Chapter 49

Bayesian models of perceptual organization

Jacob Feldman

Inference in Perception

One of the central ideas in the study of perception is that the proximal stimulus—the pattern of energy that impinges on sensory receptors, such as the visual image—is not sufficient to specify the actual state of the world outside (the distal stimulus). That is, while the image of your grandmother on your retina might look like your grandmother, it also looks like an infinity of other arrangements of matter, each having a different combination of three-dimensional structures, surface properties, colour properties, etc., so that they happen to look just like your grandmother from a particular viewpoint. Naturally, the brain generally does not perceive these far-fetched alternatives, but rapidly converges on a single solution, which is what we consciously perceive. A shape on the retina might be a large object that is far away, or a smaller one that is closer, or anything in between. A mid-grey region on the retina might be a bright white object in dim light, a dark object in bright light, or anything in between. An elliptical shape on the retina might be an elliptical object face-on, a circular object slanted back in depth, or anything in between. Every proximal stimulus is consistent with an infinite family of possible scenes, only one of which is perceived.

The central problem for the perceptual system is to quickly and reliably decide among all these alternatives, and the central problem for visual science is to figure out what rules, principles, or mechanisms the brain uses to do so. This process was called *unconscious inference* by Helmholtz, perhaps the first scientist to appreciate the problem, and is sometimes called *inverse optics* to convey the idea that the brain must in a sense invert the process of optical projection—to take the image and recover the world that gave rise to it.

The modern history of visual science contains a wealth of proposals for how exactly this process works, far too numerous to review here. Some are very broad, like the Gestalt idea of *Prägnanz* (infer the simplest or most reasonable scene consistent with the image). Many others are narrowly addressed to specific aspects of the problem, like the inference of shape or surface colour. But historically, the vast majority of these proposals suffer from one or both of the following two problems. First, many (like Prägnanz and many other older suggestions) are too vague to be realized as computational mechanisms. They rest on central ideas, like the Gestalt term 'goodness of form', that are subjectively defined and cannot be implemented algorithmically without a host of additional assumptions. Second, many proposed rules are arbitrary or unmotivated, meaning that is unclear exactly why the brain would choose them rather than an infinity of other equally effective ones. Of course, it cannot be taken for granted that mental processes are principled in this sense, and some have argued for a view of the brain as a 'bag of tricks' (Ramachandran 1985). Nevertheless, to many theorists, a mental function as central and evolutionarily ancient as perceptual inference seems to demand a more coherent and principled explanation.

Inverse Probability and Bayes' Rule

In recent decades, Bayesian inference has been proposed as a solution to these difficulties, representing a principled, mathematically well-defined, and comprehensive solution to the problem of inferring the most plausible interpretation of sensory data. Bayesian inference begins with the mathematical notion of *conditional probability*, which is simply probability restricted to some particular set of circumstances. For example, the conditional probability of A conditioned on B, denoted $p(A|B)$, means the probability that A is true given that B is true. Mathematically, this conditional probability is simply the ratio of the probability of A and B both being true, $p(A$ and $B)$, divided by the probability that B is true, $p(B)$, hence

$$p(A \mid B) = \frac{p(A \text{ and } B)}{p(B)} \tag{1}$$

Similarly, the probability of B given A is the ratio of the probability that B and A are both true divided by the probability that A is true, hence

$$p(B \mid A) = \frac{p(B \text{ and } A)}{p(A)} \tag{2}$$

It was the Reverend Thomas Bayes (1763) who first noticed that these mathematically simple observations can be combined to yield a formula[1] for the conditional probability $p(A|B)$ (A given B) in terms of the inverse conditional probability $p(B|A)$ (B given A)

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)} \tag{3}$$

This formula is now called *Bayes' theorem* or *Bayes' rule*.[2] Before Bayes, the mathematics of probability had been used exclusively to calculate the chances of a particular random outcome of a stochastic process, like the chance of getting ten consecutive heads in ten flips of a fair coin [p(ten heads|fair coin)]. Bayes realized that his rule allowed us to *invert* this inference and calculate the probability of the conditions that gave rise to the observed outcome—here, the probability, having observed ten consecutive heads, that the coin was fair in the first place [p(fair coin|10 heads)]. Of course, to determine this, you need to assume that there is some *other* hypothesis we might entertain about the state of the coin, such as that it is biased towards heads. Bayes' logic, often called *inverse probability*, allows us to evaluate the plausibility of various hypotheses about the state of the world (the nature of the coin) on the basis of what we have observed (the sequence of flips). For example, it allows us to quantify the degree to which observing ten heads in a row might persuade us that the coin is biased towards heads.

1 More specifically, note that $p(B$ and $A) = p(A$ and $B)$ (conjunction is commutative). Substitute the latter for the former in Eq. (1) to see that $p(A|B)p(B)$, and likewise $p(B)p(A|B)$, are both equal to $p(A$ and $B)$ and thus to each other. Divide both sides of $p(A|B)p(B) = p(B|A)p(A)$ by $p(B)$ to yield Bayes' rule.

2 The rule does not actually appear in this form in Bayes' essay. But Bayes' focus was indeed on the underlying problem of inverse inference and deserves credit for the main insight (see Stigler 1983).

Bayes and his followers, especially the visionary French mathematician Laplace, saw how inverse probability could form the basis of a full-fledged theory of inductive inference (see Stigler 1986). As David Hume had pointed out only a few decades previously, much of what we believe in real life—including all generalizations from experience—cannot be proved with logical certainty, but instead merely seems intuitively plausible on the basis of our knowledge and observations. To philosophers seeking a deductive basis for our beliefs, this argument was devastating. But Laplace realized that Bayes' rule allowed us to *quantify* belief—to precisely gauge the plausibility of inductive hypotheses.

By Bayes' rule, given any data D which has a variety of possible hypothetical causes H_1, H_2, etc., each cause H_i is plausible in proportion to the product of two numbers: the probability of the data if the hypothesis is true $p(D|H_i)$, called the *likelihood*, and the *prior probability* of the hypothesis, $p(H_i)$, that is, how probable the hypothesis was in the first place. If the various hypotheses are all mutually exclusive, then the probability of the data D is the sum of its probability under all the various hypotheses:

$$p(D) = p(H_1)p(D \mid H_1) + p(H_1)p(D \mid H_1) + \cdots \sum_i p(H_i)p(D \mid H_i) \tag{4}$$

Plugging this into Bayes' rule (with H_i playing the role of A, and D playing the role of B), this means that the probability of hypothesis H_i given data D, called the *posterior probability, $P(H_i|D)$*, is

$$p(H_i \mid D) = \frac{p(H_i)p(D \mid H_i)}{p(D)} = \frac{p(H_i)p(D \mid H_i)}{\sum_i p(H_i)p(D \mid H_i)}, \tag{5}$$

or in words

$$\text{posterior for } H_i = \frac{\text{prior for } H_i \times \text{likelihood of } H_i}{\text{sum of (prior } \times \text{ likelihood) over all hypotheses}} \tag{6}$$

The posterior probability $p(H_i|D)$ quantifies how much we should believe H_i after considering the data. It is simply the ratio of the probability of the evidence under H_i (the product of its prior and likelihood) relative to the *total* probability of the evidence arising under all hypotheses (the sum of the prior–likelihood products for all the hypotheses). This ratio measures how plausible H_i is relative to all the other hypotheses under consideration.

