Rutgers Center for Cognitive Science Technical Report #80, 2004

Bayes and the Simplicity Principle in Perception

Jacob Feldman
Dept. of Psychology, Center for Cognitive Science
Rutgers University

Two governing principles of perceptual inference, the Likelihood principle and the Simplicity principle, have been much discussed by perceptual theorists, and often placed in opposition. Recently, Chater (1996) has argued that the two principles are actually consistent in that their decisions tend to agree asymptotically. This article seeks to relate in a more mathematically direct way what is arguably the most plausible version of a likelihood theory, Bayesian inference, with a recently-proposed generalized formal minimum principle (the maximum-depth rule of Minimal Model theory). Assuming qualitative information on the part of the observer, and equal prior probabilities among among all competing hypotheses, maximizing the Bayesian posterior probability turns out to be mathematically equivalent to choosing the maximum-depth interpretation from a hierarchical space of possible scene interpretations. That is, the maximum-depth rule is analytically equivalent to Bayes with a particular choice of priors. Thus this version of the Simplicity principle, as well as "full-blown Bayes," each constitute distinct points in a well-defined continuum of possible perceptual decision rules. In a very literal mathematical sense, the observer's position in this continuum—and, consequently, the perceptual decision rule it employs—reflect the nature of its tacit assumptions about the environment.

Simplicity vs. Likelihood principles in Perception

A recurrent theme in the study of human visual perception is the idea that the visual system selects the simplest interpretation consistent with the visual image—sometimes referred to as the *Simplicity principle*, sometimes by the Gestalt term *Prägnanz*, and sometimes as the *minimum principle*. The principle has taken many forms, from a relatively vague preference for the maximization of "regularity" (Kanizsa, 1979), to more concrete systems in which the image is described in some fixed coding language, and the interpretation whose code is of minimal length is selected (Hochberg & McAlister, 1953; Leeuwenberg, 1971; Buffart, Leeuwenberg, & Restle, 1981). Many phenomena of visual perception seem to be explained at least in part by appeals to the minimum principle, in one form or another. Yet at the same time many authors have been troubled over the motivation or justification of the principle (Hatfield & Epstein, 1985), paralleling an analogous debate about the rationale of Occam's razor in the selection of scientific theories (Quine, 1965; Sober, 1975).

Another well-known principle of perceptual inference, sometimes held up in opposition to the Simplicity principle, is the *Likelihood principle*: choose the interpretation most likely to be true. The rationale behind this idea seems relatively self-evident, in that it is clearly desirable (say, from an evolutionary point of view) for the organism to achieve veridical percepts of the world. Yet the mere statement of the principle begs the question of how the visual system actually determines the relative likelihood of various candidate interpretations, and hence it has not been clear exactly how the Likelihood principle might translate into concrete computational procedures.

Historically, the minimum principle and the likelihood principle have usually been regarded as competitors, or at least roughly contradictory (Perkins, 1976; Hatfield & Epstein, 1985; Leeuwenberg & Boselie, 1988; van der Helm, 2000). Recently, however Chater (1996), using mathematical arguments paralleling those from Minimum Description Length (MDL) theory (Rissanen, 1989), has shown that the two principles can be regarded as equivalent. Under very general assumptions, the visual interpretation whose description is of minimum length is, in fact, the one that is most likely to be the correct in an objective sense. This remarkable demonstration combines the most appealing aspects of both principles, giving the minimum principle a clear rationale

I am grateful to Rajesh Kasturirangan, Whitman Richards, Manish Singh, Matthew Stone, and Josh Tenenbaum for helpful discussions and comments. This research was supported in part by NSF SBR-9875175.

Please direct correspondence to Jacob Feldman, Dept. of Psychology & Center for Cognitive Science, Rutgers University - New Brunswick, Busch Campus, Piscataway, New Jersey, 08854, or by e-mail at jacob@ruccs.rutgers.edu.

(namely, veridicality) while suggesting a criterion by which the most likely interpretation can be identified (namely, minimum length).

Nevertheless Chater's demonstration is pitched at a very high level of abstraction. The argument connecting probability to complexity is based on the notion of *Kolmogorov complexity* (the length of the shortest computer program that could generate a given string). The beauty of the mathematics surrounding Kolmogorov complexity is that it doesn't depend on details of the coding language used—all so-called universal codes give approximately the same complexity value. But the flip side of this same universality is that while one can make *general* statements about the Kolmogorov complexity of a given string, you never know the specific value—the length of the actual shortest program; in fact it is uncomputable in general (see Schöning & Pruim, 1998 for a simple proof of this). The agreement between the probability of an interpretation and its complexity (like all statements about Kolmogorov complexity) is necessarily asymptotic: they tend to match in the limit as number of stimulus elements grows infinitely large. But for any given stimulus and any given coding language, the disagreement can be arbitrarily large, and thus potentially overshadow the agreement.[1] The exact discrepancy for realistic stimuli depends on the coding language, meaning that different coding languages may in practice achieve the most veridical conclusion with extremely different degrees of success.

Hence Chater's argument, while persuasive in an abstract sense, leaves open narrower—but crucial—questions such as the nature of the actual coding language used by the visual system, and the exact form of the associated minimization rule. The current paper seeks to demonstrate a stronger and more specific connection between a particular minimization rule recently proposed (the maximum-depth rule of Minimal Model theory; see Feldman, 1997c, 1997b, 1999, 2003) and probabilistically optimal inference, i.e. Bayesian theory. The conditions invoked in this rule are less general than in Chater's formulation, but are important in wide variety of perceptual situations, especially those involving perceptual organization, grouping, and the inference of three-dimensionality. Hence while consistent with Chater's general conclusion, the connections between minimum principles and Bayes described below give a more concrete account of why perceptually realistic simplicity minimization tends to yield the true state of the world.

Bayesian formulation

Bayesian theory is particularly attractive formulation of the likelihood principle in perception (for examples see Bülthoff & Yuille, 1991; Feldman, 2001; Knill & Richards, 1996; Landy, Maloney, Johnston, & Young, 1995). Its attractiveness stems largely from the fact that it provides provably optimal inferences under conditions of uncertainty (see Jaynes, 1957/1988 for an especially lucid demonstration of this). Thus Bayes provides "rational" perceptual decisions.

In Bayesian theory, the subjective belief in a particular hypothesis given particular data is associated with the *posterior probability*, i.e. the conditional probability of the hypothesis given the data. In the context of visual perception, the data is the visual image I and the hypotheses are the various scene interpretations among which the observer will choose. In what follows I will assume an image I chosen from an image space $\mathbf{I}$, and a finite[2] set $\mathcal{S} = \{S_1, S_2 \ldots S_n\}$ of distinct candidate interpretations, i.e. categories of distal scenes. Each interpretation S_i has an associated *likelihood function* $p(I|S_i)$ indicating how likely a given possible image is under that hypothesis; and each scene occurs with a certain scalar prior probability $p(S_i)$. The priors must sum to unity ($\sum_i p(S_i) = 1$), and each likelihood function integrates to unity over $\mathbf{I}$ ($\int_{\mathbf{I}} p(I|S_i) d\mathbf{I} = 1$).

