

Bayes and the Simplicity Principle in Perception

Jacob Feldman
Rutgers University—New Brunswick

Discussions of the foundations of perceptual inference have often centered on 2 governing principles, the likelihood principle and the simplicity principle. Historically, these principles have usually been seen as opposed, but contemporary statistical (e.g., Bayesian) theory tends to see them as consistent, because for a variety of reasons simpler models (i.e., those with fewer dimensions or free parameters) make better predictors than more complex ones. In perception, many interpretation spaces are naturally hierarchical, meaning that they consist of a set of mutually embedded model classes of various levels of complexity, including simpler (lower dimensional) classes that are special cases of more complex ones. This article shows how such spaces can be regarded as algebraic structures, for example, as partial orders or lattices, with interpretations ordered in terms of dimensionality. The natural inference rule in such a space is a kind of simplicity rule: Among all interpretations qualitatively consistent with the image, draw the one that is lowest in the partial order, called the maximum-depth interpretation. This interpretation also maximizes the Bayesian posterior under certain simplifying assumptions, consistent with a unification of simplicity and likelihood principles. Moreover, the algebraic approach brings out the compositional structure inherent in such spaces, showing how perceptual interpretations are composed from a lexicon of primitive perceptual descriptors.

Keywords: perceptual organization, perceptual grouping, likelihood, Bayes, simplicity

Supplemental materials: <http://dx.doi.org/10.1037/a0017144.supp>

Simplicity Versus 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 (Buffart, Leeuwenberg, & Restle, 1981; Hochberg & McAlister, 1953; Leeuwenberg, 1971; Van der Helm & Leeuwenberg, 1996). Many phenomena of visual perception seem to rest in whole or in part by a preference for simple interpretations (Chater, 2005; Chater & Vitányi, 2003; Pomerantz & Kubovy, 1986). 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).

This research was supported in part by National Science Foundation Grants SBR-9875175 and SBR-0339062 and by National Institutes of Health (National Eye Institute) Grant EY15888. I am grateful to Nick Chater, Julian Hochberg, Rajesh Kasturirangan, Michael Kubovy, Whitman Richards, Manish Singh, Matthew Stone, Josh Tenenbaum, and Johan Wagemans for helpful discussions and comments.

Correspondence concerning this article should be addressed to Jacob Feldman, Department of Psychology & Center for Cognitive Science, Rutgers University—New Brunswick, Busch Campus, Piscataway, NJ 08854. E-mail: jacob@ruccs.rutgers.edu

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 an organism to achieve veridical percepts of the world (see Geisler & Diehl, 2002). 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 as roughly incompatible (Hatfield & Epstein, 1985; Leeuwenberg & Boselie, 1988; Perkins, 1976; Van der Helm, 2000). More recently however, Chater (1996), using mathematical arguments paralleling those from minimum description length theory (Li & Vitányi, 1997; 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 correct in an objective sense. This remarkable demonstration combines the most appealing aspects of both principles, giving the minimum principle a clear rationale (namely, veridicality) while suggesting a criterion by which the most likely interpretation can be identified (namely, minimum description length). More broadly, Bayesian theory has come to be generally understood to involve a built-in bias toward simpler (e.g., lower dimensional) models, sometimes referred to as the Bayesian Occam factor (e.g., see Duda, Hart, & Stork, 2001; MacKay, 2003, and discussion below).

Nevertheless, the relationship between complexity and Bayesian optimality is less than completely satisfying because it is essentially *asymptotic* in nature, reflecting the behavior of ideal perceptual codes in the limit. The universality of the connection stems from 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 does not 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, one never knows its specific value—the length of the actual shortest program is uncomputable in general (see Schöning & Pruim, 1998, for an elegant proof). 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 the 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 can 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 (1996) argument, while persuasive in an abstract sense, leaves open the narrower (but crucial) question of the nature of the actual coding language used by the visual system, and thus the exact form of the associated minimization rule. The current article seeks to demonstrate a stronger and more specific connection between a certain type of minimization rule and Bayesian theory, in a way that also makes more explicit the nature and meaning of the associated coding language. The *maximum-depth* or *lattice minimum rule* has been developed previously (Feldman, 1997b, 1997c, 1999, 2003a), but in nonprobabilistic terms, and its intimate connection to Bayesian theory has not previously been developed. The conditions invoked in this rule are less general than in Chater's formulation but are important in a 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 connection between the minimum principles and Bayesian inference described below sheds new light on why perceptually realistic simplicity minimization tends to identify the state of the world correctly.

Bayesian Formulation

Bayesian theory is a particularly attractive formulation of the likelihood principle in perception (see Bühlhoff & Yuille, 1991; Kersten, Mamassian, & Yuille, 2004; Knill & Richards, 1996), in that it provides provably optimal inferences under conditions of uncertainty (see Jaynes, 2003). Thus Bayes provides optimal solutions to the ambiguities inherent in perception. In Bayesian theory, the subjective belief in a particular hypothesis given particular data is associated with the *posterior probability*, that is, the conditional probability of the hypothesis given the data. In the context of visual perception, the data are 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 set $\mathbf{S} = \{S_1, S_2, \dots\}$ of distinct scene models, that is, categories of distal scenes. Each interpreta-

tion 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} ($\int_{\mathbf{I}} p(I|S_i) d\mathbf{I} = 1$).²

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

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

Because the denominator is the same for all interpretations, the winning interpretation will be the one that maximizes the numerator $p(S_i)p(I|S_i)$, the product of the prior and the likelihood. The central principle of Bayesian theory is that the observer's degree of belief should be proportional to the posterior, and thus to this product. Another idea important in what follows is the *support* $\sigma(S_i)$ of an interpretation S_i , defined as the region of \mathbf{I} where S_i 's likelihood is nonzero,

$$\sigma(S_i) = \{I \in \mathbf{I} | p(I|S_i) > 0\}. \quad (2)$$

(Note that if we chose to be less strict we could set the criterion to some $\delta > 0$ instead of 0.) The support of a scene model S_i is the set of images that it could have produced, that is, the set of images that are qualitatively consistent with it.

Hierarchies of Candidate Interpretations

Bayesian theory has often been criticized for requiring the calculation of the posterior for a large (and indeed potentially infinite) number of distinct hypotheses, and thus a great deal of computation. Much of the Bayesian literature is devoted to finding well-motivated approximations to the optimal Bayesian solution, but even so many approaches require computations that may seem unrealistic for biological computation.

In some perceptual situations, though, preexisting formal relations among the S_i may make such an elaborate computation unnecessary. Many perceptual interpretation spaces are hierarchical, meaning that certain interpretations are special cases of others. An example comes from the well-known recognition-by-components (RBC) theory of Biederman (1987) in which many of the basic part types, called *geons*, are special cases of each other (see Kurbat, 1994). For example, straight bricks are special cases of curved bricks (in which curvature is zero) and also of tapered bricks (in which taper is zero), a relation that may even be reflected in neurophysiological tuning curves for volumetric primitives (De Baene, Ons, Wagemans, & Vogels, 2008; De Baene, Premereur, & Vogels, 2007). Figure 1 diagrams a partial set of geons in terms of

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 complex, the constant difference between coding lengths for two machines can be arbitrarily large.

² Note though that the likelihood function generally does not sum to unity over models \mathbf{S} .

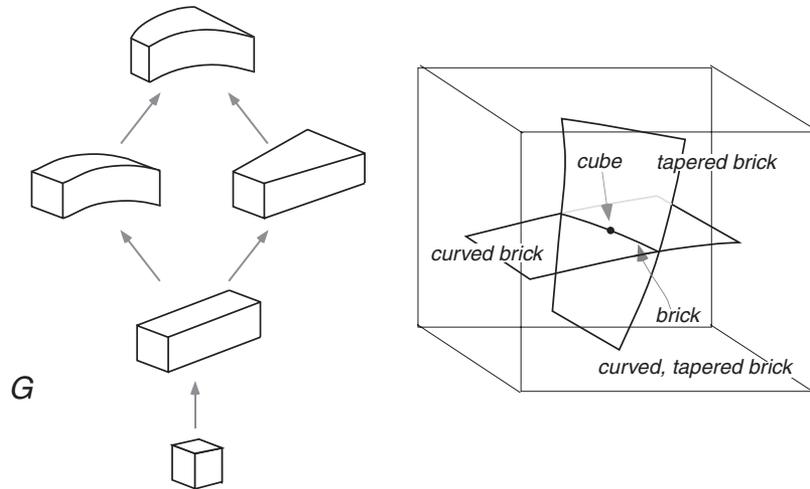


Figure 1. A set of geons (Biederman, 1987), depicted (left) as a hierarchical interpretation space (lattice) G (note isomorphism to Figure 3b), and (right) as a set of intersecting manifolds in 3-space. Curved, tapered bricks form a three-parameter volume, in which curved bricks and tapered bricks are each two-parameter submanifolds. Their intersection (bricks) is a one-parameter curve (the dimension being the brick's length). Cubes are a point on this curve. A color version of this figure is available on the Web at <http://dx.doi.org/10.1037/a0017144.supp>.

the subset relations among the types. In the diagram, nodes lower in the figure represent subsets of nodes higher in the figure. Each step down the diagram removes one parameter or degree of freedom from the model class, and each step up adds one parameter. The classes thus form a set of intersecting spaces or manifolds of various dimensions, each corresponding to a family of scene models (illustrated schematically at the right in the figure).