But Laplace's ambitious account was followed by a century of intense controversy about the use of inverse probability (see Howie 2004). In modern retellings, the critics' objections to Bayesian inference are often reduced to the idea that to use Bayes' rule we need to know the prior probability of each of the hypotheses (for example, the probability that the coin was fair in the first place), and that we often don't have this information. But their criticism was far more fundamental, and relates to the meaning of probability itself. They argued that many propositions—those whose truth value is fixed but unknown—can't be assigned probabilities at all, in which case the use of inverse probability to assign them probabilities would be nonsensical. This criticism reflects a conception of probability, often called *frequentism*, in which probability refers exclusively to *relative frequency* in a repeatable chance situation. Thus, in their view, you can calculate the probability of a string of heads for a fair coin because this is a random event that occurs on some fraction of trials; but you can't calculate a probability of a *non*-repeatable state of nature, like *this coin is fair*, or *the Higgs' boson exists* because such hypotheses are either definitely true or definitely false, and are not 'random'. The frequentist objection was not just that we don't know the prior for many hypotheses, but that most hypotheses don't *have* priors—or posteriors, or any probabilities at all.

But, in contrast, Bayesians generally thought of probability as quantifying the *degree of belief*, and were perfectly content to apply it to any proposition at all, including non-repeatable ones. To Bayesians, the probability of any proposition is simply a characterization of our state of knowledge about it, and can freely be applied to *any* proposition as a way of quantifying how strongly we believe it. This conception of probability, sometimes called *subjectivist* (or *epistemic* or sometimes just *Bayesian*), is thus essential to the Bayesian programme. Without it, one cannot calculate the posterior probability of a non-repeatable proposition because such propositions simply don't have probabilities—and this would rule out most uses of Bayes' rule to perform induction. But to subjectivists, Bayesian inverse probability can be used to determine the posterior probability, and thus the strength of belief, for any hypothesis at all.[3]

Bayesian Inference as a Model of Perception

The use of Bayesian inference as a model for perception rests on two basic ideas. The first, just mentioned, is the basic idea of inverse probability as a general method for determining belief under conditions of uncertainty. Bayesian inference allows us to quantify the degree to which different *scene models*—hypotheses about what is actually going on in the world—should be believed on the basis of sensory data. Indeed, to many researchers, the subjectivist attitude towards probability resonates perfectly with the inherently 'subjective' nature of perception—that is, that by definition it involves understanding belief from the observer 's point of view

The other attribute of Bayesian inference that drives enthusiasm in its favour is its *rationality*. Cox (1961) showed that Bayesian inference is unique among inference systems in satisfying basic considerations of internal consistency, such as invariance to the order in which evidence is considered. If one wishes to assign degrees of belief to hypotheses in a rational way, one must inevitably use the conventional rules of probability, and specifically Bayes' rule. Later de Finetti (1970/1974) demonstrated that if a system of inference differs from Bayesian inference in any substantive way, it is subject to catastrophic failures of rationality. (His so-called Dutch book theorem shows, in essence, that any non-Bayesian reasoner can be turned into a 'money pump'.) In recent decades these strong arguments for the uniqueness of Bayesian inference as a system for fixing belief were brought to wide attention by Jaynes (2003). Though there are of course many subtleties surrounding the putatively optimal nature of Bayesian inference (see Earman 1992), most modern statisticians now regard Bayesian inference as a normative method for making inferences on the basis of data.

This characterization of Bayesian inference—as an optimal method for deciding what to believe under conditions of uncertainty—makes it perfectly suited to the central problem of perception, that of estimating the properties of the physical world based on sense data. The basic idea is to think of the stimulus (e.g. the visual image) as reflecting both stable properties of the world (which

[3] This philosophical disagreement underlies the recent debate between traditional statistics centred on null hypothesis significance testing (NHST) and Bayesian inference (see Lee and Wagenmakers 2005). NHST was invented by fervent frequentists (Fisher, Neyman, and Pearson) who insisted that scientific hypotheses, being non-repeatable, cannot have probabilities. This position rules out the application of Bayes' rule to estimate the posterior probability of a hypothesis, leading them to propose alternative ways of evaluating hypotheses such as 'rejecting the null'.

we would like to infer) plus some uncertainty introduced in the process of image formation (which we would like to disregard). Bayesian inference allows us estimate the stable properties of the world conditioned on the image data. The aptness of Bayesian inference as a model of perceptual inference was first noticed in the 1980s by a number of authors, and brought to wider attention by the collection of papers in Knill and Richards (1996). Since then the applications of Bayesian inference to perception have multiplied and evolved, while always retaining the core idea of associating perceptual belief with the posterior probability as given by Bayes' rule. Several excellent introductions are already available (e.g. Bülthoff and Yuille 1991; Kersten et al. 2004) each with a slightly different emphasis or slant. The current chapter is intended as an introduction to the main ideas of Bayesian inference as applied to human perception and perceptual organization. The emphasis will be on central principles rather than on mathematical details or recent technical advances.

Basic Calculations in Bayesian Inference

All Bayesian inference involves a comparison among some number of hypotheses H, drawn from a hypothesis set H, each of which has associated with it a prior probability $p(H)$ and a likelihood function $p(X|H)$ which gives the probability of each possible dataset X conditioned on H.[4] In many cases, the hypotheses H are qualitatively distinct from each other (H is finite or countably infinite). In other cases the hypotheses form a continuous family of hypotheses (the *hypothesis space*) distinguished by the setting of some number of parameters. In this case the problem is often called *parameter estimation*, because the observer's goal is to determine, based on the data at hand, the most probable value of the parameter(s), or, more broadly, the distribution of probability of over all possible values of the parameter(s) (called the *posterior distribution*). The mathematics of discrete hypothesis comparison and parameter estimation can look quite different (the former involving summation while the latter involves integration) but the logic is essentially the same: in both cases the goal is to infer the posterior assignment of belief to hypotheses, conditioned on the data.

The hypothesis with greatest posterior probability, the mode (maximum value) of the posterior distribution, is called the *maximum a posteriori* (MAP) hypothesis. If we need to reduce our posterior beliefs to a single value, this is by definition the most plausible, and casual descriptions of Bayesian inference often imply that Bayes' rule dictates that we choose the MAP hypothesis. But Bayes' rule does not actually authorize this reduction; it simply dictates how much to believe *each* hypothesis—that is, the full posterior distribution. In many situations use of the MAP be quite undesirable: for example, broadly distributed posteriors that have many other highly probable values, or multimodal posteriors that have multiple peaks that are almost as plausible as the MAP.

[4] Students are often warned that the likelihood function is *not* a probability distribution, a remark that in my experience tends to cause confusion. In traditional terminology, likelihood is a property of the model or hypothesis, not the data, and one refers, for example, to the likelihood of H (and not the likelihood of the data under H). This is because the term 'likelihood' was introduced by frequentists (specifically Fisher 1925), who insisted that hypotheses did not have probabilities, and sought a word other than 'probability' to express the degree of support given by the data to the hypothesis in question. To Bayesians, however, the distinction is unimportant, since both data and hypotheses can have probabilities. So Bayesians have tended (especially recently) to refer to the likelihood of the data under the hypothesis, or the likelihood of the hypothesis, in both cases meaning the probability $p(D|H)$. In this sense, likelihoods are indeed probabilities. However, note that the likelihoods of the various hypotheses do not have to sum to one; for example, it is perfectly possible for many hypotheses to have likelihood near one given a dataset that they all fit well. In this sense, the distribution of likelihood over *hypotheses* (models) is certainly not a probability distribution. But the distribution of likelihood over the data for a single fixed model is, in fact, a probability distribution and sums to one.