By Bayes' rule, given image I, the posterior probability that the interpretation S_i is correct is

$$p(S_i|I) = \frac{p(S_i)p(I|S_i)}{\sum_j p(S_j)p(I|S_j)} \tag{1}$$

Because the denominator is the same for all interpretations, the winning interpretation will be that S_i that maximizes the numerator

$$p(S_i)p(I|S_i), \tag{2}$$

i.e. the product of the prior and the likelihood. (Notice this depends both on the likelihood of the given image under S_i, as well as the prior probability $p(S_i)$ that S_i was true before the image was observed.) Another definition that will be important below is the *support* of an interpretation S_i, defined as the region of $\mathbf{I}$ where S_i's likelihood is non-zero[3] and denoted $\sigma(S_i)$,

$$\sigma(S_i) = \{I \in \mathbf{I} | p(I|S_i) > 0\}. \tag{3}$$

The support of a scene model S_i is the set of images that *could* have been produced by it.

A fully Bayesian observer would select an interpretation by first computing the product (2) for all S_i, and then selecting the largest.[4] Bayesian theory has at times been criticized for requiring this global maximization, which may involve a great deal of computation.

---

[1] Technically, the agreement between two complexity measures is bounded by a constant that does not depend on the stimulus. But the size of the constant depends on the amount of information needed to specify the design of a complete Turing machine, enabling one universal Turing machine to simulate another. Because Turing machines, or Turing-equivalent computing systems such as human brains, can be arbitrarily large and complex, the constant difference between coding lengths for two machines can be arbitrarily large.

[2] We will not generally require that $\mathcal{S}$ be finite, but it is more notationally convenient to assume so.

[3] The support of S can alternatively be defined as the region where $p(I|S) > \varepsilon$ for some arbitrarily small number ε; this makes no difference in the ensuing theory.

[4] Technically this is the *maximum-a-posteriori* (MAP) interpretation. This is the most straightforward, but not the only, way of choosing a single best interpretation in Bayesian theory.

Hierarchies of candidate interpretations

In some perceptual situations, though, pre-existing formal relations among the S_i may make such an elaborate computation unnecessary. The argument developed below shows that when the set $\mathcal{S}$ of candidate interpretations is *hierarchical* in a way that perceptual interpretations often are, Bayes' rule reduces to a simpler form that requires far less computation. This simpler form turns out to be equivalent to a particular minimum rule in the literature, thus showing in a new way the relationship between the Likelihood and Simplicity principles.

Many discussions and controversies about Bayesian reasoning, and in particular its applicability to human intuitions, revolve around the way the prior probabilities and likelihood functions are assigned. The discussion below focuses on how probability is distributed among the priors $p(S_i)$, but analogous arguments can be made about the way probability is assigned within each likelihood function $p(I|S_i)$.

In most treatments, the set of candidate interpretations $\mathcal{S}$ is treated as "flat", i.e. with no hierarchy among the interpretations. However in many perceptual situations, some candidate interpretations may actually be *special cases* of others. For example, the interpretation S_2 (*I is Fido*) is a proper subset of another interpretation S_1 (*I is a dog*): if S_2 holds, then by its very nature S_1 must hold as well. I will denote this situation by $S_2 \rightarrow S_1$. In probabilistic terms the relation "$\rightarrow$" can be defined by

$$S_2 \rightarrow S_1 \text{ iff } \sigma(S_2) \subsetneq \sigma(S_1), \tag{4}$$

i.e. whenever S_2 is possible (has non-zero likelihood) S_1 is possible too, but not necessarily vice versa (cf. Bennett, Hoffman, & Murthy, 1993). The situation is graphically depicted in Fig. 1. The relation $\rightarrow$ can hold hierarchically as well, with interpretations embedded within interpretations within interpretations; or some interpretations may be embedded in others, while still others in $\mathcal{S}$ are disjoint. The structure of the interpretation space can be depicted diagrammatically; Fig. 2a show some of the possibilities. I will call interpretation spaces in which some of the interpretations are embedded in others *hierarchical*.

Examples of hierarchical interpretation spaces abound in the perceptual literature, often involving the notion of "non-accidentalness." Non-accidental relations (Witkin & Tenenbaum, 1983; Lowe, 1987) are geometric relations between image elements that are unlikely to occur by accident, and which consequently are perceptually salient (Wagemans, 1992). Examples include parallelism, collinearity (Caelli & Umansky, 1976; Smits & Vos, 1986; Feldman, 1997a), and skew symmetry (Kanade, 1981; Wagemans, 1993) (see Fig. a). Non-accidental configurations can be regarded as special cases of, and hence embedded in, generic relations: for example the set of line segment pairs that are *parallel* is a subset of the set of *all* line segment pairs. Moreover non-accidental relations can be embedded hierarchically: for example collinear lines segments are necessarily parallel, though not vice versa (Fig. a). Geons, the part representation units in the well-known theory of Biederman (1987), which are built out of non-accidental relations, are another example of embedded interpretation categories (Fig. b). Again the main idea in these examples is simply that some candidate interpretations are special cases of others.

Neutral prior probabilities

We would like to consider how hierarchical interpretation spaces work when placed in a Bayesian framework. As suggested above, it turns out that their embedded structure makes Bayesian decisions reduce to a particularly simple form. Before we start, we must consider how prior probabilities ought to be assigned to interpretations in a hierarchy.

In the perceptual literature a great deal of emphasis has been put on the idea of "neutral" prior probabilities: that is, prior assumptions that entail as the least possible commitment on the part of the perceiver. A formal scheme that accounts for human perception without making any ad hoc assumptions about the perceiver's knowledge—for example, by positing neutral priors—would seem especially convincing.

Non-accidental inference has often been held up as such a system, and there appears to be a widespread belief in the literature that non-accidental interpretation is justified by Bayesian theory under very neutral probabilistic assumptions. There are, however, several different ways of assigning prior probability "neutrally" or uniformly, and it turns out that non-accidental inference is justified under some ways but not others.

For explicitness, consider two ways to assign priors uniformly: over the image space $\mathbf{I}$; or over the hypothesis space $\mathcal{S}$. I will refer to these two distinct assumptions as two versions of the *Equal Priors Assumption* (EPA):

EPA-$\mathbf{I}$: $p(I)$ is constant for all $I \in \mathbf{I}$.[5]
EPA-$\mathcal{S}$: $p(S)$ is constant for all $S \in \mathcal{S}$.

These two assumptions are generally *not* equivalent, unless the supports of all interpretations have equal areas in the image space (which is emphatically not true in many important cases, such as non-accidental properties).

Assumption EPA-$\mathbf{I}$ is more truly "neutral" because it makes no assumptions about what hypotheses the observer is entertaining. EPA-$\mathcal{S}$ by contrast explicitly refers to the specific hypotheses under consideration, and hence assigns priors in a different way depending on what these hypotheses are. EPA-$\mathcal{S}$ in effect "squeezes" an equal amount of probability mass into each interpretation, regardless of its intrinsic size in the space or its relation to other interpretations—in the case of non-accidental properties, squeezing that same amount of mass even into some areas that are *infinitesimally*

[5] Note that the integral of such a prior over any infinite image space $\mathbf{I}$ will be infinite (and not unity as required), making this an "improper prior." Traditional Bayesian theory includes methods for dealing with this situation that make it unproblematic for purposes of the current paper (Box & Tiao, 1973). In any case this version of the EPA is not pursued in what follows.