Another important example of a hierarchical interpretation space in the perceptual organization literature is the set of nonaccidental relations (Lowe, 1987; Witkin & Tenenbaum, 1983; see Freeman, 1994). Nonaccidental relations are geometric relations between image elements that are unlikely to occur by accident, and which consequently are perceptually salient (Feldman, 2007; Kukkonen, Foster, Wood, Wagemans, & van Gool, 1996; Wagemans, 1992; Wagemans, van Gool, Lamote, & Foster, 2000). Examples include parallelism, collinearity (Caelli & Umansky, 1976; Claeysens & Wagemans, 2005; Feldman, 1997a; Feldman & Singh, 2005; Smits & Vos, 1986), and skew symmetry (Kanade, 1981; Wagemans, 1993). A lattice of nonaccidental relations is shown in Figure 2. As with the geon lattice, cases lower on the lattice are special cases of, and hence embedded in, cases higher on the lattice: For example the set of line segment pairs that are parallel is a subset of the set of all line segment pairs.

The relation depicted in these diagrams is a *partial order*, a relation in which some (but not necessarily all) of the elements are ranked. Partial orders can have a variety of different structures, several of which are illustrated in Figure 3. If S_2 precedes (or is equal to) S_1 in the partial order (notated $S_2 \leq S_1$) then S_2 is ad descendant of (or is equal to) S_1 , that is, hangs somewhere down the chain from S_1 in the diagram (or they may be equal). If S_2 is the immediate child of S_1 then we write $S_2 \rightarrow S_1$; these arrows correspond directly to the arrows depicted in the diagram. Notice that some model classes are not special cases of each other: Neither is the other's descendant in the diagram, like tapered bricks and curved bricks in Figure 1. This particular partial order

is also a *lattice*, a partial order in which every two nodes S_1 and S_2 have both a greatest common child (called their *meet* and denoted $S_1 \wedge S_2$) and a least common ancestor (called their *join* and denoted $S_1 \vee S_2$). (See Davey & Priestley, 1990, for an introduction to orders and lattices.) The main goal of the current article is to explore how these algebraic structures and relations relate to Bayesian inference.

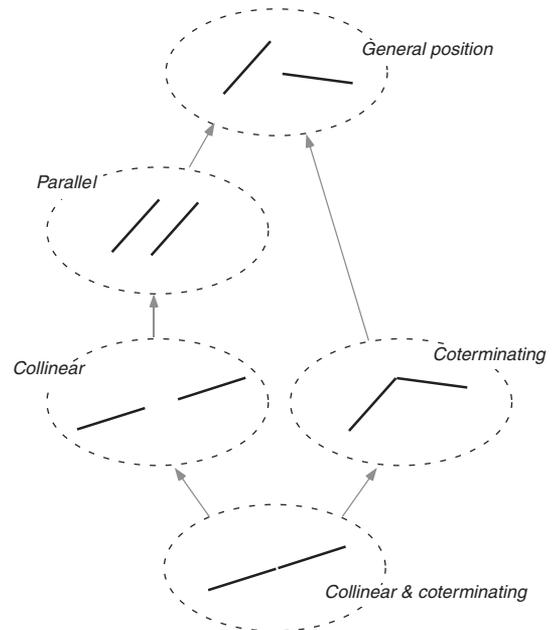


Figure 2. An example of a hierarchical interpretation space from the perceptual grouping literature: spatial relations between two line segments. See Feldman (2007) for discussion.

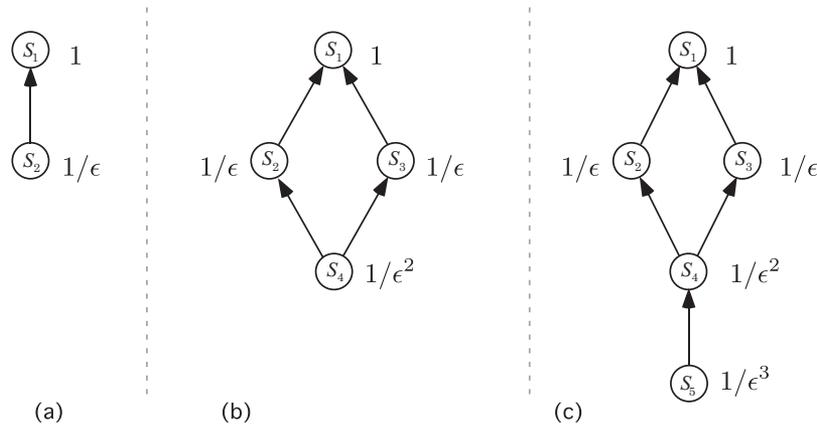


Figure 3. Examples of hierarchical interpretation space diagrams. (a) $S_2 \rightarrow S_1$; (b) $S_2 \rightarrow S_1$, $S_3 \rightarrow S_1$, $S_4 \rightarrow S_2$, $S_4 \rightarrow S_3$; (c) $S_2 \rightarrow S_1$, $S_3 \rightarrow S_1$, $S_4 \rightarrow S_2$, $S_4 \rightarrow S_3$, $S_5 \rightarrow S_4$. Note isomorphism between Panel (c) and the lattice in Figure 1. Adjacent to each interpretation is its likelihood ratio (see Equation 17 and surrounding text).

Model classes that are subsets of other model classes pose a special problem for any theory of inference, because the larger model always fits better than (or at least as well as) the smaller. This means that in the absence of any mitigating factor, the broadest model will always be chosen. This is undesirable, because it leads to poor generalization (sometimes called *overfitting*; see Hastie, Tibshirani, & Friedman, 2001, for extensive discussion), and is clearly a poor model for what the perceptual system actually does. A common solution is to penalize each model as a function of how many fittable parameters it contains, for example, by the Akaike information criterion (AIC; Akaike, 1974), which imposes a penalty proportional to the number of parameters, or the Bayesian information criterion (BIC), which imposes a somewhat larger penalty.

The AIC (and by extension, the BIC, which shares critical properties) has been criticized for both its motivation, which some feel is ad hoc, and its performance, which is inconsistent (Dowe, Gardner, & Oppy, 2007). And indeed Bayesians have argued that Bayesian theory by itself handles variations in the dimension of model families in a more natural way. Tenenbaum and colleagues have extensively explored the link between Bayesian inference, Occam's razor, and the size of a model's extension, in the context of more cognitive tasks related to inductive generalization (Tenenbaum & Griffiths, 2001; Tenenbaum, Griffiths, & Kemp, 2006). MacKay (2003) has argued that Bayesian theory automatically favors models with fewer parameters over those with more (such as those in which they are embedded as special cases) because the latter necessarily spread their probability mass over a larger region of data space. MacKay's central example of two model families, one embedded in the other, is essentially the same as our S_1 and S_2 with $S_2 \rightarrow S_1$ (see Figure 3a). The main aim here is to extrapolate this point beyond a pair of models, one embedded in the other, to an entire hierarchical family of mutually embedded models with a more complex set of relationships. Such hierarchical families are naturally understood as algebraic structures, with logic-like operators and inference rules detailed below. The goal is to connect this algebraic approach—with its attractive combinatoric structures and compositional semantics—to Bayesian inference (similar to

the approach in the rational rules model of Goodman, Tenenbaum, Feldman, & Griffiths, 2008).

