Brit. J. Phil. Sci. **63** (2012), 697–723

Bayes in the Brain—On Bayesian Modelling in Neuroscience

Matteo Colombo and Peggy Seriès

ABSTRACT

According to a growing trend in theoretical neuroscience, the human perceptual system is akin to a Bayesian machine. The aim of this article is to clearly articulate the claims that perception can be considered Bayesian inference and that the brain can be considered a Bayesian machine, some of the epistemological challenges to these claims; and some of the implications of these claims. We address two questions: (i) How are Bayesian models used in theoretical neuroscience? (ii) From the use of Bayesian models in theoretical neuroscience, have we learned or can we hope to learn that perception is Bayesian inference or that the brain is a Bayesian machine? From actual practice in theoretical neuroscience, we argue for three claims. First, currently Bayesian models do not provide mechanistic explanations; instead they are useful devices for predicting and systematizing observational statements about people's performances in a variety of perceptual tasks. That is, currently we should have an instrumentalist attitude towards Bayesian models in neuroscience. Second, the inference typically drawn from Bayesian behavioural performance in a variety of perceptual tasks to underlying Bayesian mechanisms should be understood within the three-level framework laid out by David Marr ([1982]). Third, we can hope to learn that perception *is* Bayesian inference or that the brain *is* a Bayesian machine to the extent that Bayesian models will prove successful in yielding secure and informative predictions of both subjects' perceptual performance and features of the underlying neural mechanisms.

1 *Introduction*
2 *Theoretical Neuroscientists meet Bayes*
3 *Is Perception Bayesian Inference?*
4 *How Should we Understand the Inference from Bayesian Observers to Bayesian Brains?*
5 *How Could we Discover that Brains are Bayesian?*
6 *Conclusion*

doi:10.1093/bjps/axr043

Advance Access published on February 21, 2012

1 Introduction

Theoretical neuroscience uses mathematical modelling and computer simulations to understand the brain and the behaviour it generates. Following on from an insight of Hermann von Helmholtz (Helmholtz [1925]), a growing trend in theoretical neuroscience considers that the human perceptual system is akin to a Bayesian machine (Jaynes [1957]; Knill *et al.* [1996]; Kersten and Schrater [2002]; Knill and Pouget [2004]; Friston and Stephan [2007]). The function of this machine would be to infer the causes of sensory inputs in an 'optimal' way. Since sensory inputs are often noisy and ambiguous, this requires representing and handling uncertainty. In order to carry out such a function, the nervous system would encode probabilistic models. These models would be updated by neural processing of sensory information using Bayesian inference.

Work on Bayesian modelling of perception can be usefully understood within David Marr's ([1982]) three levels of analysis framework. Marr's levels include the computational, the algorithmic, and the level of implementation. The computational level specifies the problem to be solved in terms of some generic input–output mapping. In the case of Bayesian modelling in theoretical neuroscience, this is the problem of handling uncertainty. If the task is one of extracting some property of a noisy stimulus, for example, the generic input–output mapping that defines the computational problem is a function mapping the noisy sensory input to an estimate of the stimulus that caused that input. It is 'generic' in that it does not specify any class of rules for generating the output. Such class is defined at the algorithmic level. The algorithm specifies how the problem can be solved. Bayesian models belong to this level. They provide us with one class of method for producing an estimate of a stimulus variable in function of noisy and ambiguous sensory information. The level of implementation is the level of physical parts and their organization. It describes the mechanism that carries out the algorithm.

Bayesian modelling is essential in machine learning, statistics, and economics. Given the increasing influence on neuroscience of ideas and tools from such fields, it's not surprising that Bayesian modelling has a lot to offer to the study of the brain. Yet, that the brain *is* a Bayesian machine does not follow from the fact that Bayesian models are *used* to study the brain and the behaviour it generates.

The aim of this article is to clearly articulate the claims that perception can be considered Bayesian inference and that the brain can be considered a Bayesian machine; some of the epistemological challenges to these claims; and some of the implications of these claims. In order to achieve this aim, we address two questions:

(i) How are Bayesian models used in theoretical neuroscience?

(ii) From the use of Bayesian models in theoretical neuroscience, have we learned, or can we hope to learn, that perception is Bayesian inference or that the brain is a Bayesian machine?

The article is structured as follows: Section 2 explains how Bayesian models are used in theoretical neuroscience by drawing on a widely cited case-study from psychophysics. Section 3 assesses whether, and in what sense, perception is akin to Bayesian inference. It is argued that currently this claim should be understood in an instrumentalist framework. That is, currently, Bayesian models are useful devices for predicting and systematizing observational statements about people's performances in a variety of perceptual tasks. Section 4 argues that the link between the claim that people perform as if they were Bayesian observers in a variety of tasks and the claim that the brain is a Bayesian machine should be understood within Marr's three levels of analysis framework. Section 5 takes up the questions whether, and in what sense, the claim that the brain is a Bayesian machine may be justified. Section 6 offers some concluding remarks.

2 Theoretical Neuroscientists Meet Bayes

Statistical inference is the process of drawing conclusions about an unknown distribution from data generated by that distribution. Bayesian inference is a type of statistical inference where data (or new information) is used to update the probability that a hypothesis is true. To say that a system performs Bayesian inference is to say that it updates the probability that a hypothesis H is true given some data D by applying Bayes' rule:[1]

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \tag{1}$$

We can read Equation (1) thus: the probability of the hypothesis given the data ($P(H|D)$) is the probability of the data given the hypothesis ($P(D|H)$) times the prior probability of the hypothesis ($P(H)$) divided by the probability of the data ($P(D)$).

Theoretical neuroscientists have been increasingly using Bayesian modelling to address questions about biological perception (Rao *et al.* [2002]; Doya *et al.* [2007]): 'One striking observation from this work is the myriad ways in which human observers behave as optimal Bayesian observers' (Knill and Pouget [2004], p. 712). From these types of behavioural results, further hypotheses are drawn about the brain.

[1] The issue of what it is for a physical system to apply Bayes' rule is controversial (Piccinini [2010]). Here, we assume a mechanistic account, according to which a machine that executes Bayes' rule is a system of organized components and related activities that processes 'computational vehicles' according to Bayesian inferential schemes that are 'sensitive to certain vehicle properties' (Piccinini [2010], section 2.5).

'This observation,' claim Knill and Pouget ([1996], p. 712) 'along with the behavioural and computational work on which it is based, has fundamental implication for neuroscience.' The 'fundamental implication for neuroscience,' is what they call the *Bayesian coding hypothesis*: 'the brain represents information probabilistically, by coding and computing with probability density functions or approximations to probability density functions' (Knill and Pouget [2004], p. 713).

The hypothesis is two-fold:

(1) The brain performs Bayesian inference to enable us to make judgements and guide action in the world.

(2) The brain represents sensory information in the form of probability distributions.

Although there is no agreement on how the details of the hypothesis should be cashed out, much published work in theoretical neuroscience subscribes to the general claim that some neural processes can be described as Bayesian inference. The remainder of this section focuses on a widely-cited experiment. This case study will help us to do three things: first, to make clearer why and how Bayesian models are used in theoretical neuroscience; second, to explain in what sense there is evidence underwriting the idea that 'human observers behave as optimal Bayesian observers' (Knill and Pouget [2004], p. 712); third, to assess the link drawn between behavioural evidence and the Bayesian coding hypothesis.

Our senses can be viewed as independent sources of information about the properties of external objects. Object perception, hence, can be viewed as integration of information from different senses. Pursuing these ideas, Ernst and Banks ([2002]) tackled these questions: When people both touch and look at an object, why are their percepts often more affected by visual than by haptic information? How does visual information integrate with haptic information? In particular, does this integration vary with the relative reliability of the information provided by each modality?