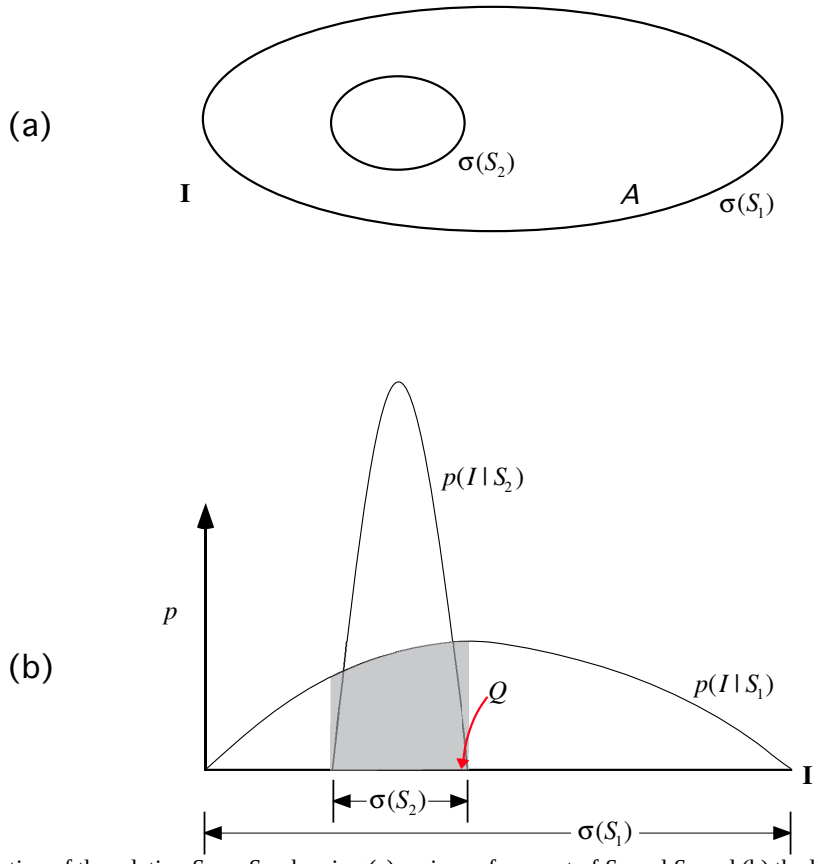

Figure 1. Schematic depiction of the relation $S_2 \rightarrow S_1$, showing (a) regions of support of S_1 and S_2 and (b) the likelihood functions $p(I|S_1)$ and $p(I|S_2)$.

smaller than others. This results in a highly non-uniform distribution of probability mass over the image space (i.e. contradicting EPA-**I**). For example straight cylinders and curved cylinders would be assigned the same prior, despite the fact that the former is a lower-dimensional subspace of the latter, and hence in a sense contains infinitely fewer possible images. Presumably, perceptual theorists usually have something like EPA-**I**—uniform over the whole image space—in mind when they refer to a truly neutral prior.

Surprisingly, though, and contrary to the widespread view, under assumption EPA-**I**, Bayes does *not* support non-accidental inference (Jepson & Richards, 1992). Rather, under EPA-**I**, Bayes does not favor *any* interpretation over any other.[6] Thus under a totally neutral prior Bayes gives no support at all for what we see—and thus no account of the plain facts of perception. Bayes only accords with human judgments if the special or non-accidental configurations are given elevated priors in some way, such as by EPA-$\mathcal{S}$ (see also Richards, Jepson, & Feldman, 1996 for discussion). Thus EPA-**I** does not appear to be psychologically plausible. Hence in what follows I will generally assume EPA-$\mathcal{S}$.

Although this means that EPA-$\mathcal{S}$ is not completely neutral in the absolutely interpretation-independent sense of EPA-**I**, it should be understood that it is legitimately neutral in a different sense: having assumed that certain types of scene structures $S_1, S_2, \ldots$ occur in the world, the perceiver now proceeds to make no distinction among them as to their prior probability.

[6] A simple example to make this clear: assume two interpretations S_1 and S_2, with S_1 the larger and S_2 the non-accidental special case, and let ε be the probability that S_2 happens "by accident" when S_1 is really correct. Then when the non-accidental relation S_2 happens, assuming EPA-**I**, the posterior for interpretation S_2 is

$$\begin{aligned} p(S_2|I) &= \frac{p(I|S_2)p(S_2)}{p(I|S_2)p(S_2)+p(I|S_1)p(S_1)} \\ &= \frac{(1)(\frac{\varepsilon}{1+\varepsilon})}{(1)(\frac{\varepsilon}{1+\varepsilon})+(\varepsilon)(\frac{1}{1+\varepsilon})} \\ &= \frac{1}{1+1} \\ &= \frac{1}{2} \end{aligned}$$

The posterior is equal between the two hypotheses, and, moreover, it does not depend on ε. In effect, under EPA-**I**, the fact that S_2 explains the image better is exactly balanced by the fact that S_2 is less likely to occur in the first place. Regardless of the magnitude of ε, the non-accidental inference is not preferred. *Equal priors give equal posteriors.*

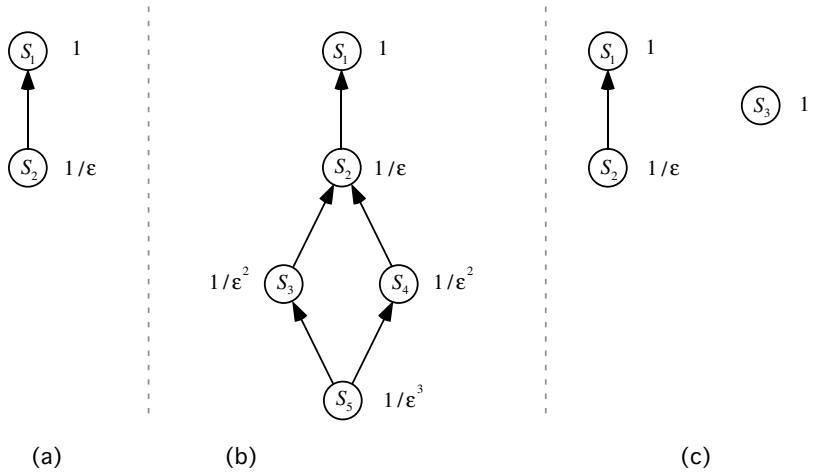

Figure 2. Various possible hierarchical interpretation space diagrams. (a) $S_2 \rightarrow S_1$ (b) $S_2 \rightarrow S_1, S_3 \rightarrow S_2, S_4 \rightarrow S_2, S_5 \rightarrow S_3, S_5 \rightarrow S_4$, (b) $\mathcal{S} = \{S_1, S_2, S_3\}, S_2 \rightarrow S_1$. Adjacent to each interpretation is its likelihood ratio (see Eqs. 18–20 and surrounding text).