As is often remarked, while Bayesian theory provides an optimal method for selecting among hypotheses it does not provide the hypothesis space itself, which must be chosen based on substantive considerations drawn from the domain in question. In perception, the choice of hypothesis space reflects the perceiver's (i.e., the brain's) tacit assumptions about what classes of events (scene models, hypotheses) tend to occur in the world—geons, spatial relations, object classes, colors, sounds, and so forth. So Bayesian accounts of perception can be tailored to the peculiar qualities of the model spaces that arise in a perceptual context.

Perception also has several peculiar demands that distinguish it from other inference problems. One is the need for very rapid and possibly pared-down computational procedures conducive to real-time updating. Another is the need for a single unitary result, the *percept*. As was first noted by the Gestalt psychologists (often in connection with bistability), perception usually converges on a single final conclusion that corresponds to the current estimate of reality (see Kanizsa, 1979, for discussion of this point). Bayesian procedures do not in general yield a single conclusion, but rather yield a full posterior distribution which assigns a degree of belief to every hypothesis in the space. When a single conclusion is required, the maximum a posteriori (MAP) interpretation is often used, though Bayesians tend to regard this as a poor substitute for the full posterior distribution.

These characteristic aims of perception—the need for rapid simplified computation, and the desire to achieve a single, unified percept—suggest the need for a simple computational procedure that might take more direct advantage of the structure inherent in hierarchical perceptual hypotheses.

The Qualitative Stance

A simple way to simultaneously satisfy these aims comes from treating the situation *qualitatively*. A qualitative interpretation of the evidence is one that ignores quantitative details of both the prior and the likelihood and instead assigns them categorically,

leading to a qualitative selection of winning model. In the now-extensive Bayesian perception literature, most analysis is oriented toward quantitative estimation of scene parameters. But an enormous literature in perceptual organization points toward the primacy of qualitative models: categorically distinct ways of organizing the image into contours (e.g., Kubovy, 1994; Kubovy, Holcombe, & Wagemans, 1998; Kubovy & Wagemans, 1995), surfaces (e.g., Gilchrist & Jacobsen, 1989), objects (e.g., Feldman, 2003b), and so forth. Thinking qualitatively, we ask which models are *consistent* with the image, rather than how well each model (quantitatively) fits the image, usually a more ambitious computation. This is actually more or less how we speak informally about certain image properties, for example, nonaccidental relations (see Kukkonen et al., 1996; Wagemans et al., 2000). When we consider the perceptual significance of observing, say, parallel lines, we are really asking what it would mean if we knew only that the lines were parallel, but did not know (or did not care about) the precise value of the angle between them. Without knowing this angle, we cannot evaluate the quantitative likelihood of any particular model, but we can make a qualitative inference about the likelihood of models that do and do not predict parallel lines (namely, that this observation supports the models that do). This is the “qualitative stance” (cf. Jepson & Mann, 1999).

Likelihood

For the likelihood, this means assigning likelihood based only on which interpretations’ support I lies within—that is, which models it satisfies, ignoring exactly how well it satisfies them. In general, we define qualitative likelihood as follows. Assume an image I that falls within some region A of image space, and assume that is all we know (or care) about it. The total likelihood of S on this data ($I \in A$) is simply the integral of the likelihood $p(I|S)$ over the whole of A , that is,

$$p(I \in A|S) = \int_A p(I|S)d\mathbf{I}, \tag{3}$$

which I will refer to as the *qualitative likelihood* of S . (Note that is really just the ordinary likelihood of S under data $I \in A$; it is only “qualitative” in the sense that we are treating the data I categorically by classifying it as “in A ” and disregarding further quantitative distinctions.) More specifically, say we observe an image I that falls within the support of a particular interpretation S_j . This qualitative event ($I \in \sigma(S_j)$) has a probability under each possible interpretation S_j , which allows us to evaluate each of these interpretations in light of the (qualitative) observation. Substituting $\sigma(S_j)$ in for A , the qualitative likelihood of the model S_j is

$$p(I \in \sigma(S_j)|S_j) = \int_{\sigma(S_j)} p(I|S_j)d\mathbf{I}, \tag{4}$$

that is, the integrated probability of all images consistent with S_j if S_j is really true. If S_i and S_j are the same ($i = j$) then this is simply unity,

$$p(I \in \sigma(S_i)|S_i) = \int_A p(I|S_i)d\mathbf{I} = 1, \tag{5}$$

meaning simply that if S_i is the true state of the world, then the image will be consistent with S_i with probability 1. If S_i and S_j are inconsistent (have disjoint support; $\sigma(S_i) \cap \sigma(S_j) = \emptyset$), then this likelihood will be zero, meaning simply that if S_i is true then S_j cannot happen.

The more interesting case is when I is consistent with both S_i and S_j . In a hierarchical interpretation space where some interpretations are special cases of others, this happens often, because any interpretation that is consistent with one interpretation is also consistent with those above it in the diagram. For example, a shape in the class *brick* is also in the class *curved brick*, though its curvature happens to be zero.

Assume two interpretations S_1 and S_2 with $S_2 \rightarrow S_1$ (recall this means that S_2 is S_1 ’s immediate child in the partial order). The qualitative likelihood of S_2 conditional on S_1 is simply the integral of S_1 ’s likelihood over the support of S_2 , which (continuing in the spirit of qualitiveness) we assume has some standard value ϵ ,

$$\epsilon = p(I \in \sigma(S_2)|S_1). \tag{6}$$

Thus ϵ is our standard value for the probability of a “coincidence” or “accidental” configuration: that a curved brick will be straight enough “by accident” to be classified as regular brick, or that two line segments generated at random orientations will happen to be approximately parallel. ϵ represents the probability of stepping one step down the partial order, that is, one step toward a more special case. If we further assume that successive steps down the partial order are independent (another common simplifying assumption; see Landy, Maloney, Johnston, & Young, 1995) then the ϵ s multiply, so the probability of d such steps will be ϵ^d . The number d is called the *depth* of interpretation S_j relative to S_i . Note that this does not mean that we assume that special image configurations generally occur independently; it means that when they occur by accident, the accidents are independent. When they occur as a normal or generic consequence of the true scene model (like the straight axis and parallel sides of a brick geon, which are generic in that model), the regularities are obviously highly nonindependent.

The exact value ϵ will depend on how strictly we have set the criterion for satisfaction of a special case. If say the threshold for parallelness is set to 1° , then ϵ will be smaller than if the criterion is set to 5° . But the exact value does not affect the logic governing how interpretations are assigned.

In summary, with several simple assumptions, the qualitative likelihood of an image I being consistent with interpretation S_i if interpretation S_j is correct is just

$$p(I \in \sigma(S_i)|S_j) = \epsilon^d, \tag{7}$$

where d is depth of S_i relative to S_j , that is, the number of steps down the diagram from S_j to S_i . As a special case, when $S_i = S_j$, then $d = 0$ and the likelihood is 1. Again, this means simply that each model S *generically* produces images that are consistent with it, whereas models above it in the partial order do so only with probability that is ϵ or some higher power of ϵ .

Prior

For the prior, a simple qualitative assumption is to set all priors equal, $p(S_i) = p(S_j)$ for all i, j . In the perceptual literature a great

deal of emphasis has been put on the idea of “neutral” prior probabilities, that is, assumptions that entail the least possible commitment on the part of the perceiver. In the Bayesian literature such priors are often referred to as *uninformative*, meaning that they tend to “stay quiet” while allowing the data to speak for themselves (via the likelihood). Setting all priors equal is a very simple way of achieving quiet priors, because when all priors are equal they cancel out of all comparisons and only the likelihood matters.

Still, as Bayesians emphasize, all inference requires assumptions, and with a hierarchical interpretation space an assumption of equal priors is by no means trivial nor without substantive consequence. Bayesian sometimes argue (e.g., see Robert, 2007) that the prior probability of any continuous parameter being exactly zero (or any other particular value) is always zero, or at least tends toward zero as our measurement resolution improves. But special cases in a hierarchical interpretation space by definition involve some parameter taking a value of zero that is *not* zero in the more generic (nonspecial) case above it in the partial order—like curvature in the case of straight bricks versus curved bricks. Hence if $S_2 \rightarrow S_1$, assuming $p(S_1) = p(S_2)$ means assuming that the parameter being zero is *just as likely* as its being nonzero. This means in effect that we are “squeezing” an equal amount of probability mass into each interpretation, regardless of its intrinsic size in the underlying image space or its relation to other interpretations. In the case of nonaccidental properties, it squeezes an equal amount of probability mass into an some areas that are *infinitely* smaller than others—or, as we have assumed above, are ϵ the size of others. This results in a highly nonuniform distribution of probability mass over the image space.