Since Knill and Pouget ([2004], p. 713) claim that '[p]erhaps the most persuasive evidence for the Bayesian coding hypothesis comes from sensory cue integration', we believe that Ernst and Banks's ([2002]) work is particularly suited to assess the sense in which perception can be considered Bayesian inference and the brain a Bayesian machine.

Ernst and Banks ([2002]) designed an experiment where human subjects were required to make discrimination judgements. Subjects had to judge which of two sequentially presented ridges was taller. There were three types of trials. First, the subjects had only haptic information: they could only touch the ridge. Then they had only visual information: they could only see the ridge. Finally, subjects had both types of information at the

same time: they could both touch and see the ridge simultaneously. The trials involving only visual information comprised four conditions that differed in the amount of noise in the visual stimuli so as to manipulate the reliability of the visual cue. To investigate cue integration quantitatively, Ernst and Banks measured the variances associated with subjects' judgements across the three types of trials. They first measured the variances associated with judgements based only on visual information, and based only on haptic information. From these, they could predict the performance of subjects for the condition where both visual and haptic cues were present, under the assumption that subjects would integrate information from the two cues in a Bayesian optimal way. They found that measured performance was in fact very similar to the Bayesian prediction. This and other results from a variety of different psychophysical experiments on perception—e.g. on colour perception (Brainard and Freeman [1997]), motion perception (Stocker and Simoncelli [2006]), visual illusions (Weiss *et al*. [2002]), and sensory-motor learning (Körding and Wolpert [2004a])—would be evidence that human observers are Bayes' optimal (Knill and Pouget [2004]).

What exactly does it mean that the subjects in Ernst and Banks's experiment behaved in a 'statistically optimal way'? How was Bayesian modelling used to reach the conclusion that 'humans integrate visual and haptic information in a statistically optimal fashion'? (Ernst and Banks [2002], p. 429).

To answer these questions, let's examine the logic underlying their experiment. Call S a random variable that takes on one of a set of possible values $S_1, \ldots, S_n$ of some physical property—e.g. colour, length, or velocity. A physical property of an object is any measurable property of that object. The value of S at a certain time describes the state of that object with respect to that property at that moment in time. Call M a sequence of measurements $M_1, \ldots, M_n$ of a physical property. M can be carried out through different measurement modalities. Call M_i a sequence of measurements obtained through modality i. Measurements M_i are typically corrupted by noise. Noise might cause a measurement M_i to yield the wrong value for a given S. An estimator $f(M_i)$ is a deterministic function that maps measurements M_i corrupted by noise to values of the physical property S. If we assume that M_i is the measurement carried out by sensory modality i—e.g. vision or touch—then perception *can* be modeled as Bayesian inference. Given a sequence of measurements M_i, the task of a Bayesian sensory system is to compute the conditional probability density function $P(S|M_i)$. We can then restate Bayes' rule (1) thus:

$$P(S|M_i) = \frac{P(M_i|S)P(S)}{P(M_i)} \qquad (1')$$

where $P(M_i|S)$ specifies the likelihood of the sensory measurements M_i for different values of the physical property S, $P(S)$ is the prior probability of different values of S, and $P(S|M_i)$ is the posterior density function. Bayesian inference here is concerned with computing the set of beliefs about the state of the world given sensory input. Bayes' rule alone does not specify how these beliefs should be used to generate a decision or a motor response. How to use the posterior distribution to generate a single response is described by Bayesian decision theory and requires the definition of a loss function $L(S, f(M_i))$, which specifies the relative cost of getting the estimate wrong. If the aim of the task is to compute a single estimate of S, the problem reduces to one of estimation.

It is worth noticing with Simoncelli ([2009]) two things: First, the estimation problem of extracting values of some physical property from noisy sensory measurements can be addressed with different classes of methods, of which exact Bayesian inference is only one possibility. This possibility, moreover, might not be biologically feasible: some researchers have argued that biological implementation of exact Bayesian inference is unfeasible because of its computational complexity and the knowledge it presupposes (e.g. Shimojo and Nakayama [1992]; Maloney [2002]; Fiser *et al.* [2010]). Alternative methods that do not require either representing or computing probabilities include: regression techniques, look-up tables, and some other supervised and unsupervised inferential strategies. Second, in the case of Bayesian modelling '[o]ptimality is not a fixed universal property of an estimator, but one that depends on [three] defining ingredients' (Simoncelli [2009], p. 529). The first two ingredients are:

- the prior $P(S)$; and
- the measurement probability density $P(M_i|S)$.

Together, they specify how to form the posterior distributions. However, they do not tell us how to use the posterior distribution to make judgements and decisions in a task. The prior and measurement probability density, that is, are not sufficient to specify which estimate should be picked in a task. The optimal choice for this depends on a third ingredient:

- the loss function $L(S, f(M_i))$.

If exact Bayesian inference is not feasible, optimality will depend on a fourth ingredient, that is:

- the family of functions F from which the estimator is to be chosen (Simoncelli [2009], p. 525), e.g. linear functions.

The most common way of choosing an estimate from the posterior distribution is known as maximum likelihood estimator (MLE). This corresponds to

choosing as an estimate $\hat{s}$ the value of the physical property that maximizes the probability of resulting in the observed measurements.

$$\hat{s} = \arg\max_{s}(P(M_i|S)) \tag{2}$$

The MLE method corresponds to optimal estimation under the assumption that the prior is flat (uniform), the loss function is 0 for $\hat{s} = s \pm \varepsilon$ (where ε is a very small quantity) and constant otherwise, and the family of functions F is not constrained.

Let's now re-examine Ernst and Banks's ([2002]) experiment in this framework. In their experiment, the physical property of interest was the height of the ridges, S. Two types of sensory measurements M_i were used: visual and haptic measurements. If we call V the sequence of visual measurements and T the sequence of haptic measurements of S, then the estimator $f(V, T)$ maps the integration of visual and haptic measurements corrupted by noise to estimated values $\hat{s}$. Ernst and Banks ([2002], pp. 429–30) reasoned that '[i]f the noises are independent and Gaussian with variance σ_i^2, and the Bayesian prior is uniform, then the maximum-likelihood estimate of the environmental property [. . .] states that the optimal means of estimation (in the sense of producing the lowest-variance estimate) is to add the sensor estimates weighted by their normalized reciprocal variances.'

The noises of different sensory modalities are independent when the conditional probability distribution of either, given the observed value of the other, is the same as if the other's value had not been observed. This assumption might be motivated by the fact that the neurons processing visual information are far apart from the cortical neurons processing haptic information in the cortex. To say that the Bayesian prior is uniform is to say that all values of S are equally likely before any measurement M_i. This assumption can be justified by noticing that the subjects in Ernst and Banks's ([2002]) experiment had no prior experience with the task, and thus no prior knowledge as to the height of the ridges in the experiment.