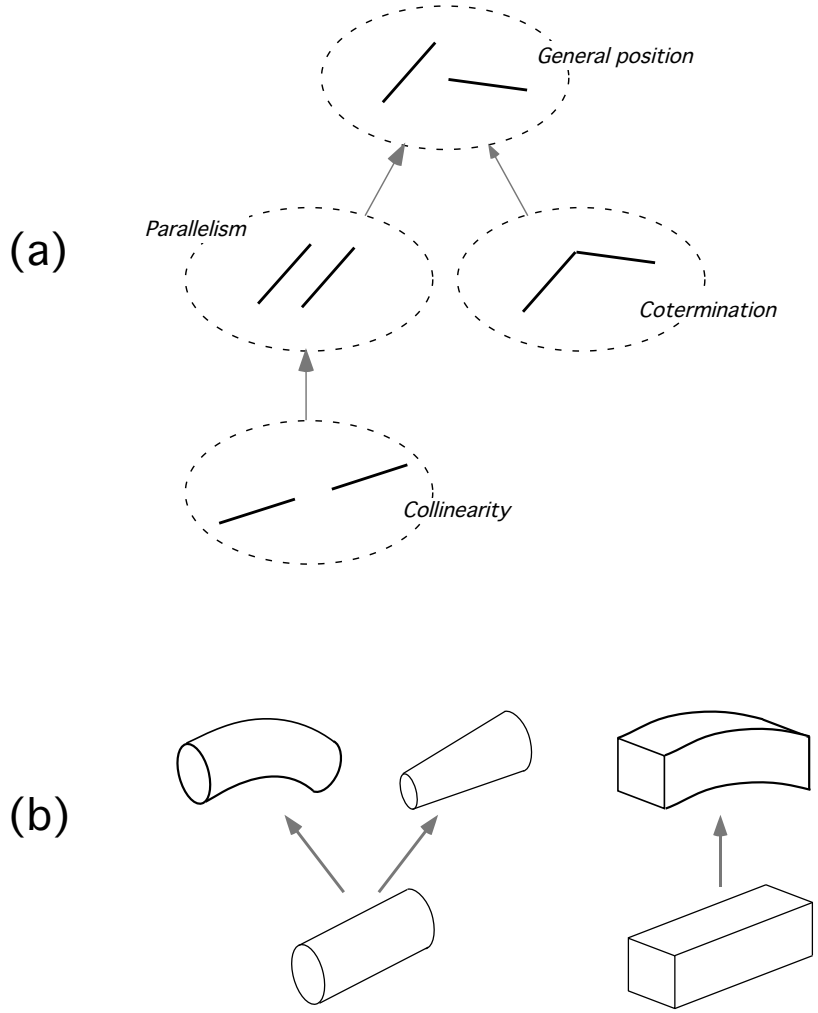

Figure 3. Examples of well-known hierarchical interpretation spaces from the literature: (a) non-accidental properties (Witkin & Tenenbaum, 1983; Lowe, 1987); (b) geons (a partial set) (Biederman, 1987). In each case lower interpretations are "special cases" of their upper neighbors in the diagram. Note that the geons form a disjoint space with hierarchical components, like Fig. 2c.

The benefit of making this assumption lies in its computational consequences: assumption EPA-$\mathcal{S}$ makes computation of the posterior probability very simple, because the prior probabilities now cancel out and become irrelevant. As discussed above, under full Bayes, the observer ought to prefer the interpretation S_i that maximizes the product of the prior and likelihood,

$$p(S_i)p(I|S_i). \tag{5}$$

But under EPA-$\mathcal{S}$, all the $p(S_i)$ are equal,

$$p(S_1) = p(S_2) = \ldots = p(S_n) = 1/n. \tag{6}$$

Hence the equal priors $1/n$ cancel out of the comparison, and the winner will be the interpretation that simply maximizes the likelihood $p(I|S_i)$.

Qualitativeness

Now imagine an observer who has access *only* to qualitative information about the image. For example imagine that the observer does not know I precisely enough to compute its likelihood $p(I|S_i)$, but *does* know which S_i's assigned I nonzero likelihood—i.e., which interpretations were consistent with the image. This assumption actually reflects the way we speak informally about non-accidental properties. For example when we ask how an observer would interpret parallel lines, what we really mean is: what would the observer think given only the knowledge that the observed line segments are parallel, but without knowing (or perhaps knowing but ignoring) any other information about their geometry? That is, what do we do when we know that the image satisfies a certain model (parallelness) but don't know exactly what its likelihood is under that model (or under any other model)?

Formally, in this situation we don't know exactly where in $\mathbf{I}$ the image I falls—and hence we don't have enough information to do Bayes "properly"—but we *do* know in which support regions $\sigma(S_i)$ it falls—that is, which interpretations are at least consistent with I. I will refer to this assumption as *qualitativeness*. An observer with qualitative knowledge about the image I knows in effect one "bit" of information about each interpretation S_i, or in other words only N bits of information altogether. This is in a very literal sense much less information than knowing the precise likelihood of I. Nevertheless, as the next section shows, this very impoverished information turns out to be very useful.

Notice that the assumption of qualitativeness attributes a perfectly well-defined state of knowledge to the observer, and hence can be realized explicitly in Bayesian terms. Specifically, say that the observer knows that that the image I falls in some region $A \subseteq \mathbf{I}$. The likelihood of this state of affairs (i.e., state of knowledge) is the integral of probability over all of A, i.e.

$$p(I \text{ falls in } A|S_i) = \int_A p(I|S_i)d\mathbf{I}. \tag{7}$$

This relation, which is fundamental to all calculations in a Bayesian framework, underlies the argument in the next section.

Bayes yields a minimum rule

This section shows that, under these two assumptions:

(i) EPA-$\mathcal{S}$
(ii) qualitativeness

Bayes' rule is analytically equivalent to a minimum rule.

Consider a simple case with only two interpretations S_1 and S_2, with $S_2 \rightarrow S_1$ (that is, S_2 is strictly more restrictive than S_1; see Fig. 2a). What does Bayes tell us in this case?

By Bayes' rule, we should choose S_2 (the more restrictive interpretation) whenever

$$p(S_2)p(I|S_2) > p(S_1)p(I|S_1). \tag{8}$$

By EPA-$\mathcal{S}$, $p(S_1) = p(S_2) = 1/2$, so this inequality holds whenever

$$p(I|S_2) > p(I|S_1). \tag{9}$$

Now, following the assumption of qualitativeness, consider an observer who knows *only* that the image I falls within the support of the smaller interpretation S_2,

$$I \in \sigma(S_2) \tag{10}$$

(e.g., knowing that two line segments are parallel, but not knowing anything else about the image). The likelihood of this situation under S_2 is

$$\int_{\sigma(S_2)} p(I|S_2)d\mathbf{I}, \tag{11}$$

the integral of S_2's likelihood over the entirety of S_2's support, which is by definition equal to unity. The likelihood of this situation under S_1, on the other hand, is the integral of S_1's likelihood over the support of S_2, i.e.

$$\int_{\sigma(S_2)} p(I|S_1)d\mathbf{I}, \tag{12}$$

(the shaded area in Fig. 1b). Because by assumption $\sigma(S_2) \subsetneq \sigma(S_1)$, this quantity must be less than unity,

$$\left[\int_{\sigma(S_2)} p(I|S_1)d\mathbf{I}\right] < 1. \tag{13}$$

Hence if $I \in \sigma(S_2)$, the total likelihood of S_2 (unity) is *always* greater than the total likelihood of S_1 (Eq. 12), and S_2 *always* wins.

Conversely, assume the observer knows only that I falls *outside* the support of S_2, $I \notin \sigma(S_2)$ (e.g., knowing that two line segments are not parallel). Now by definition $p(I|S_2) = 0$, but $p(I|S_1) > 0$, so S_1 always wins.