Reducing the posterior distribution to a single 'winner' discards useful information, and it should be kept in mind that only the full posterior distribution expresses the totality of our posterior beliefs.

Example: parameter estimation in motion

An example of parameter estimation drawn from perception is the estimation of motion based on a sequence of dynamically changing images. In everyday vision, we think of motion as a property of coherent objects plainly moving through space, in which case it is hard to appreciate the profound ambiguity involved. But in fact dynamically changing images are generally consistent with many motion interpretations, because the same changes can be interpreted as one visual pattern moving at one velocity (speed and direction), or another pattern moving at another velocity, or many options in between.

So the estimation of motion requires a comparison among a range of potential interpretations of an ambiguous collection of image data. As such, it can be placed in a Bayesian framework if one can provide (1) a prior over potential motions, indicating which velocities are more a priori plausible and which less, and (2) a likelihood function allowing us to measure the fit between each motion sequence and each potential interpretation. Weiss et al. (2002) have shown that many phenomena of motion interpretation—both under normal conditions as well as a range of standard motion illusions—are predicted by a simple Bayesian model in which (1) the prior favours *slower* speeds over faster ones, and (2) the likelihood is based on conventional Gaussian noise assumptions. That is, the posterior distribution favours interpretations that minimize speed while simultaneously maximizing fit to the observed data (leading to the simple slogan 'slow and smooth'). The close fit between human percepts and the predictions of the Bayesian model is particularly striking in that in addition to accounting for normal motion percepts, it also systematically explains certain illusions of motions as side-effects of rational inference.

Example: discrete hypotheses in contour integration

An example of discrete hypotheses in perception comes from the problem of contour integration (see Elder 2013; Singh 2013), in the question of whether two visual edges belong to the same contour (H_1) or different contours (H_2). Because physical contours can take on a wide variety of geometric forms, practically any observed configuration of two edges is consistent with the hypothesis of a single common contour. But because edges drawn from the same contour tend to be relatively collinear, the angle between two observed edges provides some evidence about how plausible this hypothesis is relative to the competing hypothesis that the two edges arise from distinct contours. This decision, repeated many times for pairs of edges throughout the image, forms the basis for the extraction of coherent object contours from the visual image (Feldman 2001).

To formalize this as a Bayesian problem, we need priors $p(H_1)$ and $p(H_2)$ for the two hypotheses, and likelihood functions $p(\alpha|H_1)$ and $p(\alpha|H_2)$ that express the probability of the angle between the two edges (called the *turning angle*) conditioned under each hypothesis. Several authors have modelled the same-contour likelihood function $p(\alpha|H_1)$ as a normal distribution centred on collinearity (0° turning angle; see Feldman 1997; Geisler et al. 2001). Figure 49.1. illustrates the decision problem in its Bayesian formulation. In essence, each successive pair of contour elements must be classified as either part of the same contour or as parts of distinct contours. The likelihood of each hypothesis is determined by the geometry of the observed configuration, with the normal likelihood function assigning higher likelihood to element pairs that are closer to collinear. The prior (in practice fitted to subjects' responses) tends to favour H_2, presumably because most image edges come from disparate objects. Bayes' rule puts these together to determine the most plausible

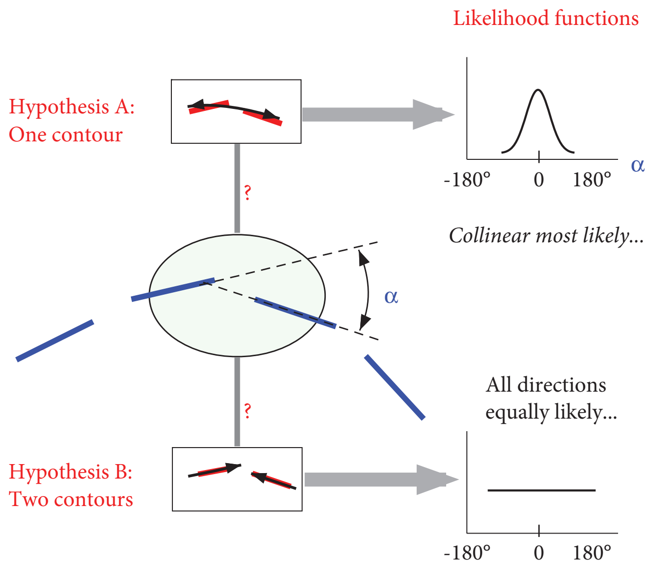

Fig. 49.1 Two edges can be interpreted as part of the same smooth contour (hypothesis A, top) or as two distinct contours (hypothesis B, bottom). Each hypothesis has a likelihood (right) that is a function of the turning angle α; with $p(\alpha|A)$ sharply peaked at 0° but $p(\alpha|B)$ flat.

grouping. Applying this simple formulation more broadly to all the image edge pairs allows the image to be divided up into a discrete collection of 'smooth' contours—that is, contours made up of elements which Bayes' rule says all belong to the same contour. The resulting parse of the image into contours agrees closely with human judgments (Feldman 2001). Related models have been applied to contour completion and extrapolation (Singh and Fulvio 2005).

Bayesian Perceptual Organization

The problems of perceptual organization—how to group the visual image into contours, surfaces, and objects—seems at first blush quite different from other problems in visual perception, because the property we seek to estimate is not a physical parameter of the world but a representation of how we choose to *organize* it. Still, Bayesian methods can be applied in a straightforward fashion as long as we assume that each image is potentially subject to many grouping interpretations, but that some are more intrinsically plausible than others (allowing us to define a prior over interpretations) and some fit the observed image better than others (allowing us to define a likelihood function). We can then use Bayes' rule to infer a posterior distribution over grouping interpretations.

More specifically, many problems in perceptual organization can be thought of as choices among discrete alternatives. Each qualitatively distinct way of organizing the image constitutes an alternative hypothesis. Should a grid of dots be organized into vertical or horizontal stripes (Zucker et al. 1983; Claessens and Wagemans 2008)? Should a configuration of dots be grouped into distinct clusters, and if so in what way (Compton and Logan 1993; Cohen et al. 2008; Juni et al. 2010)? What is the most plausible way to divide a smooth shape into a set of component parts (De Winter and Wagemans 2006; Singh and Hoffman 2001)? Each of these problems can be placed into a Bayesian framework by assigning to each distinct alternative interpretation a prior and a method for determining likelihood.