This makes sense only if we assume that the special cases to which we assign equal priors are all stable, recurring classes in the environment. Elevating a set of models for this kind of special treatment thus really amounts to adopting a basic “alphabet” of event classes in the world under observation. This point is critical and is developed further below.

Bayes Yields a Minimum Rule

Given the several qualitative assumptions above, Bayes’ rule turns out to be equivalent to a simple algebraic rule defined over the partial order. Assume two models S_1 and S_2 , with $S_2 \rightarrow S_1$ (i.e., S_2 is a special case of S_1). Qualitatively, there are two possible observations: $I \in \sigma(S_2)$ and $I \notin \sigma(S_2)$.

If $I \in \sigma(S_2)$, then the qualitative likelihood of S_1 is ϵ ,

$$p(I \in \sigma(S_2)|S_1) = \epsilon, \tag{8}$$

while the likelihood of S_2 is unity

$$p(I \in \sigma(S_2)|S_2) = 1. \tag{9}$$

Assuming equal priors, the posteriors for S_1 and S_2 are, respectively, $\epsilon/(1 + \epsilon)$ and $1/(1 + \epsilon)$. So Bayes’ rule says to pick the special case (S_2). In fact, this preference is robust against substantial deviation from equal priors, because the posterior preference for S_2 will be maintained as long as the prior ratio $p(S_1)/p(S_2)$ is less than $1/\epsilon$.

With the other possible observation, $I \notin \sigma(S_2)$, the qualitative likelihoods are

$$p(I \notin \sigma(S_2)|S_1) = 1 - \epsilon \tag{10}$$

and

$$p(I \notin \sigma(S_2)|S_2) = 0, \tag{11}$$

in the latter case meaning that the observation $I \notin \sigma(S_2)$ is inconsistent with S_2 . The posteriors for S_1 and S_2 are now, respectively, 1 and 0, so Bayes’ rule favors the nonspecial (upper) case (S_1)—regardless of the priors.

Thus the Bayesian decision in the case $S_2 \rightarrow S_1$, given the qualitative stance as outlined above, has an extremely simple form:

If S_2 is consistent with the image, choose S_2 ;
 otherwise, choose S_1 . $\tag{12}$

In other 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 a mere coincidence, and thus unexplained (cf. Griffiths & Tenenbaum, 2007). This is the basic logic of nonaccidental properties, and of Rock’s (1983) *coincidence explanation principle*, rendered in Bayesian language.

What about interpretation spaces with more than just two interpretations? It is easy to generalize the inference rule by using the structure of the partial order. Notice that the preference for children over their parents is transitive, so grandchildren (etc.) are even more favored. If $S_2 \rightarrow S_1$ and $S_3 \rightarrow S_2$, then if the image is consistent with S_3 , it is also consistent with S_2 and S_1 , but the likelihood (and posterior) for S_3 are the highest. (Posteriors for S_1 , S_2 , and S_3 are, respectively, $\epsilon^2/(1 + \epsilon + \epsilon^2)$, $\epsilon/(1 + \epsilon + \epsilon^2)$, and $1/(1 + \epsilon + \epsilon^2)$, with the last (S_3 ’s) being the highest. So generally lower nodes trump higher nodes; if $S_2 \leq S_1$ then S_2 wins. Conversely, if the image I is consistent with two interpretations S_1 and S_2 but neither $S_1 \leq S_2$ nor $S_2 \leq S_1$, then their meet $S_1 \wedge S_2$ will exist; and because $S_1 \wedge S_2 \leq S_1$ (and also $S_1 \wedge S_2 \leq S_2$), the meet $S_1 \wedge S_2$ wins. (If the meet does not exist then I cannot be consistent with both interpretations.) From there it is straightforward that the overall winner will be the interpretation that is the meet of all the interpretations consistent with I . In other words, the general rule is as follows:

Choose the lowest interpretation in the partial order
 consistent with I . $\tag{13}$

That is, among all interpretations that could have produced the image, choose the one that is most restrictive. In more formal notation, the interpretation rule is as follows. Given image I , define the set $S_I \subset S$ as the set of interpretations that are consistent with I , that is, for which $I \in \sigma(S)$. Then the interpretation rule is

$$\text{Choose interpretation } \wedge S_I. \tag{14}$$

Because the meet operator \wedge defines a formal minimization, Rule 14 is a kind of “minimum rule,” and indeed several earlier articles (Feldman, 1997b, 2003a, 2003b) have developed it as such (using nonprobabilistic arguments). The theory developing the necessary partial orders and their diagrams is called minimal model theory, and the minimum rule (Equation 14) is referred to as

the *maximum-depth rule* (or the *lattice-minimum rule*; see Feldman, 1997c; Jepson & Richards, 1991), with the chosen interpretation referred to as the *maximum-depth interpretation*, *minimal model*, or *minimal interpretation*. As mentioned above, the term *depth* in the phrase *maximum depth* refers to the number of steps down the partial order, here from the top (i.e., the row number; sometimes called the *codimension*); this is what we are maximizing by choosing the lowest interpretation. This number plays an important role in the theory, explained below.

The notion of simplicity captured by the maximum-depth rule is somewhat different in concept from the traditional notion of a minimum-length description in the tradition of “coding theory” (Hochberg & McAlister, 1953) and its more recent variants (Buffart et al., 1981; Leeuwenberg, 1971; see Wagemans, 1999, for a critique). In coding theory (as in the notions of complexity used by Chater, 1996, beginning with the approach initiated by Kolmogorov), one minimizes the *length* of the description as expressed in some fixed language. But in minimal model theory one seeks an extremal interpretation in a connected, ranked series of interpretations—or, what turns out to be equivalent, finds the minimum in certain well-defined algebras (see Feldman, 1997b). The emphasis in the minimal model approach is thus on the structural relations among the interpretations, and specifically on their inclusion relations, rather than on the length of the description or any other numeric quantity. But in effect there is a quantity minimized in minimal model theory, and it can easily be thought of as the length of a description in a certain language. Specifically, one is minimizing the number of steps on the lattice up from the bottom. This number measures the number of transformations required to generate the observed configuration from some structureless reference object corresponding to the bottom rung (Feldman, 1997c; Leyton, 1992, and see below for discussion). Thus applying the maximum-depth rule means selecting the simplest way of generating the qualitative case observed—that is, the way that requires the minimum number of generative operations. So the maximum-depth rule yields the interpretation that is both the simplest (requires the fewest generative operations) and the most likely to be correct (has maximum posterior).

The Weight of Evidence for the Winning Interpretation

We can extend the argument a bit more to quantify the strength of the qualitative evidence in favor of each interpretation. A conventional 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 an empty or “null” hypothesis, denoted L_i :

$$L_i = \frac{p(I|S_i)}{p(I|S_0)} \tag{15}$$

This ratio (or its logarithm, sometimes referred as the *weight of evidence*) quantifies the compellingness of the interpretation relative to a null baseline of “no pattern,” meaning the weakest or most general hypothesis under consideration. Griffiths and Tenenbaum (2007) showed how a similar likelihood-ratio measure captures people’s explicit judgments about the strength of a coincidence in a range of different cognitive contexts. In the minimal model approach the “null” hypothesis S_0 corresponds to S_i ’s highest

ancestor in the partial order, that is, the top node of the corresponding diagram. In the case of our two-interpretation space $\{S_1, S_2\}$, the likelihood ratio of the more restrictive interpretation S_2 is just

$$L_2 = \frac{p(I|S_2)}{p(I|S_1)} \tag{16}$$

which equals $1/\epsilon$, since the likelihoods are, respectively, 1 and ϵ . If we assume independence of accidents as discussed above, then the ϵ s multiply as we move down the diagram (or add if we take logs). Specifically each node at depth d has likelihood ratio ϵ^d relative to the top (most generic) node,

$$L_S \approx \frac{1}{\epsilon^d} = \epsilon^{-d} \tag{17}$$

or, equivalently, weight of evidence $-d \log \epsilon$. (Note that $\log \epsilon$ is always negative because $\epsilon < 1$, so $-d \log \epsilon$ is always positive.) Figure 3 illustrates several partial order diagrams with the likelihood ratios associated with each node.