Thus the orthodox Bayesian decision in the case $S_2 \rightarrow S_1$, assuming EPA-$\mathcal{S}$ and qualitativeness, has an extremely simple form:

If S_2 is consistent with the image, infer S_2; otherwise, infer S_1. (14)

In words: if the specialized configuration S_2 holds in the image, draw the more restrictive interpretation, because that would *explain* the image (the image would be 100% likely under that "story"); whereas under the less restrictive interpretation, the image would be just a coincidence, and thus unexplained. This is the basic logic of non-accidental properties, analogous to Rock (1983)'s "coincidence explanation principle," rendered in Bayesian language.

What about more complex interpretation spaces, with more than just two interpretations? It can be shown that in any hierarchical space (like the ones in Fig. 2a, or arbitrarily more complicated ones) this rule generalizes:

Choose the lowest interpretation on the diagram consistent with I. (15)

Formally, the word "diagram" here means a *partial order* defined over the interpretation space $\mathcal{S}$, and "the lowest in the diagram" means the formal minimum in this partial order among all interpretations with nonzero likelihood given the image.[7] That is, among all interpretations that could have produced the image, *choose the one that is most restrictive*. A proof of this generalization can be found in the Appendix.

Rule (15) has the form of a "minimum rule," and indeed in several earlier papers (Feldman, 1997c, 1997b, 1999), I have developed it as such (using non-Bayesian arguments). The theory describing the necessary partial orders and diagrams is called Minimal Model theory, and the minimum rule is referred as the *maximum-depth rule* (or sometimes the *lattice-minimum rule*; see Feldman, 1997c; Jepson & Richards, 1991), with the chosen interpretation called the *maximum-depth interpretation*, *minimal model* or *minimal interpretation*.

The notion of "simplest" captured by the maximum-depth rule contrasts with the more conventional notion of a minimum-length description in the tradition of "coding theory" (Hochberg & McAlister, 1953; Leeuwenberg, 1971; Buffart et al., 1981; see Wagemans, 1999 for a critique). The main difference is that instead of minimizing the *length* of the description, in Minimal Model theory one seeks an extremal interpretation in a connected, ranked series of interpretations; or, what turns out to be equivalent, to find the minimum in certain well-defined algebra (see Feldman, 1997b). One advantage of the resulting theory is that it becomes possible to explore the mathematical properties of the selection rule relatively directly—the arguments in the current paper being but one example. In a sense the advantage of algebraic techniques is that they do not depend on details of the coding language, but rather on structural properties of the relations among interpretations. By contrast, I know of no way to tie conventional complexity-minimization techniques (which are tied to a particular coding language) to Bayesian optimality *analytically* (rather than asymptotically).

The term "maximum-depth" reflects the use of the term *depth* (or *logical depth* or sometimes *codimension*) to denote the row number d of the given interpretation on the interpretation-space diagram (counting down from the top, with the top level denoted zero)—that is, just how far down the diagram it sits. This number plays an important role in the theory, and can be tied directly to Bayesian theory as follows.

The "winning margin" of the interpretation

A very useful measure of the probabilistic strength of an interpretation S_i, probably first suggested by Jeffreys (1939/1961), is the ratio between its likelihood and that of the empty or "null" hypothesis, denoted $\mathcal{L}_i$:

$$\mathcal{L}_i = \frac{p(I|S_i)}{p(I|S_0)}, \tag{16}$$

Here the "null" hypothesis S_0 is the weakest or most general under consideration: in our terms this means the highest "grandparent" of S_i in the interpretation space diagram. Thus the likelihood ratio gives the degree to which the target interpretation seems more compelling than a null or "random" pattern. In the case of our two-interpretation space $\{S_1, S_2\}$, the likelihood ratio of the more restrictive interpretation S_2 is just

$$\mathcal{L}_2 = \frac{p(I|S_2)}{p(I|S_1)}. \tag{17}$$

In order to do some thumbnail calculations about likelihood ratios, it is convenient to introduce the following common notational approximation. Every time two interpretations are connected by an edge in the diagram (i.e., every time one interpretation is embedded in another), assume that the lower one occupies about the same relative area within the upper one, and denote this relative area by ε:

$$\varepsilon = \int_{\sigma(S_{\text{lower}})} p(I|S_{\text{upper}}) d\mathbf{I}, \tag{18}$$

for any two interpretations S_{upper} and S_{lower} that adjoin each other on the diagram. That is, given that I falls in the support of one interpretation S_{upper}, ε is the probability that it *also* falls in the support of the more specialized interpretation S_{lower}. Thus ε is our standard value for the probability of a "coincidence."

In the above examples S_1 plays the role of S_{upper} and S_2 that of S_{lower}, so S_2 takes up about ε of S_1's total area (i.e. the integral (12) equals ε). That means that when S_1 is really true, there is a probability of about ε that the image will appear to be consistent with S_2 *anyway*.

Combining these equations, and again assuming qualitativeness and EPA-$\mathcal{S}$, we find immediately that the likelihood ratio for S_2 depends on ε thus:

$$\mathcal{L}_2 = \frac{1}{\varepsilon}; \tag{19}$$

that is, when S_2 appears possible, then the strength of the inference that it is actually the true state of the world is large $(1/\varepsilon)$ whenever ε is a small.

[7] Formally, before talking about "the" minimum in this partial order, we need to prove that it is unique; see Appendix.

What about when there are more than just two interpretations? Imagine an interpretation S that sits d steps down the diagram from the top. (Recall that the number d, the "row number" of the interpretation in the diagram, is called the *depth* of the interpretation.) By an obvious extension of the above argument, the likelihood ratio of this interpretation will be

$$\mathcal{L}_S \approx \frac{1}{\varepsilon^d} = \varepsilon^{-d}. \tag{20}$$

As we move down the diagram to increasingly complex "coincidences" (i.e., as d increases), the probability of the configuration having occurred by accident decreases exponentially, and the strength of our inference that the configuration is *not* a coincidence increases rapidly. For example, with $\varepsilon = 0.05$ (the conventional value for the probability of a "coincidence" in psychology) and $d = 2$ (the depth of *collinearity* in the diagram in Fig.), $\mathcal{L} = 0.05^{-2} = 400$, meaning that the inference of collinearity given a pair of collinear segments is 400 times stronger than the default interpretation of no structure. As interpretations get further down the diagram, they very rapidly increase in probabilistic compellingness, in an explicitly Bayesian sense.

Of course, Eq. 20 is approximate because it is only a convenient simplification to assume that each step down the diagram will occur by chance with the same probability ε. But it captures the intuition that successively more restrictive interpretations—being progressively less likely to occur by coincidence—are thus progressively *more* impressive and compelling when they do occur.

Hence assuming qualitativeness and EPA-$\mathcal{S}$, not only is Bayes' rule provably equivalent to the maximum-depth rule, but "depth" itself gives a numeric measure of the probabilistic strength of the interpretation. This statement brings into sharpest possible relief the direct analytic connection between the maximum-depth rule and Bayesian theory.

A quick recap, with an example

Let's quickly summarize the situation by means of an example. The image contains a pair of parallel line segments. Are they parallel *by accident* (S_1) or as a stable aspect of the structure of this world (S_2)? After all, even segments whose relative angle is determined randomly will occasionally happen to be parallel.[8]

If the configuration was created randomly, then its being parallel was only a coincidence that wouldn't always happen that way (Eq. 13); the probability of such a coincidence is ε (Eq. 18). On the other hand, if the configuration is truly (stably) parallel in the world, then the chances the lines would appear parallel in the image is 100% (Eq. 11). This means that the posterior on the "parallel" interpretation is 100%, which is higher than that of the "non-parallel" interpretation, which is ε (refer back to Eq. 13). Hence in this situation we conclude that the lines are truly parallel (Eq. 14).