Based on individual measurements T and V of the physical property S, and assuming that the two modalities are independent, we can derive the *likelihood function* $P(T, V|S)$ which describes how likely it is that any S gives rise to measurements (T, V). Once particular measurements (T, V) are obtained, by using Bayes' rule the posterior probability $P(S|T, V)$ of S being the height of the ridge can be expressed as:

$$P(S|T, V) = \frac{P(T, V|S)P(S)}{P(T, V)} = \frac{P(T|S)P(V|S)P(S)}{P(T, V)} \propto P(T|S)P(V|S)P(S) \tag{1''}$$

If we assume that the prior $P(S)$ is flat (i.e. a constant) and we know the mean estimate and variance for each modality in isolation, we can predict the mean

and variance of the Bayes-optimal bimodal estimate in this way. If σ_T^2 is the variance of the estimate of S based on haptic measurements T, and σ_V^2 is the variance of the estimate of S based on visual measurements V, and the likelihoods are Gaussian, that is:

$$P(T|S) \propto \exp\left(-\frac{(T-S)^2}{2\sigma_T^2}\right) \tag{3}$$

$$P(V|S) \propto \exp\left(-\frac{(V-S)^2}{2\sigma_V^2}\right) \tag{3'}$$

then the posterior distribution of the final visual-haptic estimate will also be described by a Gaussian. That is, from (1″), (3), and (3′) it follows that:

$$P(S|T,V) \propto \exp\left(-\frac{(T-S)^2}{2\sigma_T^2}-\frac{(V-S)^2}{2\sigma_V^2}\right) = \exp\left(-\frac{\left(S-\frac{\sigma_V^2 T+\sigma_T^2 V}{\sigma_V^2+\sigma_T^2}\right)^2}{2\frac{\sigma_V^2\sigma_T^2}{\sigma_V^2+\sigma_T^2}}\right) \tag{4}$$

If we assume that subjects' estimations correspond to extracting the maximum of this distribution (MLE), the mean of their response is given by the mean of this Gaussian:

$$\langle S\rangle = \frac{\sigma_T^2}{\sigma_T^2+\sigma_V^2}V + \frac{\sigma_V^2}{\sigma_T^2+\sigma_V^2}T \tag{5}$$

It will fall between the mean estimates given by each isolated cue (if they differ) and will tend to be pushed towards the most reliable cue. Its variance is:

$$\sigma_S^2 = \frac{\sigma_V^2\sigma_T^2}{\sigma_V^2+\sigma_T^2} \tag{6}$$

This entails that the reliability of the combined estimate is always greater than that given by the estimates of each individual modality.

Ernst and Banks tested experimentally whether the variance of the subjects' visual-haptic estimates, $\hat{s}$, was close to the variance worked out through MLE. When they found that this was in fact the case, they concluded that humans integrate visual and haptic information in a statistically optimal fashion.

3 Is Perception Bayesian Inference?

The kind of results yielded by experimental studies such as Ernst and Banks's is often taken as evidence that perception is Bayesian inference (Knill *et al.* [1996]; Kersten and Schrater [2002]; Knill and Pouget [2004]; Friston and

Stephan [2007]). This conclusion may suggest that Bayesian models are descriptions of the *mechanisms* of sensory systems, that is: of the sets of entities and associated activities organized so as to constitute perceptual phenomena (Machamer *et al*. [2000]; Craver [2007]).

In this section, we argue that, currently, Bayesian models are not descriptions of the mechanisms of sensory systems. The most we can acknowledge from existing evidence is that viewing 'perception as Bayesian inference' is useful for generating *predictions* about people's *performance* in perceptual tasks. We explain that the goal of Bayesian models in psychophysics experiments is *not* to describe sensory mechanisms. Bayesian models are used as tools for predicting, systematizing and classifying statements about people's observable performance. Hence, claims about perception as Bayesian inference should be interpreted within an instrumentalist framework.

We start by introducing the distinction between scientific realism and instrumentalism. These are two stances about scientific theories (Devitt [2008]). The contrast between scientific realism and instrumentalism can be thought as a contrast in how scientific theories and models are to be understood—as a contrast in the epistemic attitude one should have towards scientific theories and models.

Call X the target system that a scientific model aims to 'represent'. Roughly, according to instrumentalism, scientific models of X are useful instruments, heuristic devices, or tools we employ to predict observable phenomena concerning X or to summarize and systematize data about X. For instrumentalists, a good model need *not* pick out organized component entities and activities in the target system. Scientific realism contrasts with instrumentalism. For realists, good scientific models of X pick out organized component entities and activities in the target system. In this sense, for realists, good scientific models do *not* only, or mainly, aim to make predictions, summarize, or systematize data about X.

For scientific realists, a model of X is better than alternative models of X if it is more successful than alternatives at describing the *mechanism* of X. According to instrumentalists, a successful model of X need not describe any aspect of a putative mechanism of X. From this perspective, a model of X is better than alternative models of X if it is more successful than alternatives in predicting a certain set of phenomena concerning X or in summarizing and systematizing data about X. We are aware that scientific realism and instrumentalism consist of a number of more specific theses (Psillos [1999]); for our purposes, however, this general characterization should be sufficient.

A scientific model of the human perceptual system such as Ernst and Banks's is *not* aimed at describing mechanisms. Ernst and Banks's experiment was informed by abstract considerations about information processing rather

than data about some mechanism of visual or tactile perception. The considerations that informed Ernst and Banks's experiment were two-fold. On the one hand, our senses need to extract information from different cues for estimating the properties of objects. On the other, under certain conditions a certain sensory modality or a certain type of cue seems to have more weight in the final estimation of the properties of objects. Given these considerations, we can be interested in the mechanistic project of understanding how humans' perceptual systems integrate different sensory cues, or we can be interested in the project of predicting and systematizing data about people's performance in a variety of perceptual estimation tasks. The psychophysical approach to the study of the human perceptual system is concerned with the latter project.

Works in 'Bayesian psychophysics' typically proceed from the definition of a simple experimental task and the specification of how a Bayesian observer would perform in that task. The experimental task is such that experimenters can determine probability distributions necessary to test whether subjects' performance is consistent with Bayesian inference. With visual-alone and haptic-alone subjects' discrimination estimations, Ernst and Banks could measure the variability of the unimodal estimates. From these, Ernst and Banks derived the optimal bimodal estimate by applying MLE; then they compared it to the experimental data.

Thus, MLE was used to formalize the idea that perception is statistical inference. It defined how an *ideal observer* would perform in a well-defined visual task. One type of ideal observer is the one who uses Bayesian inference to make perceptual estimations. Ideal observers serve as a benchmark against which human performance in the perceptual task can be compared. However, from the fact that human performance in such tasks is consistent with an ideal observer's performance, it does not follow that human observers carry out (either consciously or unconsciously) MLE when they integrate sensory information. Nor does it follow that people's brain implement MLE. The use of Bayesian modelling in psychophysical research is aimed at predicting and systematizing data. Given this aim, the claim that perception is Bayesian inference should be understood in an instrumentalist framework. There are two reasons that justify this conclusion. First, the methodology adopted is typically performance-oriented, instead of process-oriented. Second, typically the *choice* of the prior and of the loss function, which define the Bayesian formulation of perceptual estimation problems, has a mathematical justification rather than an empirical one.

Unlike process-oriented models, performance-oriented models treat their targets as systems that exhibit overall properties. No internal structure is specified within the model. The focus is not on the mechanism that gives rise to the performance but on the relationship between performance

and a benchmark.[2] The methodology informing Bayesian modelling in psychophysical experiments such as Ernst and Banks's is performance-oriented. Their data consist of subjects' judgements under different conditions (e.g. estimations relying on visual information only versus estimations relying on haptic information only). Subjects' performances could finally be compared with the ideal observer's performance. This approach focuses on describing regularities in behavioural data. It makes no claim as to the processes underlying performance.

Maloney and Mamassian ([2009]) make an analogous point. They argue that the use of a Bayesian model 'as a benchmark model does not imply that human visuomotor processing is in any sense Bayesian inference, even when human performance is close to ideal' (p. 148). The behavioural patterns displayed by people in such perceptual tasks can result from different classes of non-Bayesian models. Maloney and Mamassian show how optimal performance in cue integration can be achieved by 'table look-up' observers who would process information in a *non*-Bayesian fashion by implementing reinforcement learning. One of their conclusions is that *if* we want to know whether people process information in a Bayesian way, we have to use Bayesian models in visuomotor experimental tasks differently. In particular, one could investigate the implications of Bayesian models in terms of representation of the underlying probability distributions. Maloney and Mamassian propose an experimental methodology, which they call 'transfer criteria', which may help us to discriminate Bayesian inference from some reinforcement learning algorithm in visuomotor experimental tasks. This methodology aims to assess whether observers can transfer knowledge about previously encountered priors, likelihoods, and loss functions to carry out novel tasks. If perception is Bayesian, observers who have learned to carry out two perceptual tasks, which are defined by two different priors, likelihoods, and loss functions, should be able to transfer knowledge of these functions to carry out a new task corresponding to novel combinations of the previously encountered priors, likelihoods, and loss functions. If observers' performances in the novel task is close to optimal without much practice, then we would have evidence that the system saves, restores, and combines in a Bayesian fashion representations of those functions. With few exceptions (e.g. Adams *et al.* [2004]), Bayesian models are not used in this way. They are typically used as a benchmark for performance in a single task.

