1):i101–109.

Weiskopf, N., Hutton, C., Josephs, O., and Deichmann, R. (2006). Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: A whole-brain analysis at 3T and 1.5T. *Neuroimage*, 33(2):493–504.

Weiss, Y., Simoncelli, E. P., and Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, 5(6):598–604.

Welch, R. B. and Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, 88(3):638–667.

Welchman, A. E., Lam, J. M., and Bulthoff, H. H. (2008). Bayesian motion estimation accounts for a surprising bias in 3D vision. *Proceedings of the National Academy of Sciences of the United States of America*.

Wertheim, A. H., Hooge, I. T., Krikke, K., and Johnson, A. (2006). How important is lateral masking in visual search? *Experimental Brain Research*, 170(3):387–402.

Westheimer, G. (1979). Spatial sense of the eye - Proctor lecture. *Investigative Ophthalmology and Visual Science*, 18(9):893–912.

Whiteley, L. and Sahani, M. (2008). Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. *Journal of Vision*, 8(3):2, 1–15.

Whiteley, L., Spence, C., and Haggard, P. (2008). Visual processing and the bodily self. *Acta Psychologica*, 127(1):129–136.

Wichmann, F. and Hill, N. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & Psychophysics*, 63(8):1293–1313.

Williford, T. and Maunsell, J. H. R. (2006). Effects of spatial attention on contrast response functions in macaque area V4. *Journal of Neurophysiology*, 96(1):40–54.

Wills, S. A. (2004). *Computation with Spiking Neurons*. PhD thesis, Cambridge University.

Wilson, M. A. and McNaughton, B. L. (1993). Dynamics of the hippocampal ensemble code for space. *Science*, 261(5124):1055–1058.

Wojciulik, E., Kanwisher, N., and Driver, J. (1998). Covert visual attention modulates face-specific activity in the human fusiform gyrus: fMRI study. *Journal of Neurophysiology*, 79(3):1574–1578.

Wolfe, J. M. (1998). Visual search. In Pashler, H., editor, *Attention*, pages 13–74. Psychology Press Ltd, London, UK.

Wolfe, J. M. and Cave, K. R. (1999). The psychophysical evidence for a binding problem in human vision. *Neuron*, 24(1):11–17.

Wolfe, J. M., Cave, K. R., and Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3):419–433.

Wolpert, D. M., Ghahramani, Z., and Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882.

Womelsdorf, T., Schoffelen, J. M., Oostenveld, R., Singer, W., Desimone, R., Engel, A. K., and Fries, P. (2007). Modulation of neuronal interactions through neuronal synchronization. *Science*, 316(5831):1609–1612.

Wonnacott, T. H. and Wonnacott, R. J. (1990). *Introductory Statistics*. Wiley, Cambridge, MA, 5th edition.

Wood, J. N. and Grafman, J. (2003). Human prefrontal cortex: Processing and representational perspectives. *Nature Reviews Neuroscience*, 4(2):139–147.

Wyss, R., Konig, P., and Verschure, P. F. (2004). Involving the motor system in decision making. *Proceedings of the Royal Society of London Series B: Biological Sciences*, 271((Suppl. 3)):S50–52.

Yacubian, J., Glascher, J., Schroeder, K., Sommer, T., Braus, D. F., and Buchel, C. (2006). Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain. *Journal of Neuroscience*, 26(37):9530–9537.

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

Yang, T. and Zemel, R. S. (2000). Managing uncertainty in cue combination. *Advances in Neural Information Processing Systems*, 12.

Yin, H. H., Knowlton, B. J., and Balleine, B. W. (2005). Blockade of NMDA receptors in the dorsomedial striatum prevents action-outcome learning in instrumental conditioning. *European Journal of Neuroscience*, 22(2):505–512.

Yu, A. J. (2007). Optimal change-detection and spiking neurons. *Advances in Neural Information Processing Systems*, 19:1545–1552.

Yu, A. J. and Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, 46(4):681–692.

Yu, A. J., Dayan, P., and Cohen, J. D. (2008). Dynamics of attentional selection under conflict: Toward a rational Bayesian account. *Journal of Experimental Psychology: Human Perception and Performance*, In press.

Yuille, A. and Bulthoff, H. (1996). Bayesian decision theory and psychophysics. In Knill, D. and Richards, W., editors, *Perception as Bayesian Inference*. Cambridge University Press, Cambridge, UK.

Zeki, S. M. (1976). The functional organization of projections from striate to prestriate visual cortex in the rhesus monkey. *Cold Spring Harbor Symposia on Quantitative Biology*, 15:591–600.

Zeki, S. M. (1978). Functional specialisation in the visual cortex of the rhesus monkey. *Nature*, 274:423–428.

Zelinksy, G. (2005). Specifying the components of attention in a visual search task. In Itti, L., Rees, G., and Tsotsos, J. K., editors, *The neurobiology of attention*. Elsevier Academic Press, Oxford, UK.

Zemel, R. and Dayan, P. (1999). Distributional population codes and multiple motion models. *Advances in Neural Information Processing Systems*, 11.

Zemel, R. S. and Dayan, P. (1997). Combining probabilistic population codes. In *JCAI-97: 15th International Joint Conference on Artificial Intelligence*, volume 2, pages 1114–1119, Nagoya, Japan. Morgan Kauffman.

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

Zhang, K., Ginzburg, I., McNaughton, B. L., and Sejnowski, T. J. (1998). Interpreting neuronal population activity by reconstruction: Unified framework with application to hippocampal place cells. *Journal of Neurophysiology*, 79(2):1017–1044.