Given the simplifying assumptions we have made, the strength of each interpretation depends only on where it sits in the partial order. As we move down the diagram to increasingly complex “coincidences” (i.e., as d increases), the likelihood ratio in favor of the inference that the configuration is not a coincidence increases exponentially. For example, with $\epsilon = 0.05$ (the conventional value for the probability of a “coincidence” in psychology) and $d = 2$ (the depth of collinearity in the diagram in Figure 2), $L = 0.05^{-2} = 400$, 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 rapidly increase in probabilistic compellingness, in an explicitly Bayesian sense. Similarly the weight of evidence (log-likelihood ratio) increases linearly with the depth. Feldman (2007) found experimental evidence that perceptual groups become progressively stronger (more tightly bound into “objects”) in direct proportion to the depth d of the group, exactly as theory would predict.

Of course, Equation 17 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.

A Richer Example

A more complicated, but essentially similar, example concerns configurations of multiple oriented elements, such as line segments or Gabor patches. Here the interpretations—qualitative descriptions of the spatial arrangement of the elements—are not simply qualitative predicates such as “collinear” or “parallel,” but more complex, possibly hierarchical combinations of such relations among various combinations of elements. For simplicity, consider four elements at a time, with only collinearity recognized as a possible relation. Following the assumption of qualitiveness, we classify each pair of elements as collinear or not on the basis of some simple angular threshold (e.g., 30° deviation from perfect

collinearity; the exact criterion does not matter for the example). From a more thorough probabilistic point of view, this threshold acts as a decision boundary separating turning angles α that were likely to have been generated by a collinear-centered distribution (e.g., $\alpha \sim N(0^\circ, \sigma^2)$; see Feldman, 2001) from those that were likely not. From a qualitative point of view, they are simply “collinear” or “not collinear.” Considering each element pair (chosen from the set of four elements) this way gives a set of five possible interpretations, each a tree with leaves corresponding to individual elements, and other nodes the collinearity classification of the spatial relations among the children (see Feldman, 1997b, 2003a, for further discussion of trees like these). The five interpretations (trees) form a partial order of interpretation strength (see Figure 4), in this case exactly isomorphic to Figure 3b). Interpretations lower on the diagram are stronger, in the sense that they dominate when more than one interpretation applies. Thus for an arbitrary configuration, we select the lowest interpretation that applies. Intuitively, the illustrated configurations (Gabor fields) form stronger or more “prägnant” patterns as one moves down the

lattice. The totally generic (top) interpretation treats each of these angles as random events; the totally curvilinear (bottom) interpretation treats the entire configuration as a single chain. Each step down the lattice amounts to one additional collinearity expected under the model and thus explained by the model. The chain interpretation is stronger than the null interpretation by an amount that is proportional to the number of coincidences it explains—its depth. More specifically, if we assume equal priors, the log-likelihood ratio increases linearly as we move down the lattice, following the epsilon powers in Figure 3b. This is the simplicity principle *and* the likelihood principle.

It is instructive to compare this qualitative analysis with a more full-blown Bayesian analysis of this situation, in which we evaluate the quantitative likelihoods of each configuration under each hypothesis, instead of considering the configurations only qualitatively. Consider a chain of n oriented elements, corresponding to a sequence of turning angles $\{\alpha_1, \alpha_2, \dots, \alpha_{n-1}\}$ (corresponding to the lowest interpretation in the diagram in Figure 4, where $n = 4$). This is the crucial case of the integration of an elongated contour, extensively studied by many authors (e.g., Field, Hayes, & Hess, 1993). Along such a contour, we assume the angles α_i are generated independently and identically from a collinear-centered Gaussian³ $N(0^\circ, \sigma^2)$, which means that the likelihood will be given by

$$p(\{\alpha_i\}|\text{smooth chain}) = \prod_i \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{\alpha_i^2}{2\sigma^2}}, \quad (18)$$

the product of $n - 1$ independent Gaussian deviates. By contrast, the generic or null interpretation (the topmost node in the lattice) treats these same angles $\{\alpha_i\}$ as a series of independent accidents, each with probability ϵ , and thus has likelihood

$$p(\{\alpha_i\}|\text{null}) = \epsilon^{n-1}. \quad (19)$$

As suggested above, the strength of the chain interpretation relative to the null is given by the ratio of its likelihood to that of the null interpretation,

$$L_{\text{smooth chain}} = \log\left(\frac{\text{likelihood of smooth chain}}{\text{likelihood of null}}\right), \quad (20)$$

or the logarithm of this ratio. Dividing Equation 18 by Equation 19 and taking the log yields

$$\log L_{\text{smooth chain}} = \overbrace{-\frac{1}{2} \sum_i \left(\frac{\alpha}{\sigma}\right)^2}^{\text{smoothness term}} + \overbrace{(n-1) \log\left(\frac{1}{\epsilon \sigma \sqrt{2\pi}}\right)}^{\text{depth term}}. \quad (21)$$

This expression breaks down in an edifying way. The first term on the right-hand side (the negative sum of the squared z scores of the angles along the chain, labeled the *smoothness term*) gets more negative as angles in the chain get larger, and thus the chain gets

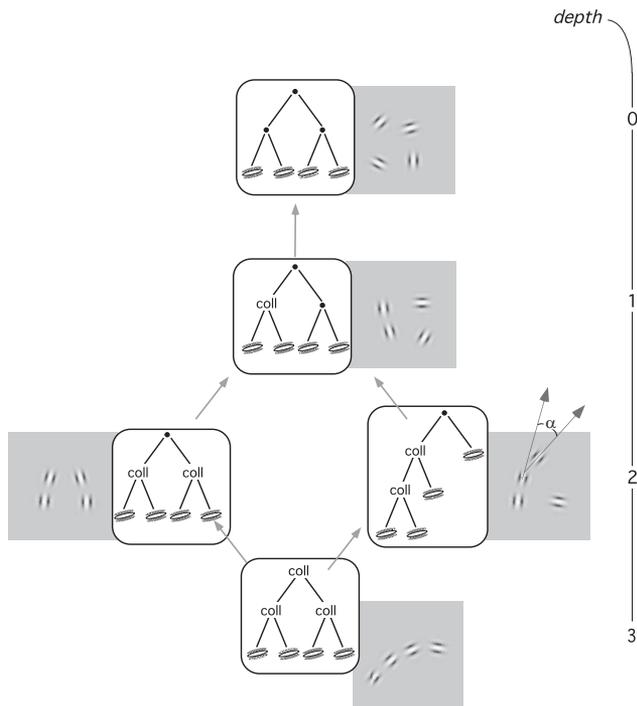


Figure 4. The lattice of interpretations (in this case, trees) of four oriented elements, along with typical instances (Gabor fields). Each tree describes a distinct scene qualitative interpretation; stronger interpretations are lower on the lattice, weaker ones higher. The perceived interpretation is the lowest one consistent with the image. Nodes in each tree represent classifications of the angle α . An angle is classified as (qualitatively) collinear (denoted *coll*) if α is less than some threshold θ ; otherwise it is generic, indicated here by a dot (\cdot). Thus each configuration of Gabors is assigned a tree on the basis of its qualitative pattern of component angles. Notice how moving one rung down the lattice always involves “fixing” (in this case setting to near collinear) one of the angles α . This particular lattice is isomorphic to that in Figure 3b, and its likelihood ratios follow those given there, assuming $\epsilon = p(\alpha < \theta|\text{generic})$. A color version of this figure is available on the Web at <http://dx.doi.org/10.1037/a0017144.supp>.