Notice that this does not depend on the value of ε! It is purely a consequence of the fact that S_2 is embedded in S_1 (i.e., random configurations can come out parallel, but not vice versa). This is important, because in the conventional wisdom it is sometimes suggested that non-accidental inference works because the probability of an unlikely viewpoint (etc.) is nearly zero. This is wrong. Non-accidental inference works because more restrictive interpretations are stronger—even those that are *just a little* more restrictive.

All of the above will sound extremely familiar to anyone familiar with the standard story of non-accidentalness. The point here is that (a) there are a few important hidden assumptions in the standard story, (b) by adding those assumptions you can put the whole thing on a firm Bayesian footing and (c) when you generalize the story to more complicated hierarchies you get the maximum-depth rule.

Summarizing, the above argument shows that the maximum-depth rule instantiates a kind of qualitative Bayesian perceptual inference (cf. Jepson & Mann, 1999). It is not that the maximum-depth rule is vaguely or approximately equivalent to Bayes (as Chater showed equivalences between simplicity principles and Bayes always are at least); rather it is *exactly* equivalent to Bayesian reasoning using assumptions and information that is qualitative in a well-defined sense. The maximum-depth rule is literally a restatement of Bayes' rule under certain assumptions about the observer's knowledge and beliefs.

A continuum of perceptual inference rules

Taking stock, we see that the maximum-depth rule is not only consistent with Bayes' rule, but actually *is* Bayes rule when the priors are set a certain way. This leads naturally to a different way of viewing the range of possible perceptual inference principles. Rather than viewing different perceptual decision rules as representing distinct and mutually inconsistent procedures, instead view them as representing *alternative choices of priors.* Then, the potentially open-ended catalog of conceivable distinct rules materializes as a very concrete and bounded parametric space: namely, the $n-1$-dimensional[9] space of possible prior probabilities, which I will refer to as *observer space* (Fig. 4).

This representation of the range of perceptual rules is attractive in that it makes explicit that alternative rules may all be realizations of a common comprehensive procedure—i.e., a Bayesian one—but all manifesting distinct assumptions on the part of the observer, yielding different decisions. This makes very explicit the connections among alternative principles, and at the same time consummates Chater's insight that reasonable principles may all represent different sides of the same coin.

[8] Note that the non-accidental question would usually be posed in terms of three-dimensionality: is this pair of line segments truly parallel in 3D, or is the parallelness just a coincidence of viewpoint? In my view this is really just a special case of the (more basic) way the question is posed here: does the configuration have property P by accident or (as it were) "on purpose"—i.e. because of some stable causal process?

[9] Recall that the priors must sum to unity, removing one degree of freedom from n, the number of interpretations.

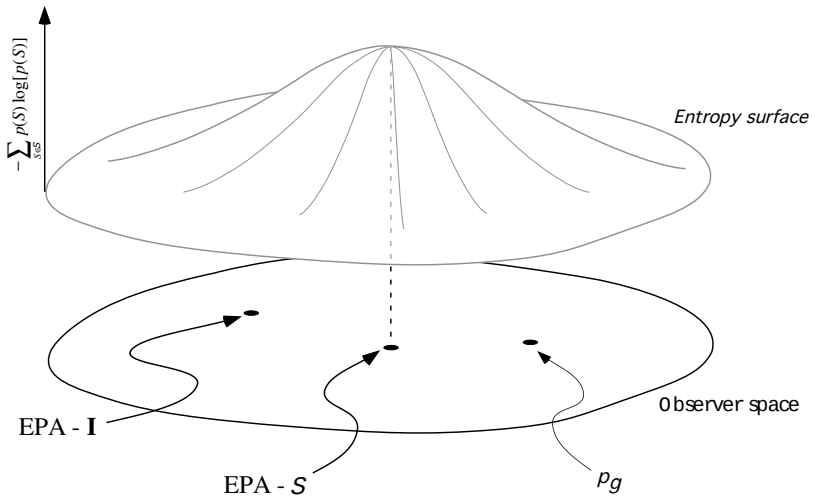

Figure 4. Observer space. Each point in this $n-1$-dimensional space represents a particular choice of prior probabilities. The figure also shows the information or entropy surface defined over this space.

Among the infinity of points in observer space, several merit special mention. One such is EPA-$\mathcal{S}$, where all the $p(S_i)$ are equal (see Fig. 4). As shown above, an observer at this point in this space—whatever its algorithmic or implementation-Al details—is in effect employing the maximum-depth rule and thus a form of minimum principle.

Another point in the space is EPA-**I**, which as discussed above (a) is not generally the same as EPA-$\mathcal{S}$ (b) is what is often meant by a "neutral" prior (c) does not correspond to any particularly simple computational rule, and (d) is not plausible as a model of human observers.

Finally, another very important point somewhere in observer space, often referred to in the literature, is the point where the priors are the *correct* ones for a given environment. I will denote this point by $p_g(S_i)$ (g for "ground truth"). While EPA-$\mathcal{S}$ represents a simplified guess about the priors, $p_g(S_i)$—located, presumably, somewhere *different* from EPA-$\mathcal{S}$—represents the actual frequencies with which the various S_i occur in the observer's world. While the priors at EPA-$\mathcal{S}$ literally cancel and drop out of Bayes rule—or, equivalently, are disregarded—at p_g the priors are felt with exactly the weight Bayes prescribes. Thus the observer located at this point is executing "true Bayes."

The information content of the observer's assumptions

Distinct points in observer space differ in the nature of the assumptions about the world they embody. In a very direct way, too, they differ in the *quantity* of information they embody: namely, in the sense of Shannon's measure H:

$$H(p) = -\sum_{i=1}^{n} p(S_i) \log[p(S_i)]. \qquad (21)$$

This equation yields a scalar quantity (expressible in bits if the logarithm is taken in base 2) for each point in observer space (see Fig. 4), which represents exactly how much information the observer's assumptions embody at each point in this space. Thus this number gives a very basic property of the observer.

The value of H as the priors are varied constitutes a smooth surface defined over observer space. It is easy to show that this "information surface" reaches a unique maximum at EPA-$\mathcal{S}$[10]. That is, the observer executing the maximum-depth rule is also maximizing its own information content, as compared with other rules, including "true" Bayes. This may sound slightly paradoxical, but only due to the ambiguity of the traditional rendering of information-theoretic terms into English. Equivalently, but perhaps more intelligibly, the *entropy* or *uncertainty* of the observer's assumptions is maximal when the EPA-$\mathcal{S}$ prior is used. Thus the use of the true priors entails *knowing more* about the environment, while the EPA-$\mathcal{S}$ and the minimum rule entail knowing less but *guessing more* about the environment.

The idea of measuring perceivers' knowledge or judgments by means of Shannon information is an old idea in psychology (Attneave, 1954), though it has actually not been exploited as much as it might (though Gilden, Hiris, & Blake, 1995 and Kubovy & Wagemans, 1995 are excellent recent examples). The idea that the observer ought to maximize this quantity in its assumptions, after taking into account whatever affirmative knowledge it possesses, is the essence of the idea of Maximum-Entropy inference, which has been very influential in probability theory, machine learning, and physics (see Skilling, 1988). Here the application of an information-theoretic measure simply gives mathematical precision to the argument that the Simplicity and Likelihood principles both represent "rational" inference but with different quantities—as well as types—of background knowledge.