[2] A different way to put the distinction between process- and performance-oriented models is in terms of models that are intended to be constrained by the details of the underlying mechanism versus models that are intended to be used to summarise/systematize data and make predictions about some outcome. Thanks to an anonymous referee for drawing our attention to this way to put the distinction.

That the aim of Bayesian modelling is predictive, since the underlying methodology is performance-oriented, can also be justified thus. Typically, the models used in psychophysical experiments are empirically underconstrained. Ernst and Banks's model, for example, does not provide any clue about how exactly information is acquired, represented, and transformed by the perceptual system. Models of perception that are constrained by incorporating empirical data about information acquisition, for example, may help us to understand in which order information is searched, when a search terminates, and why a certain class of models is suitable to predict subjects' performance in certain types of circumstances rather than others. *If* we want models that instruct us about the nature of certain phenomena, such models need to be constrained to incorporate data about some putative mechanism underlying subjects' performance. Without such constraints we have few grounds for maintaining that the model has counterparts in the world, even though people's performance in a given task is consistent with the performance the model predicts. If we have little reason to maintain that the model has counterparts in the world, we cannot conclude that people's perception *is* Bayesian inference from evidence that people often behave as ideal observers. As shown by Maloney and Mamassian ([2009]), a good fit between predictions about what an ideal observer will do in a given task and people's performances in that task does not necessarily mean that the Bayesian model describes the cognitive processes behind people's performances.

Secondly, to formulate cue combination in terms of Bayesian integration it is necessary to choose a prior and a loss function. The prior is assumed to capture the statistical structure of the environment. The loss function defines the goal of a given task by specifying the costs and benefits to the observers of their estimations. *If* Bayesian models were intended to represent some feature of the mechanisms of sensory perception, the experimenters' choice of prior and loss functions should be informed by empirical considerations. But typically prior and loss functions are chosen on theoretical grounds only, in order to keep the assumptions of the model as simple as possible. Hence Bayesian models are not aimed at representing some feature of the mechanisms of sensory perception. The choice of prior and loss function is aimed at making prediction in experimental tasks simple.

Stocker and Simoncelli ([2006], p. 578) underwrite this last claim by arguing that 'the prior distribution used in most Bayesian models to date was chosen for simplicity and/or computational convenience'. Ernst and Banks ([2002]), for example, chose uniform prior distributions for the physical property being estimated in their experiment (i.e. height cues). This decision can be justified on 'intuitive' grounds by observing that their subjects had no prior knowledge of the size of the ridges in that experiment. In general, however, sensory system processes are adapted to the perceptual signals to which they are exposed at

evolutionary, developmental, and behavioural timescales. Not all sensory signals are equally likely in one's environment, 'it is natural to assume that perceptual systems should be able to best process those signals that occur most frequently. Thus, it is the statistical properties of the environment that are relevant for sensory processing' (Simoncelli and Olshausen [2001], pp. 1193–4). Hence, in general, if the aim of Bayesian modelling is to acquire knowledge about underlying mechanisms of perception, then the criterion for choosing the prior should include some 'ecological' consideration since neural processing is influenced by the statistical properties of the environment.

An analogous set of issues arises in choosing the loss function. Ideal observers are those who minimize the average loss. Thus, the choice of the loss function defines what counts as optimal performance in a given perceptual task. The criterions for the choice of the loss function could be whether it represents the true gains (or losses) of the observer, or whether it facilitates prediction of certain types of performance in a given task because of its mathematical tractability. Discussing models of sensorimotor control, Körding and Wolpert ([2004b], p. 9839) argue that '[l]oss functions have been assumed to be quadratic in error in almost all the models of sensorimotor control.' This assumption is typically made purely for mathematical convenience. If the loss function is quadratic in error, that is $L(S, f(M_i)) = (S - f(M_i))^2$, then the optimal estimate is the mean of the posterior. A quadratic loss is simple to solve since it is differentiable, whereas, say, absolute error is not. Ernst and Banks ([2002]) made this assumption, and thereby showed that their subjects' performance was consistent with the claim that observers combine cues linearly with the choice of weights that minimize quadratic loss. However, Ernst and Banks gave no empirical reason for the claim that human observers *in fact* penalize the errors they make in that way.

One possible approach to the choice of prior and loss function is to develop psychophysical tasks that allow us to estimate them. Once we have gained some knowledge of the prior and loss function of the subjects performing in a task, we may constrain the Bayesian model and use it to predict the subjects' behaviour in *different* psychophysical tasks, as Maloney and Mamassian ([2009]) recommend with their 'transfer criteria'. This use of Bayesian modelling would give us grounds to maintain that the claim that perception is Bayesian inference is intended to offer an approximately true account of the mechanisms of perception. Some recent work in psychophysics (e.g. Körding and Wolpert [2004b]; Stocker and Simoncelli [2006]; Chalk *et al.* [2010]) pursued this approach by 'reverse-engineering' the shape of the prior and of the loss function directly from people's perceptual behaviour. The question underlying this 'reverse-engineering' approach is: for what choices of prior and loss function would the subject's performance be considered optimal?

Körding and Wolpert ([2004b]), for example, adopted this approach. They measured the loss associated with different errors in a sensorimotor task. The task was such that the subjects' choices were associated with different patterns of errors. The subjects were required to make their choice so as to be 'on average as accurate as possible'. Körding and Wolpert estimated the loss function used by subjects from their choice behaviour in the task. They assumed that 'people are able to optimize an inherent loss function and that we can systematically measure this function' (p. 9841). From the distribution of errors in the task, Körding and Wolpert found that their subjects seemed to use a loss function in which the cost increases approximately quadratically for small errors, and significantly less than quadratically for large errors.

The estimation of the loss function or of the prior, however, does *not* contribute by itself to the justification of a realistic interpretation of Bayesian models of sensory perception. In order to give grounds to such an interpretation, the *choice* of prior and loss function should be systematically constrained by evidence from independent studies. Once researchers fit parameters in a Bayesian model to one set of data, they should try to predict subjects' performances—now with these parameters fixed—in a further new set of circumstances. However, this kind of validation presents a number of challenges and might be unfeasible. As pointed out both by Stocker and Simoncelli ([2006]) and Körding and Wolpert ([2004b]), it is likely that the prior and the loss function are specific to the particular experimental task and the details of the particular physical property to be estimated in the task. Thus the prior and loss function estimated for certain subjects performing in a particular task may not generalize to different experimental conditions.

In contrast to current Bayesian models of perception, mechanistic models have different purposes. They aim to be explanatory, as they aim to give us genuine insight into the way perceptual systems work by describing their physical implementation. Mechanistic models, unlike current Bayesian models, purport to produce explanations that are potentially useful for intervention and control. So if we want to understand what we can currently learn about the brain with Bayesian models and how we can use them in cognitive neuroscience, it is important for us to mark their difference from mechanistic models.