³ Actually, for technical reasons it would be a von Mises density (see Feldman & Singh, 2005; Swindale, 1998, for discussion); it makes virtually no difference to the current discussion.

less smooth. (Because it is a negative sum of squared numbers, it is always negative; the smoother the angle chain, the higher, i.e., closer to zero, it is.) Thus the smoothness term reflects the degree to which the observed angles have high likelihood under the hypothesis of a smooth chain (see Claessens & Wagemans, 2008; Feldman, 2001; Singh & Fulvio, 2005). The second term, labeled the *depth term*, increases linearly with $n - 1$, with step size $\log(1/\epsilon\sigma\sqrt{2\pi})$, which is positive as long as $\epsilon < (\sigma\sqrt{2\pi})^{-1}$. The number $n - 1$ —the number of angles in the chain—is just the depth of the chain, that is, the row number on which this hypothesis would sit in the lattice of interpretations; this is the number of coincidences the chain hypothesis would explain. So this component of the log-likelihood ratio increases linearly with the strength of the hypothesis in question. The step size (the term multiplied by $n - 1$) is the margin by which explaining each nearly collinear angle is better than not explaining it. In essence the smoothness term reflects how well the angles fit a curve-chain hypothesis—the *quantitative* goodness of fit to the hypothesis—whereas the depth term reflects the strength of the curve-chain hypothesis itself—the *qualitative* strength of the hypothesis, which increases with its depth. In the literature on contours, smoothness has usually been emphasized (not generally formalized this way) because research tends to focus on close cases with fixed qualitative geometry, and the influence that slight changes in angle might have on them. But when the *ns* among competing hypotheses are not equal—as when one is deciding among different ways of qualitatively subdividing a field of elements into chains of various lengths and geometries—the depth term will dominate. (Remember that in this situation the data still exert a strong influence on decisions via which angles are classified as “approximately collinear” and which are not; we are choosing only among hypotheses that qualitatively fit the data in this sense.) Exactly as discussed above, if one evaluates all hypotheses quantitatively, an elaborate numerical maximization is required. But if one chooses to ignore niceties of the likelihood (i.e., to adopt the “qualitative stance”), then the depth term tells you how strong the hypothesis is, and the maximum-depth rule applies.

Creating a Vocabulary for Scene Description

Above, scene descriptions have been portrayed as “holistic” classifications of the entire image. But in most contexts they are more usefully regarded as *attributes* of scenes, picking out those scenes that satisfy a particular (possibly local) attribute. As such they can occur in combination and compose to form complex scene descriptions. In this sense, individual attributes that received elevated priors as discussed above constitute a kind of alphabet or vocabulary of scene description. Algebraic rules then govern the mechanisms of legal combination of these attributes, whereas the associated Bayesian interpretation developed here means that each composite scene interpretation comes with a well-motivated likelihood ratio. As in Goodman et al. (2008), this approach shows how we can have a productive system of complex scene descriptions coupled with a rational Bayesian inference procedure—again, reconciling ideas that (in the guise of the traditional simplicity and likelihood principles) once seemed like competing approaches.

A simple but important example comes again from Biederman’s RBC (geon) theory. As portrayed in Figure 1, geons form a partially ordered set of inclusion classes, in this case a lattice G . (G

includes only a subset of all geons for purposes of discussion; a more complete set would form a more complex partial order.) But geons do not appear in isolation, but rather in the context of other geons, with which they combine via a variety of spatial relations. A representative sample might include collinear parts, coterminating parts with arbitrary joint angles, and T-junctions. These three spatial relations form a lattice R of inclusion classes, illustrated with straight bricks in Figure 5. As with the geon lattice, in this lattice, relations lower in the order (e.g., collinear parts) are special cases (measure-zero subsets) of relations higher in the lattice (e.g., coterminating parts). Just as the geons themselves form a vocabulary of part types, the relations in the lattice form a vocabulary of relation types. And as with the geons, relations lower in the lattice have a Bayesian (likelihood ratio) advantage over those higher in the lattice, meaning that the preference order in the lattice embodies a rational relation classification procedure.

To build complete object representations, the classification of an individual geon can be composed with those of another geon, and with the spatial relation between them, forming a complex set of object classifications (in this case, of two-part objects). The new set of complex interpretations then forms a partial order composed from the geon and relation lattices. Following standard rules for composing partial orders (Davey & Priestley, 1990), the lattice for two-part objects would then simply be the Cartesian product of the lattices,

$$G \times R \times G, \tag{22}$$

with a partial order inherited from those on G and R , and defined by

$$(g_1, r_1, h_1) \leq (g_2, r_2, h_2) \text{ if } g_1 \leq g_2, r_1 \leq r_2, h_1 \leq h_2 \tag{23}$$

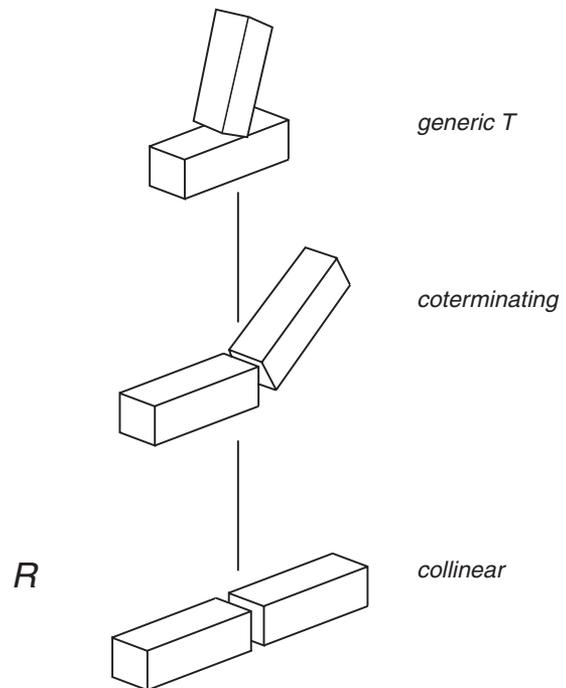


Figure 5. A simple lattice R of relations between two geons. Relations are illustrated with simple “bricks,” but these are placeholders; each role can be played by any geon class (e.g., those in Figure 1) to create a multiplicative space of ordered object types.

(with $g_i, h_i \in G, r_i \in R$). This large lattice (not depicted as it would contain $5 \times 3 \times 5 = 75$ nodes) gives a systematic enumeration of the set of possible two-part objects that can be constructed from these five part types and these three relation types (obviously a small subset of those envisioned by RBC theory). If as above we assume independent “accidents,” all the ϵ s multiply and these classes now come equipped with well-motivated Bayesian likelihood ratios. This allows a rational classification of images into object classes—a Bayesian realization of RBC.

An interesting wrinkle concerns how the nodes in these lattices are “spelled.” Abstractly, any lattice can be regarded as having been generated by a small subset of its elements coupled with the meet (\wedge) and join (\vee) operators. Most elements on the lattice can be expressed as meets (or joins) of other elements. But a special subset are irreducible, meaning that they cannot be expressed as combinations of others. For example, immediate children of the top node are atomic in terms of the meet operator, because they cannot be expressed in terms of meets of others. Dually, immediate parents of the bottom node cannot be expressed as joins of others. In addition, depending on the structure of the lattice, certain other internal nodes may be irreducible in terms of either meets or joins. The set of irreducible elements thus forms a kind of atomic vocabulary for expressing the complete set of elements, just as the prime numbers constitute the atoms for forming unique representations of the natural numbers (via multiplication). The entire lattice can be re-expressed either down from the top (using meet-irreducible elements and the meet operator) or up from the bottom (using join-irreducible elements and the join operator). Either way, the lattice gives an explicit, generative compositional structure to the space of interpretations.

But in a perceptual context, spelling from the bottom and spelling from the top entail substantively different semantics. Meet-irreducible elements (spelling from the top; see Figure 6, left) are *regularities*—special configurations of the image, like nonaccidental properties. Join-irreducible elements (spelling from the bottom; see Figure 6,

right) are *operations*, that is, transformations of one structure to yield a family of transformed versions. Regularities fix a free parameter, whereas operations set a fixed parameter free. The algebra thus makes explicit the relationship between traditional feature-based representations of image structures, which describe the image in terms of its properties, with “generative” approaches, which describe it in terms of the operations that might have produced it (e.g., Feldman & Singh, 2006; Leyton, 1989, 1992). In Figure 6, the geon lattice is shown spelled both ways, with semantic labels on the geon classes illustrating the two types of semantics. For example, the second node from the bottom can be regarded as a brick with a straight axis and parallel sides (left) or as a stretched cube (right). The first way, it is two degrees down from (more regular than) the top; the second way, it is one operation from (more general than) the bottom. Either way, the node has depth $d = 2$ and thus likelihood ratio $1/\epsilon^2$. No matter how it is spelled, the model has a strength that can be quantified in Bayesian terms.

Again, adopting a particular set of geons, spatial relations, or contour properties (etc.) means elevating these choices to a kind of vocabulary of object construction. By adopting them we have in effect elevated this vocabulary from the infinite space of possible alternative vocabularies, giving them priors that are elevated relative to arbitrary classes. After all, *any* spatial relation could, if we chose, be regarded as a primitive type. But if all relations are regarded as types, then there can be no generalization, no recognition of like kinds, and thus no meaningful image description. The configurations we anoint as types are the ones we think tend to occur, and whose combinations correspond approximately to the cases prevailing in the environment.