[10] Indeed, it is axiomatic in the derivation of Shannon's information measure that it is maximal when all probabilities are equal (see Khinchin, 1957 for a derivation).

The tradeoff between precision and optimality

It might seem that the observer ought to aspire to be located at p_g: after all, that is by definition the *right* answer. This simple dictum brings, however, several inherent problems. One problem with the true priors is that, in any given environment, *the observer has no way of knowing what they are.* Evolution may contribute to setting them, as often speculated in Bayesian treatments, but there is no guarantee that the needed priors refer to stable categories that have existed over evolutionary time-scales. They may, in principle, change from environment to environment, from day to night, or indeed from image to image. Storing priors tuned to different situations leads quickly to a combinatorial explosion as all factors that may influence the priors need to be explicitly tabulated and fully crossed.[11]

It should be emphasized, in fact, that the use of the true priors p_g comes at a substantial computational cost. While the maximum-depth-rule observer ignores the priors, and only needs to determine one bit of information about each likelihood function (namely, whether I falls within its scope), the p_g needs to store or estimate the actual numeric value of $p_g(S_i)$ for each S_i, in principle tuned to each environment, and needs to know the details of I well enough, and the structure of each likelihood function in enough detail, to determine the precise numeric value of $p(I|S_i)$ for every S_i. By contrast, the observer using EPA-$\mathcal{S}$ needs simply to perform an easy minimization. Yet this simple procedure may give rise to little loss in the accuracy of the final decisions, because the maximum-depth rule will only conflict with the true Bayesian decision in the rare cases when the details of the priors or likelihood function priors overwhelm the embedding relations (e.g. point Q in Fig. 1).

Putting this another way, with hierarchical interpretation spaces, most of the inferential leverage comes from the structure of the hierarchy—the details about which interpretations are embedded in which others. The precise quantitative details of the priors and likelihood functions add little in the way of correctness to the final decision, but cause most of the computational difficulties. Doing away with them by means of EPA-$\mathcal{S}$ causes relatively few errors but greatly simplifies the computation.

A reasonable gloss on the situation, then, is that the observer at p_g holds the correct answer, while the one at EPA-$\mathcal{S}$ can calculate its answer the most conveniently, and end up with an answer that is not too far off in a wide range of situations, including novel environments where the priors are in fact totally unknown. Indeed, by following a reasonable and cheap strategy, it may be that the observer is minimizing some combined accuracy-and-expense loss function and thus following a truly "optimal" meta-strategy, although that is admittedly pure speculation. In the words of the medieval philosopher Fabricius, *nature does what is best.*

Conclusions

For those perceptual theorists who have fretted over the philosophical justification for the Simplicity principle, the idea of complexity minimization has seemed at times no more than a handy but totally unjustified calculating trick, whose empirical success was essentially mystifying (again see Hatfield & Epstein, 1985). Chater's (1996) argument goes a long way towards clearing up this mystery: in a very general but also somewhat abstract sense, complexity minimization serves the purpose of building a veridical representation in the Bayesian sense. But because this property is shared by any reasonable complexity measure (any one that is universal in Kolmogorov's sense), Chater's argument does not help clear up the ambiguity in specifying exactly which minimum rule the visual system uses.

The argument in the current paper shows that a certain mathematically well-defined minimum rule—the maximum-depth rule of Minimal Model theory—achieves Bayesian optimality under very reasonable assumptions. Again, the argument is analytic, not asymptotic: the maximum-depth interpretation is precisely the Bayesian interpretation assuming neutral priors over the interpretations and qualitative knowledge about the image. This provides a rationale for the use of the maximum depth rule, in that it shows mathematically why the rule tends to produce correct interpretations. Moreover, this version of the Simplicity principle and ordinary Bayesian inference can be regarded as distinct points in a well-defined continuous space of possible perceptual rules. The two principles (as well as every other point in observer space) differ in the prior probabilities they assume, and thus differ literally in the magnitude of their information content.

Finally, these arguments emphasize that different rules for perceptual inference can differ not only in the computations they specify, but in the knowledge and assumptions they embody. This idea recasts the historic debate between the Simplicity and Likelihood principles: the question is not what computational tricks the visual system uses, but rather what assumptions about the world are embedded in the rules it employs.

References

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*, 183–193.

Bennett, B. M., Hoffman, D. D., & Murthy, P. (1993). Lebesgue logic for probabilistic reasoning and some applications to perception. *Journal of Mathematical Psychology*, *37*(1), 63–103.

Biederman, I. (1987). Recognition by components: a theory of human image understanding. *Psychological Review*, *94*, 115–147.

Box, G. E. P., & Tiao, C., George. (1973). *Bayesian inference in statistical analysis.* Reading, Massachussetts: Addison-Wesley.

Buffart, H., Leeuwenberg, E. L. J., & Restle, F. (1981). Coding theory of visual pattern completion. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(2), 241-274.

[11] These are, by the way, well-known arguments in the Bayesian literature. Rather than being seen as arguments against Bayes, they are better regarded as arguments in favor of a subjectivist rather than frequentist approach to the interpretation and construction of priors. See for example Jaynes (1973).

Bülthoff, H. H., & Yuille, A. L. (1991). Bayesian models for seeing shapes and depth. *Comments on Theoretical Biology*, *2*(4), 283–314.

Caelli, T. M., & Umansky, J. (1976). Interpolation in the visual system. *Vision Research*, *16*(10), 1055–1060.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, *103*(3), 566–581.

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

Feldman, J. (1997b). Regularity-based perceptual grouping. *Computational Intelligence*, *13*(4), 582–623.

Feldman, J. (1997c). The structure of perceptual categories. *Journal of Mathematical Psychology*, *41*, 145–170.

Feldman, J. (1999). The role of objects in perceptual grouping. *Acta Psychologica*, *102*, 137–163.

Feldman, J. (2001). Bayesian contour integration. *Perception & Psychophysics*, *63*(7), 1171–1182.

Feldman, J. (2003). Perceptual grouping by selection of a logically minimal model. *International Journal of Computer Vision*, *55*(1), 5–25.

Gilden, D., Hiris, E., & Blake, R. (1995). The informational basis of motion coherence. *Psychological Science*, *6*(4), 235–240.

Hatfield, G., & Epstein, W. (1985). The status of the minimum principle in the theoretical analysis of visual perception. *Psychological Bulletin*, *97*(2), 155–186.

Hochberg, J., & McAlister, E. (1953). A quantitative approach to figural "goodness". *Journal of Experimental Psychology*, *46*, 361–364.

Jaynes, E. T. (1957/1988). How does the brain do plausible reasoning? In G. J. Erickson & C. R. Smith (Eds.), *Maximum-entropy and Bayesian methods in science and engineering* (Vol. 1, pp. 1–24). Dordrecht: Kluwer. (First appeared as a Stanford Microwave Laboratory Report, 1957.)

Jaynes, E. T. (1973). The well-posed problem. *Foundations of physics*, *3*, 477-491.

Jeffreys, H. (1939/1961). *Theory of probability (third edition).* Oxford: Clarendon Press.

Jepson, A., & Mann, R. (1999). Qualitative probabilities for image interpretation. In *Proceedings of the International Conference on Computer Vision* (Vol. II, pp. 1123–1130).

Jepson, A., & Richards, W. A. (1991). *What is a percept?* (Occasional Paper No. 43). MIT Center for Cognitive Science.