Although Bayesian models are currently not mechanistic, they are still useful *epistemic devices*. For example, a model showing good predictive success in a given psychophysical task can give us reason to investigate why this is the case (Schrater and Kersten [2002]). Researchers in theoretical neuroscience typically make the inference that if people behave as Bayesian observers in psychophysical tasks, then their brains *must* implement some Bayesian estimation scheme and somehow represent probability distributions over possible

states of the sensory world (see e.g. Ma *et al*. [2006], p. 1432). But how should we understand the *inference* from psychophysics to brain mechanisms? The next section addresses this question.

4 How Should we Understand the Inference from Bayesian Observers to Bayesian Brains?

From their psychophysical results, Ernst and Banks ([2002], p. 431) drew a conclusion about brain functioning: 'we found that height judgements were remarkably similar to those predicted by the MLE integrator. Thus, the nervous system seems to combine visual and haptic information in fashion similar to the MLE rule.' It is not clear that this claim should be read in a realist fashion since the use of 'similar' can be compatible with an instrumentalist reading of Ernst and Bank's position.

Some authors seem to display more explicitly a realist stance towards results like those obtained by Ernst and Banks. Knill and Pouget ([2004], p. 718), for example, write that 'these [psychophysical] data strongly suggest that the brain codes even complex patterns of sensory and motor uncertainty in its internal representations and computations'. Knill writes that '[a]n emerging consensus from the perceptual work [in psychophysics] is that the visual brain is a near-optimal Bayesian estimator of object properties, for example, by integrating cues in a way that accounts for differences in their reliability' (Knill [2005], p. 103). Ma *et al*. ([2006]) claim that behavioural studies showing that subjects often behave as Bayesian observers have 'two important *implications*. First, neural circuits *must represent* probability distributions . . . Second, neural circuits *must be able to combine probability distributions nearly optimally*, a process known as Bayesian inference' (p. 1432, emphases added). Beierholm *et al*. ([2008]), after having introduced behavioural results on multisensory perception, write that 'cue combination has become the poster child for Bayesian inference in the nervous system'.

According to the realist stance, the model used in the psychophysical task would pick out at least some features of the mechanism that gave rise to the psychophysical performance. Hence a realist interpretation of Bayesian models would be apt to motivate the inference from behavioural performance to brain mechanism. Moreover, according to the so-called 'no miracle argument' for scientific realism (see e.g. Putnam [1975]), the inference from behavioural performance to brain mechanism would be *necessary* to account for the success of the model in the psychophysical task. The no-miracle argument starts from the premise that the success of Bayesian models in predicting behavioural performance in a wide range of tasks calls for explanation. We would have an explanation of why the predictions of Bayesian models hold only if they picked out organized component entities and activities

responsible for the system's performance. People's performance would therefore be explained in virtue of Bayesian models picking out features of brain mechanisms. The success of Bayesian models would appear miraculous if scientific realism were not endorsed.

Thus, under a realist interpretation, the fact that people's performance can be described in terms of MLE integration of sensory information would justify the inference that the nervous system combines information in a way 'similar to MLE.' A realist, it is worth noting, need not assume that for every model of behavioural performance there is a model of the neural processing carried out in some part of the brain such that the two models are isomorphic. The realist can hold that the Bayesian algorithms describing behavioural performance are not implemented in a particular neural population. Visual processing, for example, takes place along a cascade of many processing stages: 'If the system as a whole performs Bayesian inference, it seems unlikely that any one stage in this cascade represents a single component of the Bayesian model (e.g. the prior) or performs one of the mathematical operations in isolation (e.g. multiplying the prior and the likelihood)' (Rust and Stocker [2010], p. 384). In general, it may be the case that whole brains solve certain 'computational problems' in a distributed way such that their solutions are visible only at the level of behavioural performance, and the performance does not depend on any particular process in any specific part of the brain (see e.g. Dennett [1991]).

However, no-miracle arguments are controversial (see e.g. Lipton [2004], Ch. 11). And, more importantly, current practice in theoretical neuroscience—we have argued—shows that it is premature to endorse a realist attitude towards Bayesian models of sensory perception. Bayesian models do not, and do not purport to, represent features of the mechanisms of perception. How should we understand the inference from Bayesian observers to Bayesian brains then?

We argue that understanding this inference within Marr's three-level framework (Marr [1982]) fits nicely with an instrumentalist attitude towards Bayesian models. Marr's framework of levels of analysis is in fact often used to understand probabilistic models of cognition (e.g. Griffiths *et al.* [2010]). It can also be used to understand Bayesian models of sensory perception. The adoption of Marr's framework can both motivate the inference from behavioural performance to underlying mechanisms, *and* an instrumentalist attitude towards the Bayesian model. This is because of two features of Marr's framework: First, the relationship between the three levels is one of *non-decompositional realization*. This relationship does not necessitate inferences across levels. Nonetheless, it can motivate us to ask what type of mechanism may implement a given algorithm. Second, questions at the computational level are *formally independent* of issues at the other levels,

and therefore they can be tackled with little or no concern for constraints at lower levels.

Craver ([2007]) explains that the relationship between levels in a *mechanism* is a type of part-whole relationship where entities and activities at one level are *components* of organized entities and activities at a higher level (Craver [2007], Ch. 5). Craver notes that the relationship between Marr's three levels is not one of composition, but rather one of *realization*. Computational, algorithmic, and implementation levels are not decompositional. Algorithms are not components of computations. Rather, the algorithmic level *realizes* the computational level. Thus, the estimation of environmental properties in the human sensory system may be realized by certain algorithmic Bayesian transformations, which in turn may be realized by certain organized collections of neural structures and activities. This relationship of realization allows one to avoid questions concerning biological mechanisms, as it suggests that algorithmic and biological implementation approaches are just different ways of looking at the same system.

Adapting Craver's ([2007], p. 218) reconstruction of Marr's framework to our topic it may be said that the human sensory system as a whole is at once an estimator, a Bayesian manipulator of sensory information, and an organized collection of certain patterns of neural spikes. The system is an estimator in virtue of being a Bayesian manipulator of sensory information. It is a Bayesian manipulator of sensory information in virtue of being an organized set of neural circuits. The computational process of estimation, the Bayesian transformations, and the organized collection of neural circuits are all properties of the same system. But if they are all properties of the same system, the predictive success of a Bayesian model in a given psychophysical task can *motivate* us to investigate why this is the case. Hence, the discovery that the sensory system can solve the problem of sensory cue integration by using Bayesian inference, or its approximation using MLE, motivates the Bayesian coding hypothesis at the neural level.

Marr emphasized the formal independence of the three levels because different algorithms can solve the same computational problem and different hardwares (or mechanisms) can implement the same algorithms. In this sense, the discovery that people behave as though they were Bayesian observers does *not* compel us to make any specific claim at the neural level of implementation. Theoretical neuroscientists pursuing Marr's methodological approach can continue to work at the algorithmic level independently of findings about the hardware that implements it. It is important to note, however, that the formal independence of algorithmic and implementation levels does not entail that the algorithms used by the human cognitive system are best *discovered* independently of a detailed understanding of its neurobiological mechanisms. The formal independence of algorithms and

neural hardware *allows* theoretical neuroscientists to be able to use Bayesian models to generate predictions and systematize data, while remaining agnostic about the underlying mechanism (on instrumentalism and Marr's framework, see Danks [2008]).

The formal independence of the three levels emphasized by Marr, however, is not a claim about how the algorithms implemented by our cognitive system are best *discovered*. Ultimately, knowledge of the nervous system is essential to discovering what types of algorithms are carried out by our cognitive system. Therefore, knowledge of the nervous system should inform the Bayesian models we use to study perception if, by using those models, we aim to discover whether our cognitive system implements some Bayesian algorithm in solving a given perceptual task. The next section explains this last claim by exploring the issue of how Bayesian models could be *used* so as to be gradually informed and constrained by knowledge at the neural level of implementation.