Discussion

Summarizing, the above argument shows that the maximum-depth rule instantiates a kind of qualitative Bayesian perceptual inference. Many instances of visual inference can thus be regarded

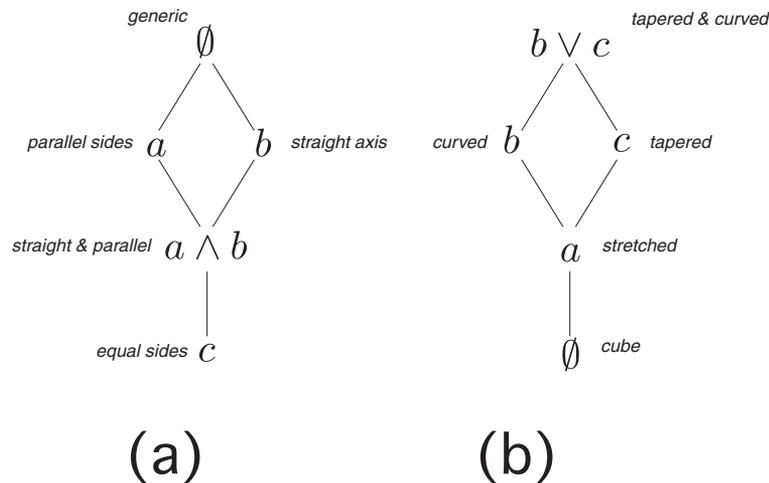


Figure 6. Two alternate ways of “spelling” the geon lattice G (see Figure 1) in terms of atomic elements: (a) down from the top, in terms of meet-irreducible elements (regularities), or (b) up from the bottom, in terms of join-irreducible elements (generative operations). Nodes marked by letters (a, b, c) are irreducible elements; other nodes ($a \wedge b, b \vee c$) are combinations (meets or joins). The hierarchical relations among the meet-irreducible elements and the hierarchical relations among the join-irreducible elements are related by a famous theorem of lattice theory, Birkhoff’s representation theorem (see Davey & Priestley, 1990).

as rational, albeit approximate, attempts to infer the best explanation for image data. Key aspects of this account have intriguing parallels in the literature on Bayesian models of higher level cognition, perhaps suggesting common principles at work across domains. Chater (1996) had shown that perceptual simplicity principles are at least asymptotically equivalent to Bayes; here we see a perceptual simplicity principle that is exactly equivalent to Bayes 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. More broadly, this argument demonstrates the utility of bringing algebraic structures and concepts to bear in a Bayesian context, allowing inference to be representationally expressive and both probabilistically optimal.

Maximizing the depth of an interpretation is perhaps best viewed as a realization of Rock's (1983) coincidence explanation principle. Each image satisfies some number of properties that are unlikely to occur by accident (i.e., occur independently with probability ϵ). These properties are "explained" under any model in which they are generic, that is, that assigns them high probability; but they are left unexplained by more generic models, which treated them as ϵ -probability coincidences. In this sense the maximum-depth interpretation is simply the one that explains as many image properties as possible. Indeed it may not be possible to explain all the coincidences in the image, but the maximum-depth interpretation is simply the one that explains as many as possible (Jepson & Richards, 1991; Richards, Jepson, & Feldman, 1996). From a Bayesian point of view, it is also the maximum a priori model from among those that can be expressed in the description language we have chosen to adopt. The "words" in this language are simply the meet-irreducible elements in the corresponding partial order or lattice—or, dually, the join-irreducible elements if we choose to express our interpretations as generative models. This establishes a close relationship between these compositional representations (along with their associated preference ordering) and maximization of the Bayesian posterior.

Of course, qualitative inference yields only qualitative conclusions. Nevertheless, in many perceptual settings crucial *quantitative* parameters are modulated by qualitative factors. For example, perceived luminance (a quantitative parameter) changes depending on whether or not the surface perceptually completes with another (a qualitative decision; Gilchrist, 1977; Gilchrist & Jacobsen, 1989). The perceived orientation of a shape (quantitative) depends on how it decomposes into parts (qualitative; Cohen & Singh, 2006). Historically, though, these types of qualitative perceptual decisions have been more poorly understood than quantitative ones. A more complete understanding of qualitative perceptual decision making thus has the potential of furthering (rather than preempting) a more complete quantitative understanding of perceptual function.

The close relationship between depth minimization and posterior maximization is predicated on a number of simplifying assumptions. Naturally, if key assumptions (like equality of priors) are incorrect in a given environment, the resulting conclusions will deviate from optimality. As with any inference theory, the correctness of the conclusions requires that the assumptions be suitably "tuned" to the environment (see Brunswik, 1956; Geisler & Diehl, 2002). Indeed, the compositional structure of a given interpretation space *depends* on adopting a description language that correctly

reflects environmental regularities, in the sense that it assigns high prior to just those events that tend to occur frequently (which in the algebra become the irreducible elements and combinations thereof). Of course, in any given environment, certain legal combinations may in reality occur more often than others, in which case the assumption of equal priors over the partial order would be incorrect and the resulting interpretations nonoptimal. An obvious example would be certain combinations of geons that occur more than other combinations, such as the standard quadruped body plan. (Such cases form a sublattice with its own combinatorics.) In the end, if the aim is to achieve quantitatively optimal inferences, then a more conventionally quantitative approach is required. But with approximately correct assumptions, algebraic minimization can achieve approximately correct results with a minimum of computation, which under real time-pressured conditions might be the more adaptive strategy (Brighton & Gigerenzer, 2008).

Conclusions

For those perceptual theorists who have fretted over the philosophical justification of the simplicity principle, the idea of complexity minimization has seemed at times no more than a handy but unjustifiable calculating trick, whose empirical success was essentially mystifying (Hatfield & Epstein, 1985). Chater's (1996) argument goes a long way toward clearing up this mystery: in a very general but also somewhat abstract sense, complexity minimization serves the purpose of building a veridical representation. But because this property is shared by any reasonable complexity measure (any one that is universal in Kolmogorov's sense), this argument does not help clear up the ambiguity in specifying exactly which minimum rule the visual system uses. The argument in the current article is that if the image description vocabulary is well-chosen, maximizing interpretation depth likewise maximizes the posterior. This brings out in a more explicit way how the selection of an optimal interpretation relates to the assumptions one has adopted—regarded either as a set of models assigned high priors, or as an image-description vocabulary.

In this sense, this brings out the idea that the various rules and strategies imputed in perceptual inference differ not only in the computations they specify but in the knowledge and assumptions they implicitly embody. Different assumptions license different computational mechanisms. Here, certain assumptions about priors and likelihoods transmuted Bayes' rule into a very different-looking—but mathematically equivalent—formal minimization rule. This idea recasts the historic debate between the simplicity and likelihood principles as one about alternative *assumptions* rather than one about alternative *principles*. Indeed, the question we should be asking is not what computational tricks the visual system uses, but rather what constraints and assumptions about the world are embodied in the tricks (Barlow, 1994; Marr, 1982; Richards, 1988).

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.
- Barlow, H. (1994). What is the computational goal of the neocortex? In C. Koch & J. L. Davis (Eds.), *Large-scale neuronal theories of the brain* (pp. 1–22). Cambridge, MA: MIT Press.