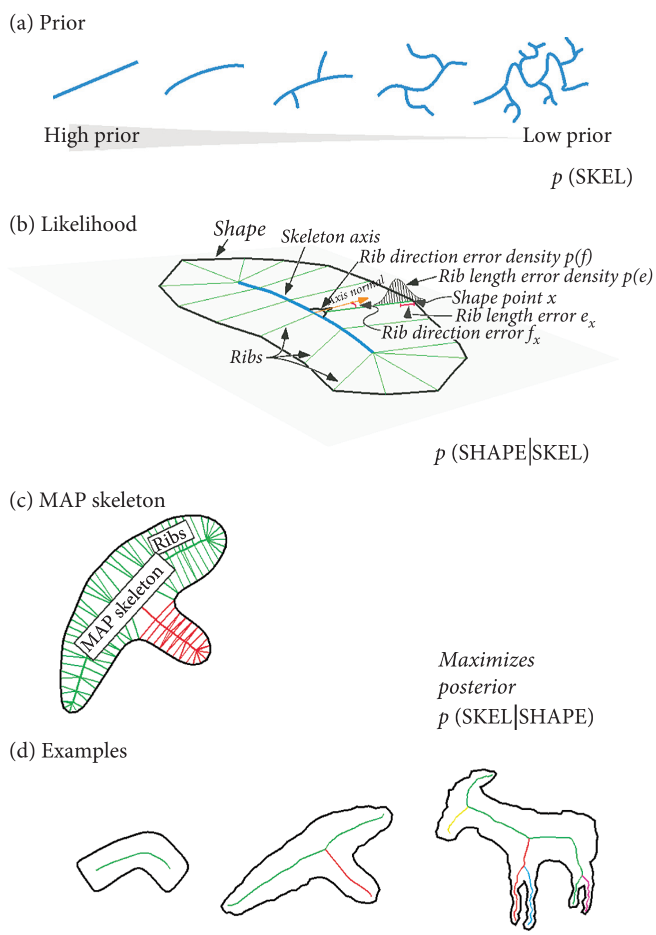

Fig. 49.2 Generative model for shape from Feldman and Singh (2006), giving: (a) prior over skeletons, (b) likelihood function, (c) MAP skeleton, the maximum posterior skeleton for the given shape, and (d) examples of the MAP skeleton.

Each of these problems requires its own unique approach, but broadly speaking a Bayesian framework for any problem in perceptual organization flows from a *generative model* for image configurations (Feldman et al. 2013). Perceptual organization is based on the idea that the visual image is generated by regular processes that tend to create visual structures with varying probability, which can be used to define likelihood functions. The challenge of Bayesian perceptual grouping is to discover psychologically reasonable generative models of visual structure.

For example, Feldman and Singh (2006) proposed a Bayesian approach to shape representation based on the idea that shapes are generated from axial structures (skeletons) from which the shape contour is understood to have 'grown' laterally. Each skeleton consists of a hierarchically organized collection of axes, and generates a shape via a probabilistic process that defines a probability distribution over shapes (Fig. 49.2). This allows a prior over skeletons to be defined, along with a likelihood function that determines the probability of any given contour shape conditioned on the skeleton. This in turn allows the visual system to determine the MAP skeleton (the skeleton most likely to have generated the observed shape) or, more broadly, a posterior distribution over skeletons. The estimated skeleton in turn determines the perceived decomposition into parts, with each section of the contour identified with a distinct generating axis perceived as a distinct 'part'. This shape model is certainly oversimplified relative to the myriad factors that influence real shapes, but the basic framework can be augmented with a more elaborate generative model, and tuned to the properties of natural shapes (Wilder et al. 2011). Because the framework is Bayesian, the resulting representation of shape is, in the sense discussed above, optimal given the assumptions specified in the generative model.

Discussion

This section raises issues that often arise when Bayesian models of cognitive processes are considered.

Bayesian updating

Bayesian inference is sometimes referred to as Bayesian *updating* because of the inherently progressive way that the arrival of new data leads the observer's belief to evolve from the prior towards the ultimate posterior. The initial prior represents the observer's beliefs before any data have been encountered. When data arrive, belief in all hypotheses is modified to reflect them: the likelihood of each hypothesis is multiplied by its prior (Bayes' rule) to yield a new, updated posterior belief distribution. From there on, the state of belief continues to evolve as new data are acquired, with the posterior at each step becoming the prior for the next step. In this way, belief is gradually pushed by the data away from the initial prior and towards the beliefs that better reflect the data.

More specifically, because of the way the mathematics works, the posterior distribution tends to get narrower and narrower (more and more sharply peaked) as more and more data come in. That is, belief typically evolves from a broad prior distribution (representing uncertainty about the state of the world) towards a progressively narrower posterior distribution (representing increasingly well-informed belief). In this sense, the influence of the prior gradually diminishes over the course of inference—in a Bayesian cliché, the 'likelihood swamps the prior'. Partly for this reason, though the source of the prior can be controversial (see Where do the priors come from?), in many situations (thought not all) its exact form is not too important, because the likelihood eventually dominates it.

Where do the priors come from?

As already mentioned, a great deal of controversy has centred on the epistemological status of prior probabilities. Frequentists long insisted that priors were justified only in the presence of 'real knowledge' about the relative frequencies of various hypotheses, a requirement that they argued ruled out most uses. A similar attitude is surprisingly common among present-day Bayesians in cognitive science (see Feldman 2013), many of whom aim to validate priors with respect to tabulations of relative frequency in natural conditions (e.g. Geisler et al. 2001; Burge et al. 2010; see Dakin 2013). However, as mentioned above, this restriction would limit the application of Bayesian models to hypotheses which (1) can be objectively tabulated and (2) are repeated many times under essentially identical conditions; otherwise objective relative frequencies cannot be defined. Unfortunately, these constraints would rule out many hypotheses which are of central interest in cognitive science, such as interpreting the intended meaning of a sentence (itself a belief, and not subject to objective measurement, and in any event unlikely ever to be repeated) or choosing the 'best' way to organize the image (again subjective, and again dependent on possibly unique aspects of the particular image). However, as already discussed, Bayesian inference is not

OUP UNCORRECTED PROOF – FIRSTPROOFS, Mon Nov 03 2014, NEWGEN

really limited to such situations if (as is traditional for Bayesians) probabilities are treated simply as quantifications of belief. In this view, priors do not represent the relative frequency with which conditions in the world obtain, but rather the observer's *uncertainty* (prior to receiving the data in question) about the hypotheses under consideration.

There are many ways of boiling this uncertainty down to a specific prior. Many descend from the Laplace's *principle of insufficient reason* (sometimes called the *principle of indifference*), which holds that a set of hypotheses, none of which one has any reason to favour, should be assigned equal priors. The simplest example of this is the assignment of uniform priors over symmetric options, such as the two sides of a coin or the six sides of a die. More elaborate mathematical arguments can be used to derive specific priors from more generalized symmetry arguments. One is Jeffreys' prior, which allows more generalized equivalences between interchangeable hypotheses (Jeffreys 1939/1961). Another is the maximum entropy prior (Jaynes 1982), which prescribes the prior that introduces the least information (in the technical sense of Shannon) beyond what is known.

Bayesians often favour so-called *uninformative* priors, meaning priors that are as 'neutral' as possible; this allows the data (via the likelihood) to be the primary influence on posterior belief. Exactly how to choose an uninformative prior can, however, be problematic. For example, to estimate the probability of success of a binomial process, like the probability of heads in a coin toss, it is tempting to adopt a uniform prior over success probability (i.e. equal over the range 0 to 100 per cent).[5] But mathematical arguments suggest that a truly uninformative prior should be relatively peaked at 0 and 100 per cent (the beta(0,0) distribution, sometimes called the Haldane prior; see Lee 2004). But recall that as data accumulate, the likelihood tends to swamp the prior, and the influence of the prior progressively diminishes. Hence while the choice of prior may be philosophically controversial, in some real situations the actual choice is moot.