5 How Could we Discover that Brains are Bayesian?

Theoretical neuroscientists are ultimately interested in how a system actually works. Hence they are ultimately interested in building *mechanistic* models where findings about the hardware inform investigations at the algorithmic and computational levels. Mechanistic models of sensory perception describe entities, activities, and organizational features that are relevant to represent and explain perceptual phenomena. Granted this goal of theoretically neuroscience, and granted that currently Bayesian models should be understood as no more than toolboxes for making predictions and systematizing data, how can an instrumentalist *use* of Bayesian models lead to gradually transforming them into mechanistic models so that a realist attitude towards such models can be justified?

A growing number of theoretical studies have started to explore how neural mechanisms could implement the types of Bayesian models used in psychophysical perceptual tasks (Rao [2004]; Ma *et al*. [2006]; Beck *et al*. [2008]; Deneve [2008]). To carry out this project, three issues need be addressed: (i) How might neurons represent uncertainty? (ii) How might they represent probability distributions? (iii) How might they implement different approximations to Bayesian inference?

Recall our case study above. Ernst and Banks ([2002]) derived psychometric functions from subjects' estimations. That is, they derived functions that described the relationship between a parameter of the physical stimulus (the height of a ridge) and the discrimination performances of the subjects. At the neural level, the probability that the physical stimulus takes any particular value can be estimated from firing activity. If one adopts Marr's framework the psychophysical model and the neural model are isomorphic. If the

algorithm that solves a problem such as sensory integration uses certain probability distributions, then that algorithm and those probability distributions are to be implemented neurally since it is on the neural hardware that this algorithm would run.

Although for a given experimental task the two models can be taken to be isomorphic, that does not mean that there is only one way that probability distributions could be neurally encoded. There are a number of proposals about how populations of neurons might code probability distributions (Ma *et al*. [2008]; Fiser *et al*. [2010]). These proposals consist of neural models aimed at predicting and systematizing statements about neural data. The current challenge for these models is to yield good, clear, and testable predictions at the neural level, a goal that has yet to be satisfactorily reached (Fiser *et al*. [2010]).

In general, good predictions have two epistemic virtues: they are secure and informative. Secure predictions are based on reliable, solid grounds. The more adjustable parameters a proposed model has, the more secure its predictions are, but the greater the risk of merely accommodating the data used to construct the model. In general, models should accommodate the data used to formulate them. But the risk for models that merely aim to accommodate some known data set is to overfit the data (Hitchcock and Sober [2004]). If a model fits perfectly the data of a given data set, its predictive power can be undermined. By over-fitting the data, it would be too sensitive to the idiosyncrasies or noise in the particular data set and would be unlikely to generalize across samples drawn from the same underlying distribution. Due to over-fitting the data, a model can yield predictions that are either uninformative or inaccurate.

Bayesian models are often simpler and depend on fewer parameters than other types of models designed to fit the same data (see e.g. Weiss *et al*. [2002]; Chalk *et al*. [2010]). In that sense, they are not particularly prone to over-fitting the noise. However, they suffer from a related concern: they have sometimes been accused of merely accommodating the data due to the use of ad hoc priors, thereby running the risk of yielding uninformative predictions. Hammett *et al*. ([2007], p. 565) emphasize this problem when they argue that 'a Bayesian model [of speed perception] might be seen as little more than a re-description of the data with little predictive power.' Their point is that it is not clear that, in general, a given Bayesian model can yield informative predictions of perceptual phenomena like speed perception. The model might accommodate *any* experimental result by moulding 'the shape of the prior to observed data' ([2007], p. 565). Theoretical neuroscientists, like Stocker and Simoncelli, using Bayesian models of sensory perception are aware of this problem. Their methodological advice is that 'in order to realize its potential for explaining biology, [a Bayesian model] needs to be

constrained to the point where it can make quantitative experimentally testable predictions' (Stocker and Simoncelli [2006], p. 583). The idea is that Bayesian models can be more than mere 'descriptions of the observed data.' They can yield good predictions of both subjects' perceptual performance in a variety of tasks *and* features of the underlying mechanisms to the extent they are able to incorporate knowledge of relevant neurophysiological constraints. By yielding good predictions, the models can then gradually 'explain biology.' If Bayesian models 'explain biology,' we would have more grounds for a realistic attitude towards them.

Good predictions are typically quantitatively accurate and informative, in that they match some *novel* phenomena, thereby avoiding over-fitting. Musgrave ([1974]) distinguishes three ways in which predictions match novel phenomena. According to the temporal view, a phenomenon is novel for a model only if it was unknown at the time the model was construed. According to the heuristic view, a phenomenon is novel if the model was not constructed only to accommodate it. According to the theoretical view, a phenomenon is novel for a model if it is not predicted by any of the model's extant alternatives. As the predictions yielded by a model neural network become more secure and informative, the model is gradually transformed into a mechanistic model. Whether a literal understanding of the claim that the brain *is akin to* a Bayesian machine is justified depends ultimately on the success of the transformation from Bayesian models as predictive tools to Bayesian models as mechanistic models.

Let's illustrate the logic underlying such a transformation by considering Ma *et al.*'s ([2006]) work. They tackled the questions of how neuronal activity can encode probability distributions and perform Bayesian inference by building a model network. Their methodology is top-down: the computational problem that motivates their work is analogous to Ernst and Banks's ([2002]). They relied on the finding that human observers perform in a Bayesian fashion in a variety of psychophysical tasks to claim that neurons 'must represent probability distributions' and 'must be able' to implement Bayesian inference ([2006], p. 1432).[3] They approached this implementational

[3] As already noted, the way Ma *et al.* ([2006]) put this point suggests a decidedly realist attitude towards Bayesian models of sensory perception. They write: 'Behavioral studies have confirmed that human observers not only take uncertainty into account in a wide variety of tasks, but do so in a way that is nearly optimal [. . .] This has two important *implications*. First, neural circuits *must represent* probability distributions [. . .] Second, neural circuits *must be able to combine probability distributions nearly optimally*, a process known as Bayesian inference' (p. 1432, emphases added). Interestingly, the way the same authors phrase the same point in a subsequent paper does *not* underwrite the same realist attitude. Now they write: 'models of neural representation and computation have started to explore the *possibility* that neurons encode probability distributions and that neural computation is *equivalent* to probabilistic inference. This work was *inspired* by psychophysical findings showing that human perception and motor control are nearly optimal in a Bayesian sense.' (Ma, Beck and Pouget [2008], p. 217, emphases added).

problem by observing that the response of cortical neurons has high variability, namely: firing responses of cortical neurons to the same stimulus vary dramatically from one presentation to the next. This variability can be described by Poisson statistics. Ma *et al.* ([2006], p. 1432) gave a specific interpretation to neural firing-rate variability: 'it allows neurons to represent probability distributions in a format that reduces optimal Bayesian inference to simple linear combinations of neural activities.' When a population of neurons displays Poisson-like firing-rate statistics, Bayesian cue integration can be implemented by a network of neurons by using linear operations on population activities. Ma *et al.* ([2006], p. 1436) claimed that their model makes a number of specific predictions about neural activations and behavioural performance in psychophysical tasks. For example, if an observer performs in a Bayesian fashion in a cue combination task, and the variability of multisensory neurons is Poisson-like, then 'the responses of these neurons to multisensory inputs should be the sum of the responses to the unisensory inputs' (p. 1436). Hence the model network is used as a tool to interpret existing neural data and to yield predictions based on such an interpretation.