- Biederman, I. (1987). Recognition by components: A theory of human image understanding. *Psychological Review*, *94*, 115–147.
- Brighton, H., & Gigerenzer, G. (2008). Bayesian brains and cognitive mechanism: Harmony or dissonance? In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for a Bayesian cognitive science* (pp. 189–208). New York, NY: Oxford University Press.
- Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley, CA: University of California Press.
- Buffart, H., Leeuwenberg, E. L. J., & Restle, F. (1981). Coding theory of visual pattern completion. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 241–274.
- Bülthoff, H. H., & Yuille, A. L. (1991). Bayesian models for seeing shapes and depth. *Comments on Theoretical Biology*, *2*, 283–314.
- Caelli, T. M., & Umansky, J. (1976). Interpolation in the visual system. *Vision Research*, *16*, 1055–1060.
- Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, *103*, 566–581.
- Chater, N. (2005). A minimum description length principle for perception. In P. Grünwald, M. Pitt, & I. Myung (Eds.), *Advances in minimum description length: Theory and applications* (pp. 372–398). Cambridge, MA: MIT Press.
- Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science. *Trends in Cognitive Sciences*, *7*, 19–22.
- Claessens, P. M. E., & Wagemans, J. (2005). Perceptual grouping in Gabor lattices: Proximity and alignment. *Perception & Psychophysics*, *67*, 1446–1459.
- Claessens, P. M. E., & Wagemans, J. (2008). A Bayesian framework for cue integration in multistable grouping: Proximity, collinearity, and orientation priors in zigzag lattices. *Journal of Vision*, *8*, 1–23.
- Cohen, E. H., & Singh, M. (2006). Perceived orientation of complex shape reflects graded part decomposition. *Journal of Vision*, *6*, 805–821.
- Davey, B., & Priestley, H. (1990). *Introduction to lattices and order*. Cambridge, England: Cambridge University Press.
- De Baene, W., Ons, B., Wagemans, J., & Vogels, R. (2008). Effects of category learning on the stimulus selectivity of macaque inferior temporal neurons. *Learning & Memory*, *15*, 717–727.
- De Baene, W., Premereur, E., & Vogels, R. (2007). Properties of shape tuning of macaque inferior temporal neurons examined using rapid serial visual presentation. *Journal of Neurophysiology*, *97*, 2900–2916.
- Dowe, D. L., Gardner, S., & Oppy, G. (2007). Bayes not bust! Why simplicity is no problem for Bayesians. *British Journal for the Philosophy of Science*, *58*, 709–754.
- Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. New York, NY: Wiley.
- Feldman, J. (1997a). Curvilinearity, covariance, and regularity in perceptual groups. *Vision Research*, *37*, 2835–2848.
- Feldman, J. (1997b). Regularity-based perceptual grouping. *Computational Intelligence*, *13*, 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*, 1171–1182.
- Feldman, J. (2003a). Perceptual grouping by selection of a logically minimal model. *International Journal of Computer Vision*, *55*, 5–25.
- Feldman, J. (2003b). What is a visual object? *Trends in Cognitive Sciences*, *7*, 252–256.
- Feldman, J. (2007). Formation of visual “objects” in the early computation of spatial relations. *Perception & Psychophysics*, *69*, 816–827.
- Feldman, J., & Singh, M. (2005). Information along contours and object boundaries. *Psychological Review*, *112*, 243–252.
- Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences, USA*, *103*, 18014–18019.
- Field, D. J., Hayes, A., & Hess, R. F. (1993). Contour integration by the human visual system: Evidence for a local “association field.” *Vision Research*, *33*, 173–193.
- Freeman, W. T. (1994, April 7). The generic viewpoint assumption in a framework for visual perception. *Nature*, *368*, 542–545.
- Geisler, W. S., & Diehl, R. L. (2002). Bayesian natural selection and the evolution of perceptual systems. *Philosophical Transactions of the Royal Society of London, Series B*, *357*, 419–448.
- Gilchrist, A. L. (1977, January 14). Perceived lightness depends on perceived spatial arrangement. *Science*, *195*, 185–187.
- Gilchrist, A. L., & Jacobsen, A. (1989). Qualitative relationships are decisive. *Perception & Psychophysics*, *45*, 92–94.
- Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*, 108–154.
- Griffiths, T. L., & Tenenbaum, J. B. (2007). From mere coincidences to meaningful discoveries. *Cognition*, *103*, 180–226.
- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.
- Hatfield, G., & Epstein, W. (1985). The status of the minimum principle in the theoretical analysis of visual perception. *Psychological Bulletin*, *97*, 155–186.
- Hochberg, J., & McAlister, E. (1953). A quantitative approach to figural “goodness.” *Journal of Experimental Psychology*, *46*, 361–364.
- Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, England: Cambridge University Press.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, England: Clarendon Press. (Original work published 1939)
- Jepson, A., & Mann, R. (1999). Qualitative probabilities for image interpretation. In *Proceedings of the International Conference on Computer Vision* (Vol. 2, pp. 1123–1130). Washington, DC: IEEE Computer Society.
- Jepson, A., & Richards, W. A. (1991). *What is a percept?* (Occasional Paper No. 43). Cambridge, MA: MIT Center for Cognitive Science.
- 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, NY: Praeger.
- Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*, 271–304.
- Knill, D., & Richards, W. (Eds.). (1996). *Perception as Bayesian inference*. Cambridge, England: Cambridge University Press.
- Kubovy, M. (1994). The perceptual organization of dot lattices. *Psychonomic Bulletin & Review*, *1*, 182–190.
- Kubovy, M., Holcombe, A. O., & Wagemans, J. (1998). On the lawfulness of grouping by proximity. *Cognitive Psychology*, *35*, 71–98.
- Kubovy, M., & Wagemans, J. (1995). Grouping by proximity and multistability in dot lattices: A quantitative gestalt theory. *Psychological Science*, *6*, 225–234.
- Kukkonen, H. T., Foster, D. H., Wood, J. R., Wagemans, J., & van Gool, L. (1996). Qualitative cues in the discrimination of affine-transformed minimal patterns. *Perception*, *25*, 195–206.
- Kurbat, M. A. (1994). Structural description theories: Is RBC/JIM a general-purpose theory of human entry-level object recognition? *Perception*, *23*, 1339–1368.
- 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*, 389–412.
- Leeuwenberg, E. L. J. (1971). A perceptual coding language for visual and auditory patterns. *American Journal of Psychology*, *84*, 307–349.

- Leeuwenberg, E. L. J., & Boselie, F. (1988). Against the likelihood principle in visual form perception. *Psychological Review*, *95*, 485–491.
- Leyton, M. (1989). Inferring causal history from shape. *Cognitive Science*, *13*, 357–387.
- Leyton, M. (1992). *Symmetry, causality, mind*. Cambridge, MA: MIT Press.
- Li, M., & Vitányi, P. (1997). *An introduction to Kolmogorov complexity and its applications*. New York, NY: Springer.
- Lowe, D. G. (1987). Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, *31*, 355–395.
- MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge, England: Cambridge University Press.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: Freeman.
- Perkins, D. (1976). How good a bet is good form? *Perception*, *5*, 393–406.
- Pomerantz, J. R., & Kubovy, M. (1986). Theoretical approaches to perceptual organization. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance*, Vol. 2: *Cognitive processes and performance* (pp. 36–46). New York, NY: Wiley.
- 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, NY: Odyssey Press.
- Richards, W. A. (1988). The approach. In W. A. Richards (Ed.), *Natural computation* (pp. 3–17). Cambridge, MA: MIT 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, England: Cambridge University Press.
- Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Singapore: World Scientific.
- Robert, C. (2007). *The Bayesian choice* (2nd ed.). New York, NY: Springer.
- Rock, I. (1983). *The logic of perception*. Cambridge, MA: MIT Press.
- Schöning, U., & Pruim, R. (1998). *Gems of theoretical computer science*. Berlin, Germany: Springer.
- Singh, M., & Fulvio, J. M. (2005). Visual extrapolation of contour geometry. *Proceedings of the National Academy of Sciences, USA*, *102*, 939–944.
- 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, England: Oxford University Press.
- Swindale, N. V. (1998). Orientation tuning curves: Empirical description and estimation of parameters. *Biological Cybernetics*, *78*, 45–56.
- Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.
- Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*, 309–318.
- Van der Helm, P. (2000). Simplicity versus likelihood in visual perception: From surprisals to precisals. *Psychological Bulletin*, *126*, 770–800.
- Van der Helm, P. A., & Leeuwenberg, E. L. J. (1996). Goodness of visual regularities: A nontransformational approach. *Psychological Review*, *103*, 429–456.
- Wagemans, J. (1992). Perceptual use of non-accidental properties. *Canadian Journal of Psychology*, *46*, 236–279.
- Wagemans, J. (1993). Skewed symmetry: A nonaccidental property used to perceive visual forms. *Journal of Experimental Psychology: Human Perception and Performance*, *19*, 364–380.
- Wagemans, J. (1999). Toward a better approach to goodness: Comments on Van der Helm and Leeuwenberg (1996). *Psychological Review*, *106*, 610–621.
- Wagemans, J., van Gool, L., Lamote, C., & Foster, D. H. (2000). Minimal information to determine affine shape equivalence. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 443–468.
- 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). New York, NY: Academic Press.

Received June 11, 2008

Revision received July 8, 2009

Accepted July 9, 2009 ■