Jepson, A., & Richards, W. A. (1992). What makes a good feature? In L. Harris & M. Jenkin (Eds.), *Spatial vision in humans and robots* (pp. 89–125). Cambridge University Press.

Kanade, T. (1981). Recovery of the three-dimensional shape of an object from a single view. *Artificial Intelligence*, *17*, 409–460.

Kanizsa, G. (1979). *Organization in vision: essays on Gestalt perception.* New York: Praeger Publishers.

Khinchin, A. I. (1957). *Mathematical foundations of information theory.* New York: Dover.

Knill, D., & Richards, W. (Eds.). (1996). *Perception as Bayesian inference.* Cambridge: Cambridge University Press.

Kubovy, M., & Wagemans, J. (1995). Grouping by proximity and multistability in dot lattices: a quantitative gestalt theory. *Psychological Science*, *6*(4), 225–234.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Research*, *35*(3), 389–412.

Leeuwenberg, E. L. J. (1971). A perceptual coding language for visual and auditory patterns. *American Journal of Psychology*, *84*(3), 307–349.

Leeuwenberg, E. L. J., & Boselie, F. (1988). Against the likelihood principle in visual form perception. *Psychological Review*, *95*, 485–491.

Lowe, D. G. (1987). Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, *31*, 355–395.

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

Quine, W. (1965). On simple theories of a complex world. In M. H. Foster & M. L. Martin (Eds.), *Probability, confirmation, and simplicity: Readings in the philosophy of inductive logic* (pp. 250–252). New York: Odyssey Press.

Richards, W. A., Jepson, A., & Feldman, J. (1996). Priors, preferences, and categorical percepts. In D. Knill & W. Richards (Eds.), *Perception as Bayesian inference* (pp. 93–122). Cambridge: Cambridge University Press.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry.* Singapore: World Scientific.

Rock, I. (1983). *The logic of perception.* Cambridge: MIT Press.

Schöning, U., & Pruim, R. (1998). *Gems of theoretical computer science.* Berlin: Springer.

Skilling, J. (Ed.). (1988). *Maximum entropy and Bayesian methods.* Dordrecht: Kluwer.

Smits, J. T. S., & Vos, P. G. (1986). A model for the perception of curves in dot figures: the role of local salience of "virtual lines". *Biological Cybernetics*, *16*, 407–416.

Sober, E. (1975). *Simplicity.* London: Oxford University Press.

van der Helm, P. (2000). Simplicity versus likelihood in visual perception: From surprisals to precisals. *Psychological Bulletin*, *126*(5), 770–800.

Wagemans, J. (1992). Perceptual use of non-accidental properties. *Canadian Journal of Psychology*, *46*(2), 236–279.

Wagemans, J. (1993). Skewed symmetry: a nonaccidental property used to perceive visual forms. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(2), 364–380.

Wagemans, J. (1999). Toward a better approach to goodness: Comments on Van der helm and Leeuwenberg (1996). *Psychological Review*, *106*(3), 610–621.

Witkin, A. P., & Tenenbaum, J. M. (1983). On the role of structure in vision. In J. Beck, B. Hope, & A. Rosenfeld (Eds.), *Human and machine vision* (pp. 481–543). Academic Press.

Appendix: Generalization of the minimum rule (maximum-depth rule) to arbitrary n

An interpretation space $\mathcal{S} = \{S_1 \ldots S_n\}$ is said to be *hierarchical* if for each $S_1, S_2 \in \mathcal{S}$, one of the following holds:

(a) $\sigma(S_1) \cap \sigma(S_2) = \emptyset$ (S_1's and S_2's supports are disjoint)
(b) $S_1 \rightarrow S_2$,
(c) $S_2 \rightarrow S_1$.

That is, in every pair of interpretations, either one contains the other the two do not overlap at all. This means that every

hierarchical interpretation space is partially ordered by support inclusion. Note that this trivially includes "flat" spaces with all interpretations disjoint.

We seek to prove that for an arbitrary hierarchical interpretation space $\mathcal{S}$ with likelihood functions $p(I|S_i)$, and assuming EPA-$\mathcal{S}$ priors $p(S_i) = 1/n$, then given any image $I \in \mathbf{I}$, the minimal interpretation $S_{\min}$ of I is

(i) unique, and
(ii) the maximum a posteriori interpretation of I.

Proof.

Assume in what follows that all I are contained in the union of the supports of all interpretations; that is, all I under discussion are have non-zero likelihood in at least one S_i.

(i) *Uniqueness.* Given I, define a *minimal interpretation* $S_{\min} \in \mathcal{S}$ as an interpretation such that for every S_i with $I \in \sigma(S_i)$ (i.e. that could have produced I), either $S_{\min} = S_i$ or $S_{\min} \rightarrow S_i$. ($S_{\min}$ either *is* or *implies* every interpretation that might have produced I.) We seek to prove that $S_{\min}$ is unique.

By induction. Assume some hierarchical interpretation space $\mathcal{S}$ such that I has a unique minimal interpretation $S_{\min}$. An example of such a space is the one-element space $\{S\}$, in which case $S = S_{\min}$ is clearly unique.

Now add a new interpretation S_{new}, such that the new interpretation space $\mathcal{S} \cup \{S_{\text{new}}\}$ is also hierarchical. Now, if $I \notin \sigma(S_{\text{new}})$ (i.e., $p(I|S_{\text{new}}) = 0$), then $S_{\min}$ is still the unique minimal interpretation in the new interpretation space $\mathcal{S} \cup \{S_{\text{new}}\}$. Conversely, assume $I \in \sigma(S_{\text{new}})$ (i.e., $p(I|S_{\text{new}}) > 0$). First, $\sigma(S_{\min})$ and $\sigma(S_{\text{new}})$ cannot be disjoint because they both contain I. Hence either $S_{\text{new}} \rightarrow S_{\min}$, in which case S_{new} is the new minimum and is unique, or else $S_{\min} \rightarrow S_{\text{new}}$, in which case $S_{\min}$ is still the minimum and still unique.

All hierarchical interpretation spaces can be built up from the one-element space by the above induction. Hence all hierarchical interpretation spaces have unique minimal interpretations for all I, completing the proof of uniqueness.

(ii) *Bayesian correctness.* We seek to show that given I and $\mathcal{S}$ the unique minimal interpretation $S_{\min}$ is the maximum a posteriori interpretation, i.e., maximizes $p(S)p(I|S)$, assuming EPA-$\mathcal{S}$ and qualitativeness as discussed in the text.

Again we proceed by induction from a space $\mathcal{S}$ with the desired property, such as the one-element space $\{S\}$. Denote the maximum a posterior interpretation in $\mathcal{S}$ by S_{MAP}, and again denote the new interpretation by S_{new}. If $I \notin \sigma(S_{\text{new}})$ then clearly S_{MAP} continues to be the best interpretation. Conversely, assume if $I \in \sigma(S_{\text{new}})$. First, $\sigma(S_{\text{MAP}})$ and $\sigma(S_{\text{new}})$ cannot be disjoint because they both contain I. Hence either $S_{\text{new}} \rightarrow S_{\text{MAP}}$, in which case by the argument sketched in the text (Eqs. 11, 12 and ff) S_{new} has higher posterior than S_{MAP} and becomes the new best interpretation; or else $S_{\text{MAP}} \rightarrow S_{\text{new}}$, in which case S_{MAP} continues to be the best interpretation.