More specifically, certain types of simple priors occur over and over again in Bayesian accounts. When a particular parameter x is believed to fall around some value μ, but with some uncertainty that is approximately symmetric about μ, Bayesians routinely assume a Gaussian (normal) prior distribution for μ, i.e. $p(x) \propto N(\mu, \sigma^2)$. Again, this is simply a formal way of expressing what is known about the value of x (that it falls somewhere near μ) in as neutral a manner as possible (technically, this is the maximum entropy prior with mean μ and variance σ^2). Gaussian error is often a reasonable assumption because random variations from independent sources, when summed, tend to yield a normal distribution (the *central limit theorem*).[6] But it should be kept in mind that an assumption of normal error along x does not entail an affirmative assertion that repeated samples of x would be normally distributed—indeed in many situations (such as where x is a fixed quantity of the world, like a physical constant) this interpretation does not even make sense. Such simple assumptions work surprisingly well in practice and are often the basis for robust inference. Another common assumption is that priors for different parameters that have no obvious relationship are *independent* (that is, knowing the value of one conveys no information about the value of the other). Bayesian models that assume independence among parameters whose relationship is unknown are sometimes called *naïve* Bayesian models. Again, an assumption of

[5] Bayes himself suggested this prior, now sometimes called Bayes' postulate, but he was apparently uncertain of its validity, which may have contributed to his reluctance to publish his essay (which was eventually published posthumously; see Stigler 1983).

[6] More technically, the central limit theorem says that the sum of random variables with finite variances tends towards normality in the limit. In practice this means that if x is really the sum of a number of component variables, each of which is random though not necessarily normal itself, then x tends to be normally distributed.

Wagemans_PC_280814OUK.indb 1039 11/3/2014 9:22:52 PM

independence does not reflect an affirmative empirical assertion about the real-world relationship between the parameters, but rather is an expression of ignorance about their relationship.

In the context of perception, there are several ways to think of the source of the prior. Of course, perceptual data arrive in a continuous stream from the moment of birth (or before). So in one sense the prior represents belief prior to experience—that is, the innate knowledge about the environment with which evolution has endowed our brains. But in another sense it simply represents belief prior to a given perceptual act, in which case it must also reflect the updated beliefs stemming from learning over the course of life. Of course, there is a long history of controversy about the magnitude and specificity of innate knowledge (Elman et al. 1996; Carruthers et al. 2005). Bayesian theory does not intrinsically take a position on this issue, easily accommodating either very broad or uninformative 'blank slate' priors, more narrowly tuned 'nativist' priors representing more specific knowledge about the environment, or anything in between. In any case because adult perceivers benefit from both innate knowledge and experience, priors estimated by experimental techniques (e.g. Girshick et al. 2011) must be assumed to reflect both evolution and learning in combination.

Computing the Posterior

In simple situations, it is sometimes possible to derive explicit formulae for the posterior distribution. For example, normal (Gaussian) priors and likelihoods lead to normal posteriors, allowing for easy computation. (Priors and posteriors in the same model family are called *conjugate*.) But in many realistic situations the priors and likelihoods give rise to an unwieldy posterior that cannot be expressed analytically. Much of the modern Bayesian literature is devoted to developing techniques to approximate the posterior in such situations. These include expectation maximization (EM), Markov chain Monte Carlo (MCMC), and Bayesian belief networks (Pearl 1988), each appropriate in somewhat different situations. (See Griffiths and Yuille (2006) for a brief introduction to these techniques, or Hastie et al. (2001) or Lee (2004) for more in-depth treatments.) However it should be kept in mind that all these techniques share a common core principle, the determination of the posterior belief based on Bayes' rule.

Simplicity and likelihood from a Bayesian perspective

The *likelihood principle* in perceptual theory is the idea that the brain aims to select the hypothesis that is most likely to be true in the world.[7] Recently Bayesian inference has been held up as the ultimate realization of this principle (Gregory 2006). Historically, the likelihood principle has been contrasted with the *simplicity* or *minimum principle*, which holds that the brain will select the simplest hypothesis consistent with sense data (Hochberg and McAlister 1953; Leeuwenberg and Boselie 1988). Simplicity too can be defined in a variety of ways, which has led to an inconclusive debate in which examples purporting to illustrate the preference for simplicity over likelihood, or vice versa, could be dissected without clear resolution (Hatfield and Epstein 1985; Perkins 1976).

[7] This should not be confused with what statisticians call the likelihood principle, a completely different idea. The statistical likelihood principle asserts that the data should influence our belief in a hypothesis only via the probability of those data conditioned on the hypothesis (i.e. the likelihood). This principle is universally accepted by Bayesians; indeed the likelihood is the only term in Bayes' rule that involves the data. But it is violated by classical statistics, where, for example, the significance of a finding depends in part on the probability of data that did *not* actually occur in the experiment. For example, when one integrates the tail of a sampling distribution, one is adding up the probability of many events that did not actually occur.

More recently, Chater (1996) has argued that simplicity and likelihood are two sides of the same coin, for several reasons that stem from Bayesian arguments. First, basic considerations from information theory suggest that more likely propositions are automatically simpler in that they can be expressed in more compact codes. Specifically, Shannon (1948) showed that an optimal code—meaning one that has minimum expected code length—should express each proposition A in a code of length proportional to the negative log probability of A, i.e. $-\log p(A)$. This quantity is sometimes referred to as the *surprisal*, because it quantifies how 'surprising' the message is (larger values indicate less probable outcomes), or as the *description length* (DL), because it also quantifies how many symbols it occupies in an optimal code (longer codes for more unusual messages). Just as in Morse code (or for that matter approximately in English) more frequently used concepts should be assigned shorter expressions, so that the total length of expressions is minimized on average. Because the proposition with maximum posterior probability (the MAP) also has *minimum* negative log posterior probability, the MAP hypothesis is also the minimum DL (MDL) hypothesis. More specifically, while in Bayesian inference the MAP hypothesis is the one that maximizes the *product* of the prior and the likelihood $p(H)p(D|H)$, in MDL the winning hypothesis is the one that minimizes the *sum* of the DL of the model plus the DL of the data as encoded via the model [$-\log p(H) - \log p(D|H)$, a sum of logs having replaced a product]. In this sense the simplest interpretation is necessarily also the most probable—though it must be kept in mind that this easy identification rests on the perhaps tenuous assumption that the underlying coding language is optimal.

More broadly, Bayesian inference tends to favour simple hypotheses even without any assumptions about the optimality of the coding language.[8] This tendency, sometimes called 'Bayes Occam' (after Occam's razor, a traditional term for the preference for simplicity), reflects fundamental considerations about the way prior probability is distributed over hypotheses (see MacKay 2003). Assuming that the hypotheses H_i are mutually exclusive, then their total prior necessarily equals one ($\sum_i p(H_i) = 1$), meaning simply that the observer believes that one of them must be correct. This in turn means that models with more parameters must distribute the same total prior over a larger set of specific models (combinations of parameter settings) inevitably requiring each model (on average) to be assigned a *smaller* prior. That is, more highly parametrized models—models that can express a wider variety of states of nature—necessarily assign lower priors to each individual hypothesis. Hence in this sense Bayesian inference automatically assigns lower priors to more complex models and higher priors to simple ones, thus enforcing a simplicity metric without any mechanisms designed especially for the purpose. This is really an instance of the ubiquitous *bias–variance* tradeoff, that is, the tradeoff between the fit to the data (which benefits from more complex hypotheses) and generalization to future data (which is impaired by more complex hypotheses; see Hastie et al. 2001). Bayesians argue that Bayes' rule provides an ideal solution to this dilemma because it determines the *optimal* combination of data fit (reflected in the likelihood) and bias (reflected in the prior).