The security of its predictions depends both on the identification of the particular types of circumstance where people behave as Bayesian observers and on the extent to which specific neural circuits exhibit Poisson statistics. The security of the predictions of the neural model, that is, depends on research both at the psychophysical and neurobiological level. Research at the psychophysical level should provide information about the relationship between certain classes of algorithms and certain classes of tasks. Knill and Pouget ([2004], p. 712) claim that there are 'myriad ways in which humans behave as optimal Bayesian observers.' But it may be the case that the 'myriad ways' are in fact instances of the same type. It may be the case that, though the tasks where people behave as optimal Bayesian observers *seem* to be different types of perceptual tasks, they are in fact the same type. By gaining better knowledge about such a relationship, we can identify under what circumstances a certain type of algorithm is sufficient to warrant the prediction that people will behave as Bayesian in a given task. Experimental situations where human subjects are found to behave sub-optimally, violating the predictions given by the Bayesian model (e.g. Eckstein *et al.* [2004]; Seriès *et al.* [2009]; Brayanov and Smith [2010]), are thus particularly informative for two reasons: they can lead to questioning the computational goal of the system and they can shed light into the constraints on the system at the implementational level.

Research at the level of neurophysiology should provide information about the extent to which Poisson-like variability is specific to some neural circuits. Ma *et al.*'s neural model does not target a specific neural population. Yet its predictive power depends on a specific neural feature, namely Poisson-like

variability. If it is uncertain whether Poisson-like variability is a general feature of cortical neurons, or even of all neurons involved in visual processing, then the predictions yielded by the model are unsecure. As mentioned above, visual processing takes place along a cascade of processes distributed over different circuits. If the visual system as a whole represents probability distributions and performs Bayesian inference, then it is likely that their specific instantiations will vary as a function of the specific neural, non-mathematical constraints along this cascade. Once knowledge is gained of where in the brain, and to what degree, neuronal variability is Poisson-like, Ma *et al.*'s model might be revised to incorporate information about the architecture of such a circuit. The predictions of this revised neural model will be limited to a specific circuit but, because of this greater level of detail, more secure.

Ma *et al.*'s model predicts novel phenomena in the heuristic and theoretical sense, hence it is informative. In the heuristic sense, the construction of their model was motivated by a computational problem and by psychophysical findings. It was not formulated specifically to accommodate the data about the high variability of the responses of cortical neurons. In the theoretical sense, unlike alternative proposals about how probabilities can be neurally represented and how Bayesian inference may be implemented in neural activity, it predicts that Bayesian cue integration is carried out by populations of neurons *because* of the specific Poisson-like form of their variability. Insofar as neural Bayesian models such as Ma *et al.* will explicitly commit themselves to precise interpretations of specific neural features that stand in some relationship to other features, they may predict 'novel' facts about these other neural features, such as the time course of multisensory integration or the action of specific neuromodulators. As the predictions of the model become more informative, the model itself might enable us to identify candidate mechanistic features of Bayesian cue integration.

An instrumentalist *use* of Bayesian models of perception may gradually transform the models into mechanistic models. Good predictions are secure and informative. Secure predictions can be yielded by models that specify under what circumstances a phenomenon is likely to obtain. Informative predictions can be yielded by models that provide novel interpretations of known neural features. If a model enables us to learn under what circumstances, in virtue of what components and in virtue of what relationships between such components a phenomenon is to be expected, then the model provides us with information about some set of organized parts and activities that may be responsible for that phenomenon. That is, the model provides us with information about a candidate mechanism. Currently, the claim that the brain is a Bayesian machine should not be understood as taking on a commitment to the truth of the Bayesian coding hypothesis. Talk of the Bayesian brain is currently a useful locution that refers to a class of models that function

as predictive tools. They enable us to make predictions about human performance and the neural activities that may generate that performance. We have argued that as long as such tools yield increasingly better predictions they may gradually transform into models of candidate mechanisms of sensory perception. Ultimately, the status of the claim that brains are Bayesian machines will depend on the quality of the predictions that Bayesian models in theoretical neuroscience can yield.

6 Conclusion

In 1952, Hodgkin and Huxley published their work on the action potential in the squid giant axon. This is one of the first and most successful models in theoretical neuroscience. Their model can be used to predict many features of different kinds of neurons. Hodgkin and Huxley wrote: 'certain features of our equations were capable of a physical interpretation, but the success of the equations is no evidence in favor of the mechanism of permeability change that we tentatively had in mind when formulating them' ([1952], p. 541).[4] Currently, Bayesian models in theoretical neuroscience should be treated analogously. In this article, we have explained how Bayesian models are used to understand the workings of the brain and the behaviour they generate. From actual practice in theoretical neuroscience, we have argued for three claims. First, Bayesian models do not provide mechanistic explanations currently, instead they are predictive instruments. Second, the inference typically drawn from psychophysical performance to the Bayesian coding hypothesis should be understood within Marr's framework. Third, *within* Marr's framework we can hope to learn that perception *is* Bayesian inference or that the brain *is* a Bayesian machine to the extent that Bayesian models will prove successful in yielding secure and informative predictions.

Acknowledgements

We are sincerely grateful to Andy Clark, Liz Irvine, Matthew Chalk, and three anonymous referees for this journal for their insightful comments and helpful suggestions. This work was generously supported by a British Society for the Philosophy of Science Doctoral Scholarship (to M. C.) and by an Engineering and Physical Sciences Research Council (EPSRC) Studentship awarded by the School of Informatics of the University of Edinburgh (to M. C.).

[4] Bogen ([2005]) was the first, or one of the first, to argue that the Hodgkin and Huxley model describes regularities but has no explanatory force. The model played an important epistemic role in the discovery of underlying mechanisms, but the regularities it describes had no explanatory import.

Matteo Colombo
Department of Philosophy
University of Edinburgh
Dugald Stewart Building, 3 Charles Street
George Square, EH8 9AD, Edinburgh
UK
m.colombo-2@sms.ed.ac.uk

Peggy Seriès
Institute for Adaptive and Neural Computation
School of Informatics, University of Edinburgh
10 Crichton Street, EH8 9AB, Edinburgh
UK
pseries@inf.ed.ac.uk

References

Adams, W. J., Graf, E. W. and Ernst, M. O. [2004]: 'Experience can Change the "Light-from-Above" Prior', *Nature Neuroscience*, **7**, pp. 1057–8.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E. and Pouget, A. [2008]: 'Probabilistic Population Codes for Bayesian Decision Making', *Neuron*, **60**, pp. 1142–52.

Beierholm, U., Körding, K. P., Shams, S. and Ma, W. J. [2008]: 'Comparing Bayesian Models for Multisensory Cue Combination without Mandatory Integration', in J. Platt, D. Koller, Y. Singer and S. Roweis (*eds*), *Advances in Neural Information Processing Systems*, Volume 20, Cambridge, MA: MIT Press, pp. 81–8.

Bogen, J. [2005]: 'Regularities and Causality: Generalizations and Causal Explanations', *Studies in History and Philosophy of Biological and Biomedical Sciences*, **36**, pp. 397–420.

Brainard, D. and Freeman, W. [1997]: 'Bayesian Color Constancy', *Journal of Optical Society of America A*, **14**, pp. 1393–411.

Brayanov, J. B. and Smith, M. A. [2010]: 'Bayesian and "Anti-Bayesian" Biases in Sensory Integration for Action and Perception in the Size-Weight Illusion', *Journal Neurophysiology*, **103**, pp. 1518–31.

Chalk, M., Seitz, A. R. and Seriès, P. [2010]: 'Rapidly Learned Stimulus Expectations Alter Perception of Motion', *Journal of Vision*, **10**, pp. 1–18.

Craver, C. F. [2007]: *Explaining the Brain*, Oxford: Oxford University Press.

Danks, D. [2008]: 'Rational Analyses, Instrumentalism, and Implementations', in N. Chater and M. Oaksford (*eds*), *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*, Oxford: Oxford University Press, pp. 59–75.

Deneve, S. [2008]: 'Bayesian Spiking Neurons I: Inference', *Neural Computation*, **20**, pp. 91–117.

Dennett, D. [1991]: *Consciousness Explained*, Boston: Little Brown.

Devitt, M. [2008]: 'Realism/Anti-Realism', in S. Psillos and M. Curd (*eds*), *The Routledge Companion to the Philosophy of Science*, London: Routledge, pp. 224–35.

Doya, K., Ishii, S., Pouget, A. and Rao, R. P. N. (*eds*) [2007]: *Bayesian Brain: Probabilistic Approaches to Neural Coding*, Cambridge, MA: MIT Press.

Eckstein, M. P., Abbey, C. K., Pham, B. T. and Shimozaki, S. S. [2004]: 'Perceptual Learning Through Optimization of Attentional Weighting: Human versus Optimal Bayesian Learner', *Journal of Vision*, **4**, pp. 1006–19.

Ernst, M. O. and Banks, M. S. [2002]: 'Humans Integrate Visual and Haptic Information in a Statistically Optimal Fashion', *Nature*, **415**, pp. 429–33.

Fiser, J., Berkes, B., Orbán, G. and Lengyel, M. [2010]: 'Statistically Optimal Perception and Learning: From Behavior to Neural Representations', *Trends in Cognitive Sciences*, **14**, pp. 119–30.

Friston, K. and Stephan, K. E. [2007]: 'Free Energy and the Brain', *Synthese*, **159**, pp. 417–58.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A. and Tenenbaum, J. B. [2010]: 'Probabilistic models of cognition: Exploring representations and inductive biases', *Trends in Cognitive Sciences*, **14**, pp. 357–64.

Hammett, S. T., Champion, R. A., Morland, A. B. and Thompson, P. G. [2007]: 'Perceptual Distortions of Speed at Low Luminance: Evidence Inconsistent with a Bayesian Account of Speed Encoding', *Vision Research*, **47**, pp. 564–8.

Hitchcock, C. R. and Sober, E. [2004]: 'Prediction Versus Accommodation and the Risk of Overfitting', *British Journal for the Philosophy of Science*, **55**, pp. 1–34.

Hodgkin, A. L. and Huxley, A. F. [1952]: 'A Quantitative Description of Membrane Current and its Application to Conduction and Excitation in Nerve', *Journal of Physiology*, **117**, pp. 500–44.

Jaynes, E. T. [1988]: 'How does the Brain do Plausible Reasoning?' Stanford Univ. Microwave Lab. Technical Report 421; reprinted, in G. J. Erickson and C. R. Smitt (*eds*), *Maximum-Entropy and Bayesian Methods in Science and Engineering*, Volume 1, London: Kluwer Academic Publishers, pp. 1–23.

Kersten, D. and Schrater, P. R. [2002]: 'Pattern Inference Theory: A Probabilistic Approach to Vision', in D. Heyer and R. Mausfeld (*eds*), *Perception and the Physical World: Psychological and Philosophical Issues in Perception*, New York: Wiley, pp. 191–227.

Knill, D. C. [2005]: 'Reaching for Visual Cues to Depth: The Brain Combines Depth Cues Differently for Motor Control and Perception', *Journal of Vision*, **5**, pp. 103–15.

Knill, D. C., Kersten, D. and Yuille, A. [1996]: 'A Bayesian Formulation of Visual Perception', in D. C. Knill and W. Richards (*eds*), *Perception as Bayesian Inference*, Cambridge: Cambridge University Press, pp. 1–21.

Knill, D. C. and Pouget, A. [2004]: 'The Bayesian Brain: The Role of Uncertainty in Neural Coding and Computation', *Trends in Neurosciences*, **27**, pp. 712–9.

Körding, K. P. and Wolpert, D. [2004a]: 'Bayesian Integration in Sensorimotor Learning', *Nature*, **427**, pp. 244–7.

Körding, K. P. and Wolpert, D. [2004b]: 'The Loss Function of Sensorimotor Learning', *Proceedings of the National Academy of Sciences*, **101**, pp. 9839–42.

Lipton, P. [2004]: *Inference to the Best Explanation*, 2nd edn, London: Routledge.

Ma, W. J., Beck, J. M., Latham, P. E. and Pouget, A. [2006]: 'Bayesian Inference with Probabilistic Population Codes', *Nature Neuroscience*, **9**, pp. 1432–8.

Ma, W. J., Beck, J. M. and Pouget, A. [2008]: 'Spiking Networks for Bayesian Inference and Choice', *Current Opinion in Neurobiology*, **18**, pp. 217–22.

Machamer, P. K., Darden, L. and Craver, C. F. [2000]: 'Thinking about Mechanisms', *Philosophy of Science*, **57**, pp. 1–25.

Maloney, L. T. [2002]: 'Statistical Decision Theory and Biological Vision', in D. Heyer and R. Mausfeld (*eds*), *Perception and the Physical World: Psychological and Philosophical Issues in Perception*, New York: Wiley, pp. 145–89.

Maloney, L. T. and Mamassian, P. [2009]: 'Bayesian Decision Theory as a Model of Human Visual Perception: Testing Bayesian Transfer', *Visual Neuroscience*, **26**, pp. 147–55.

Marr, D. [1982]: *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, New York: Freeman.

Musgrave, A. [1974]: 'Logical versus Historical Theories of Confirmation', *British Journal for the Philosophy of Science*, **25**, pp. 1–23.

Piccinini, G. [2010]: 'Computation in Physical Systems', in E. Zalta (*ed.*), *The Stanford Encyclopedia of Philosophy*, <http://plato.stanford.edu/archives/fall2010/entries/computation-physicalsystems/>.

Psillos, S. [1999]: *Scientific Realism: How Science Tracks Truth*, New York and London: Routledge.

Putnam, H. [1975]: *Mathematics, Matter, and Method*, Cambridge: Cambridge University Press.

Rao, R. P. N. [2004]: 'Bayesian Computation in Recurrent Neural Circuits', *Neural Computation*, **16**, pp. 1–38.

Rao, R. P. N., Olshausen, B. and Lewicki, M. (*eds*) [2002]: *Probabilistic Models of the Brain: Perception and Neural Function*, Cambridge, MA: MIT Press.

Rust, N. C. and Stocker, A. A. [2010]: 'Ambiguity and Invariance: Two Fundamental Challenges for Visual Processing', *Current Opinion in Neurobiology*, **20**, pp. 382–8.

Schrater, P. and Kersten, D. [2002]: 'Vision, Psychophysics, and Bayes', in R. P. N. Rao, B. A. Olshausen and M. S. Lewicki (*eds*), *Statistical Theories of the Brain*, Cambridge, MA: MIT Press.

Shimojo, S. and Nakayama, K. [1992]: 'Experiencing and Perceiving Visual Surfaces', *Science*, **257**, pp. 1357–63.

Seriès, P., Stocker, A. and Simoncelli, E. [2009]: 'Is the Homunculus "Aware" of Sensory Adaptation?', *Neural Computation*, **21**, pp. 1–33.

Simoncelli, E. P. [2009]: 'Optimal Estimation in Sensory Systems', in M. S. Gazzaniga (*ed.*), *The Cognitive Neurosciences*, Volume 4, Cambridge, MA: MIT Press, pp. 525–35.

Simoncelli, E. P. and Olshausen, B. [2001]: 'Natural Image Statistics and Neural Representation', *Annual Review of Neuroscience*, **24**, pp. 1193–216.

Stocker, A. A. and Simoncelli, E. P. [2006]: 'Noise Characteristics and Prior Expectations in Human Visual Speed Perception', *Nature Neuroscience*, **9**, pp. 578–85.

von Helmholtz, H. [1925]: *Treatise on Physiological Optics*, Volume 3, Rochester, NY: Optical Society of America.

Weiss, Y., Simoncelli, E. P. and Adelson, E. H. [2002]: 'Motion Illusions as Optimal Percepts', *Nature Neuroscience*, **5**, pp. 598–604.