Indeed the link between probability and complexity is fundamental to information theory, and also leads to an alternative 'subjectivist' method for constructing priors. Kolmogorov (1965) and Chaitin (1966) introduced a *universal* measure of complexity (now usually called Kolmogorov complexity) which in a technical sense is invariant to differences in the language used to express messages (see Li and Vitányi 1997). This means that just as DL can be thought of as $-\log p(H)$, $p(H)$ can be defined as (proportional to) $2^{-K(H)}$ where $K(H)$ is the Kolmogorov complexity of the hypothesis H (see Cover and Thomas 1991). Solomonoff (1964) first observed that this defines a

[8] The simplest law is chosen because it is most likely to give correct predictions' (Jeffreys 1939/1961, p. 4).

'universal prior', assigning high priors to simple hypotheses and low priors to complex ones in a way that is internally consistent and invariant to coding language—another way in which simplicity and Bayesian inference are intertwined (see Chater 1996).

Though the close relationship between simplicity and Bayesian inference is widely recognized, the exact nature of the relationship is more controversial. Bayesians regard the calculation of the Bayesian posterior as fundamental, and the simplicity principle as merely a heuristic whose value derives from its correspondence to Bayes' rule. The originators of MDL and information-theoretic statistics (e.g. Akaike 1974; Rissanen 1978; Wallace 2004) take the opposite view, regarding the minimization of complexity (DL or related measures) as the more fundamental principle and dismissing as naïve some of the assumptions underlying Bayesian inference (see Burnham and Anderson 2002; Grünwald 2005). This debate roughly parallels the controversy in the perception literature over simplicity and likelihood (see Feldman 2009; van der Helm 2013).

Decision Making and Loss Functions

Bayes' rule dictates how belief should be distributed among hypotheses. But a full account of Bayesian decision making requires that we also quantify the *consequences* of each potential decision, usually called the *loss function* (or *utility function* or *payoff matrix*). For example, misclassifying heartburn as a heart attack costs money in wasted medical procedures, but misclassifying a heart attack as heartburn may cost the patient his or her life. Hence the posterior belief in the two hypotheses (heart attack or heartburn) is not sufficient by itself to make a rational decision: one must also take into account the cost (loss) of each outcome, including both ways of misclassifying the symptoms as well as both ways of classifying them correctly. More broadly, each combination of an action and a state of nature entails a particular cost, usually thought of as being given by the nature of the problem. Bayesian decision theory dictates that the agent select the action that minimizes the (expected) loss—that is, the outcome which (according to the best estimate, the posterior) maximizes the benefit to the agent.

Different loss functions entail different rational choices of action. For example, if all incorrect responses are equally penalized, and correct responses not penalized at all (called *zero–one* loss) then the MAP is the rational choice, because it is the one most likely to avoid the penalty. (This is presumably the basis of the canard that Bayesian theory requires selection of the maximum posterior hypothesis, which is correct only for zero–one loss, and generally incorrect otherwise.) Other loss functions entail other minimum-loss decisions: for example under some circumstances quadratic loss (e.g. loss proportional to squared error) is minimized at the posterior mean (rather than the mode, which is the MAP), and other loss functions are minimized at the posterior median (Lee 2004).

Bayesian models of perception have primarily focused on simple estimation without consideration of the loss function, but this is undesirable for several reasons (Maloney 2002). First, perception in the context of real behaviour subserves action, and for this reason in the last few decades the perception literature has evolved towards an increasing tendency to study perception and action in conjunction. Second, more subtly, it is essential to incorporate a loss function in order to understand how experimental data speak to Bayesian models. Subjects' responses are not, after all, pure expressions of posterior belief, but rather are choices that reflect both belief and the expected consequences of actions. For example, in experiments, subjects implicitly or explicitly develop expectations about the relative cost of right and wrong answers, which help guide their actions. Hence in interpreting response data we need to consider both the subjects' posterior belief and their perceptions of payoff. Most experimental data offered in support of Bayesian models actually show *probability matching* behaviour, that is, responses drawn in proportion to their posterior probability, referred to by Bayesians as *sampling from the posterior*. Again, only zero–one loss would require rational subjects to choose the MAP response on every trial, so probability matching generally rules out zero–one loss (but obviously does not rule out Bayesian models more generally). The choice of loss functions in real situations probably depend on details of the task, and remains a subject of research.

Loss functions in naturalistic behavioural situations can be arbitrarily complex, and it is not generally understood either how they are apprehended or how human decision making takes them into account. Trommershauser et al. (2003) explored this problem by imposing a moderately complex loss function on their subjects in a simple motor task; they asked their subjects to touch a target on a screen that was surrounded by several different penalty zones structured so that misses in one direction cost more than misses in the other direction. Their subjects were surprisingly adept at modulating their taps so that expected loss (penalty) was minimized, implying a detailed knowledge of the noise in their own arm motions and a quick apprehension of the geometry of the imposed utility function (see also Trommershauser et al. 2008).

Where do the Hypotheses Come From?

Another fundamental problem for Bayesian inference is the source of the hypotheses. Bayesian theory provides a method for quantifying belief in each hypothesis, but it does not provide the hypothesis set H, nor any principled way to generate it. Traditional Bayesians are generally content to assume that some member of the H lies sufficiently 'close' to the truth, meaning that it approximates reality within some acceptable margin of error. Such assumptions are occasionally criticized as naïve (Burnham and Anderson 2002).

But the application of Bayesian theory to problems in perception and cognition elevates this issue to a more central epistemological concern. Intuitively, we assume that the real world has a definite state which perception either does or does not reflect. If, however, the hypothesis set H does not actually contain the truth—and Bayesian theory provides no reason to believe it does—then it may turn out that *none* of our perceptual beliefs may be literally true, because the true hypothesis was never under consideration (cf. Hoffman 2009; Hoffman and Singh 2012). In this sense, the perceived world might be *both* a rational belief (in that the assignment of posterior belief follows Bayes' rule) and, in a very concrete sense, a grand hallucination (because none of the resulting beliefs are true).

Thus while Bayesian theory provides an optimal method for using all information available to determine belief, it is not magic; the validity of its conclusions is limited by the validity of its premises. Indeed this point is well understood by Bayesians, who often argue that all inference is based on assumptions (see Jaynes 2003; MacKay 2003). (This is in contrast to frequentists, who aspired to a science of inference free of subjective assumptions.) But it gains special significance in the context of perception, because perceptual beliefs are the very fabric of subjective reality.

Competence Versus Performance

Bayesian inference is a rational, idealized mathematical framework for determining perceptual beliefs, based on the sense data presented to the system coupled with whatever prior knowledge the system brings to bear. But it does not, in and of itself, specify computational mechanisms for actually calculating those beliefs. That is, Bayesian inference quantifies exactly how strongly the system should believe each hypothesis, but does not provide any specific mechanisms whereby the system might arrive at those beliefs. In this sense, Bayesian inference is a *competence theory* (Chomsky's term) or a *theory of the computation* (Marr's term), meaning it is an abstract specification of the function to be computed rather than the means to compute it. Many theorists, concurring with Marr and Chomsky, argue that competence theories play a necessary role in cognitive theory, parallel to but distinct from that of process accounts. Competence theories by their nature abstract away from details of implementation and help connect the computations that experiments uncover with the underlying problem those computations help solve. Conversely, some psychologists denigrate competence theories as abstractions that are irrelevant to real psychological processes (Rumelhart et al. 1986), and indeed Bayesian models have been criticized on these grounds (McClelland et al. 2010; Jones and Love 2011).

But to those sympathetic to competence accounts, rational models have an appealingly 'explanatory' quality precisely because of their optimality. Bayesian inference is, in a well-defined sense, the *best* way to solve whatever decision problem the brain is faced with. Natural selection pushes organisms to adopt the most effective solutions available, so evolution should tend to favour Bayes-optimal solutions whenever possible (see Geisler and Diehl 2002). For this reason, any phenomenon that can be understood as part of a Bayesian model automatically inherits an evolutionary rationale.

Conclusions

In a sense, perception and Bayesian inference are perfectly matched. Perception is the process by which the mind forms beliefs about the outside world on the basis of sense data combined with prior knowledge. Bayesian inference is a system for determining what to believe on the basis of data and prior knowledge. Moreover, the rationality of Bayesian inference means that perceptual beliefs that follow the Bayesian posterior are, in a well-defined sense, optimal given the information available. This optimality has been argued to provide a selective advantage in evolution (Geisler and Diehl 2002), driving our ancestors towards Bayes-optimal percepts. Moreover optimality helps explain *why* the perceptual system, notwithstanding its many apparent quirks and special rules, works the way it does—because these rules approximate the Bayesian posterior. Moreover, the comprehensive nature of the Bayesian framework allows it to be applied to any problem that can be expressed probabilistically. All these advantages have led to a tremendous increase in interest in Bayesian accounts of perception in the last decade.

Still, a number of reservations and difficulties must be noted. First, to some researchers a commitment to a Bayesian framework seems to involve a dubious assumption that the brain is rational. Many psychologists regard the perceptual system as a hodge-podge of hacks, dictated by accidents of evolutionary history and constrained by the exigencies of neural hardware. While to its advocates the rationality of Bayesian inference is one of its main attractions, to sceptics the hypothesis of rationality inherent in the Bayesian framework seems at best empirically implausible and at worse naïve.

Second, more specifically, the essential role of the prior poses a puzzle in the context of perception, where the role of prior knowledge and expectations (traditionally called 'top-down' influences) has been debated for decades. Indeed there is a great deal of evidence (see Pylyshyn 1999) that perception is singularly uninfluenced by certain kinds of knowledge, which at the very least suggests that the Bayesian model must be limited in scope to an encapsulated perception module walled off from information that an all-embracing Bayesian account would deem relevant.

Finally, many researchers wonder if the Bayesian framework is too flexible to be taken seriously, potentially encompassing any conceivable empirical finding. However while Bayesian accounts are indeed quite adaptable, any specific set of assumptions about priors, likelihoods, and loss functions provides a wealth of extremely specific empirical predictions, which in many specific perceptual domains have been validated experimentally.

Hence notwithstanding all of these concerns, to its proponents Bayesian inference provides something that perceptual theory has never really had before: a 'paradigm' in the sense of Kuhn (1962)—that is, an integrated, systematic, and mathematically coherent framework in which to pose basic scientific questions and evaluate potential answers. Whether or not the Bayesian approach turns out to be as comprehensive or empirically successful as its advocates hope, this represents a huge step forward in the study of perception.

References

Akaike, H. (1974). 'A new look at the statistical model identification'. *IEEE Trans Automat Contr 19*(6): 716–723.

Bayes, T. (1763). 'An essay towards solving a problem in the doctrine of chances'. *Phil Trans R Soc. Lond 53*: 370–418.

Bülthoff, H. H. and A. L. Yuille (1991). 'Bayesian models for seeing shapes and depth'. *Comm Theor Biol 2*(4): 283–314.

Burge, J., C. C. Fowlkes, and M. S. Banks (2010). 'Natural-scene statistics predict how the figure- ground cue of convexity affects human depth perception'. *J Neurosci 30*(21): 7269–7280.

Burnham, K. P. and D. R. Anderson (2002). *Model Selection and Multi-model Inference: a Practical Information-theoretic Approach* (New York: Springer).

Carruthers, P., S. Laurence, and S. Stich (2005). *The Innate Mind: Structure and Contents* (Oxford: Oxford University Press).

Chaitin, G. (1966). 'On the length of programs for computing finite binary sequences'. *J Assoc Comput Machin 13*(4): 547–569.

Chater, N. (1996). 'Reconciling simplicity and likelihood principles in perceptual organization'. *Psychol Rev 103*(3): 566–581.

Claessens, P. M. E. and J. Wagemans (2008). 'A Bayesian framework for cue integration in multistable grouping: proximity, collinearity, and orientation priors in zigzag lattices'. *J Vision 8*(7): 1–23.

Cohen, E. H., M. Singh, and L. T. Maloney (2008). 'Perceptual segmentation and the perceived orientation of dot clusters: the role of robust statistics'. *J Vision 8*(7): 1–13.

Compton, B. J. and G. D. Logan (1993). 'Evaluating a computational model of perceptual grouping by proximity'. *Percept Psychophys 53*(4): 403–421.

Cover, T. M. and J. A. Thomas (1991). *Elements of Information Theory* (New York: John Wiley).

Cox, R. T. (1961). *The Algebra of Probable Inference* (London: Oxford University Press).

Dakin, S. (2013). 'Statistical regularities'. In *Handbook of Perceptual Organization*, edited by J. Wagemans. (This volume, forthcoming.)

De Winter, J. and J. Wagemans (2006). 'Segmentation of object outlines into parts: a large-scale integrative study'. *Cognition*, *99*(3): 275–325.

Earman, J. (1992). *Bayes or Bust?: a Critical Examination of Bayesian Confirmation Theory* (Cambridge, MA: MIT Press).

Elder, J. (2013). 'Contour grouping'. In *Handbook of Perceptual Organization*, edited by J. Wagemans. (This volume, forthcoming.)

Elman, J., A. Karmiloff-Smith, E. Bates, M. Johnson, D. Parisi, and K. Plunkett (1996). *Rethinking Innateness: a Connectionist Perspective on Development* (Cambridge, MA: MIT Press).

Feldman, J. (1997). 'Curvilinearity, covariance, and regularity in perceptual groups'. *Vision Res 37*(20): 2835–2848.

Feldman, J. (2001). 'Bayesian contour integration'. *Percept Psychophys 63*(7): 1171–1182.

Feldman, J. (2009). 'Bayes and the simplicity principle in perception'. *Psychol Rev 116*(4): 875–887.

Feldman, J. (2013). 'Tuning your priors to the world'. *Top Cogn Sci 5*(1): 13–34.

Feldman, J. and M. Singh (2006). 'Bayesian estimation of the shape skeleton'. *Proc Natl Acad Sci USA 103*(47): 18014–18019.

Feldman, J., Singh, M., and Froyen, V. (2013). 'Perceptual grouping as Bayesian mixture estimation'. In *Oxford Handbook of Computational Perceptual Organization* edited by Gepshtein, Maloney and Singh, *forthcoming*.

de Finetti, B. (1970/1974). *Teoria delle Probabilita 1* (Turin: Giulio Einaudi). [Translated by A. Machi and A. Smith, 1990 as *Theory of Probability 1* (Chichester: John Wiley and Sons).]

Fisher, R. (1925). *Statistical Methods for Research Workers* (Edinburgh: Oliver and Boyd).

Geisler, W. S. and R. L. Diehl (2002). 'Bayesian natural selection and the evolution of perceptual systems'. *Phil Trans R Soc Lond B 357*: 419–448.

Geisler, W. S., J. S. Perry, B. J. Super, and D. P. Gallogly (2001). 'Edge co-occurrence in natural images predicts contour grouping performance'. *Vision Res 41*: 711–724.

Girshick, A. R., M. S. Landy, and E. P. Simoncelli (2011). 'Cardinal rules: visual orientation perception reflects knowledge of environmental statistics'. *Nat Neurosci 14*(7): 926–932.

Gregory, R. (2006). 'Editorial essay'. *Perception 35*: 143–144.

Griffiths, T. L. and A. L. Yuille (2006). 'A primer on probabilistic inference'. *Trends Cogn Sci 10*(7). Supplement to special issue on Probabilistic Models of Cognition. Available at: <http://www.stat.ucla.edu/~yuille/pubs/ucla/A204_tgriffiths_chater2007.pdf>)

Grünwald, P. D. (2005). 'A tutorial introduction to the minimum description length principle'. In *Advances in Minimum Description Length: Theory and Applications*, edited by P. D. Grünwald, I. J. Myung, and M. Pitt.(Cambridge, MA: MIT Press).

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (New York: Springer).

Hatfield, G. and W. Epstein (1985). 'The status of the minimum principle in the theoretical analysis of visual perception'. *Psychol Bull 97*(2): 155–186.

Hochberg, J. and E. McAlister (1953). 'A quantitative approach to figural "goodness" '. *J Exp Psychol 46*: 361–364.

Hoffman, D. D. (2009). 'The user-interface theory of perception: natural selection drives true perception to swift extinction'. In *Object Categorization: Computer and Human Vision Perspectives*, edited by S. Dickinson, M. Tarr, A. Leonardis, and B. Schiele.(Cambridge: Cambridge University Press).

Hoffman, D. D. and M. Singh (2012). Computational evolutionary perception. *Perception. 41*: 1073–1091.

Howie, D. (2004). *Interpreting Probability: Controversies and Developments in the Early Twentieth Century* (Cambridge: Cambridge University Press).

Jaynes, E. T. (1982). 'On the rationale of maximum-entropy methods'. *Proc IEEE 70*(9): 939–952.

Jaynes, E. T. (2003). *Probability Theory: the Logic of Science* (Cambridge: Cambridge University Press).

Jeffreys, H. (1939/1961). *Theory of Probability, 3rd edn* (Oxford: Clarendon Press).

Jones, M. and B. C. Love (2011). 'Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition'. *Behav Brain Sci 34*: 169–188.

Juni, M. Z., M. Singh, and L. T. Maloney (2010). 'Robust visual estimation as source separation'. *J Vision 10*(14): 2; doi: 10.1167/10.14.2.

Kersten, D., P. Mamassian, and A. Yuille (2004). 'Object perception as Bayesian inference'. *Ann Rev Psychol 55*: 271–304.

Knill, D. C. and W. Richards (eds) (1996). *Perception as Bayesian Inference* (Cambridge: Cambridge University Press).

Kolmogorov, A. N. (1965). 'Three approaches to the quantitative definition of information'. *Prob Inform Transmission 1*(1): 1–7.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press).

Lee, M. D. and E.-J. Wagenmakers (2005). 'Bayesian statistical inference in psychology: comment on Trafimow (2003)'. *Psychol Rev 112*(3): 662–668.

Lee, P. (2004). *Bayesian Statistics: an Introduction*, 3rd edn (Chichester: Wiley).

Leeuwenberg, E. L. J. and F. Boselie (1988). 'Against the likelihood principle in visual form perception'. *Psychol Rev 95*: 485–491.

Li, M. and P. Vitányi (1997). *An Introduction to Kolmogorov Complexity and its Applications* (New York: Springer).

McClelland, J. L., M. M. Botvinick, D. C. Noelle, D. C. Plaut, T. T. Rogers, M. S. Seidenberg, et al. (2010). 'Letting structure emerge: connectionist and dynamical systems approaches to understanding cognition'. *Trends Cogn Sci 14*: 348–356.

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

Maloney, L. T. (2002). 'Statistical decision theory and biological vision'. In *Perception and the Physical World: Psychological and Philosophical Issues in Perception*, edited by D. Heyer and R. Mausfeld, pp. 145–189 (New York: Wiley).

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (San Mateo, CA: Morgan Kauffman).

Perkins, D. (1976). 'How good a bet is good form?' *Perception 5*: 393–406.

Pylyshyn, Z. (1999). 'Is vision continuous with cognition? The case for cognitive impenetrability of visual perception'. *Behav Brain Sci 22*(3): 341–365.

Ramachandran, V. S. (1985). 'The neurobiology of perception'. *Perception 14*: 97–103.

Rissanen, J. (1978). 'Modeling by shortest data description'. *Automatica 14*: 465–471.

Rumelhart, D. E., J. L. McClelland, and G. E. Hinton (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Cambridge, MA: MIT Press).

Shannon, C. (1948). 'A mathematical theory of communication'. *Bell Syst Tech J 27*: 379–423.

Singh, M. (2013). 'Visual representation of contour geometry'. In *Handbook of Perceptual O rganization*, edited by J. Wagemans. (This volume, forthcoming.)

Singh, M. and J. M. Fulvio (2005). 'Visual extrapolation of contour geometry'. *Proc Natl Acad Sci USA 102*(3): 939–944.

Singh, M. and D. D. Hoffman (2001). 'Part-based representations of visual shape and implications for visual cognition'. In *From Fragments to Objects: Segmentation and Grouping in Vision, Advances in Psychology* Vol. *130*, edited by T. Shipley and P. Kellman, pp. 401–459 (New York: Elsevier).

Solomonoff, R. (1964). 'A formal theory of inductive inference: part II'. *Inform Control 7*: 224–254.

Stigler, S. M. (1983). 'Who discovered Bayes's theorem?' *Am Statistician 37*(4): 290–296.

Stigler S. M. (1986). *The History of Statistics: the Measurement of Uncertainty Before 1900* (Cambridge, MA: Harvard University Press).

Trommershauser, J., L. T. Maloney, and M. S. Landy (2003). 'Statistical decision theory and the selection of rapid, goal-directed movements'. *J Opt Soc Am A: Opt Image Sci Vis 20*(7): 1419–1433.

Trommershauser, J., L. T. Maloney, and M. S. Landy (2008). 'Decision making, movement planning and statistical decision theory'. *Trends Cogn Sci 12*(8): 291–297.

van der Helm, P. (2013). 'Simplicity in perceptual organization'. In *Handbook of Perceptual Organization*, edited by J. Wagemans. (This volume, forthcoming.)

Wallace, C. S. (2004). *Statistical and Inductive Inference by Minimum Message Length* (New York: Springer).

Weiss, Y., E. P. Simoncelli, and E. H.. Adelson (2002). 'Motion illusions as optimal percepts'. *Nat Neurosci 5*(6): 598–604.

Wilder, J., J. Feldman, and M. Singh (2011). 'Superordinate shape classification using natural shape statistics'. *Cognition 119*: 325–340.

Zucker, S. W., K. A. Stevens, and P. Sander (1983). 'The relation between proximity and brightness similarity in dot patterns'. *Percept Psychophys 34*(6): 513–522.