Advances in Minimum Description Length Theory and Applications

edited by
Peter D. Grünwald
In Jae Myung
Mark A. Pitt

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

Contents

15 A Minimum Description Length Principle for Perception

Nick Chater
Institute for Applied Cognitive Science
Department of Psychology
University of Warwick
Gibbet Hill Road
Coventry
CV4 7AL
UK
N.Chater@warwick.ac.uk

Perception involves inferring the structure of the environment fromsensory data. As in any process of inductive inference, there are infinitely many hypotheses about environmental structure that are compatible with sensory data. Can the perceptual system use the minimum description length (MDL) principle, which prefers hypotheses that provide a short explanation of the data, to choose between these competing explanations? This viewpoint has a long history in psychology, which can be traced back to Ernst Mach and to Gestalt psychology. This chapter considers how the MDL approach relates to apparently rival principles of perception, what types of empirical data the approach can and cannot explain, and how an MDL approach to perception might be augmented to provide an empirically adequate framework for understanding perceptual inference.

15.1 What is the Simplicity Principle?

Since Helmholtz [1925], many perceptual theorists have viewed the problem of understanding sensory input as a matter of inference (e.g., see [Gregory 1970; Marr 1982; Rock 1981]). That is, they have viewed the problem of perception as an abstract inferential problem of finding the "best" explanation of the sensory input.

The very viability of this highly abstract approach may seem in doubt: Can we really abstract away from the wealth of experimental results concerning the functioning of the perceptual system, and the detailed knowledge that is being acquired about the neurophysiology of the perceptual processing systems? The assumption in this chapter is that viewing perception as inference, and attempting to understand what kind of inference principles the perceptual system may use, is that it may provide a theoretical framework in which to organize and understand experimental and neuroscientific findings. Unless we understand, on an abstract level, *what* the perceptual system is doing, we have little hope of making sense of the psychophysical and neural data concerning *how* the perceptual system functions.

So, taking this abstract point of view, the task of the perceptual system is to find the "best" explanation of sensory input, in terms of the objects, surfaces, and lighting that may have given rise to that input. But what does it mean to be the *best* explanation? Clearly it is not just that the explanation fits the sensory input—because in perceptual problems there are invariably vast numbers of interpretations that fit the sensory data. Something other than mere "data fit" must differentiate the one (or, for rare ambiguous inputs, more than one) plausible interpretation from the plethora of implausible interpretations.

The minimum description length (MDL) principle discussed in this book provides an attractive framework for addressing the choice of interpretations. If the perceptual system follows the MDL principle, it should prefer interpretations that can be used to provide *short* descriptions of sensory data. From a purely abstract standpoint, an MDL approach to perception has a number of attractions:

- The MDL principle is widely and successfully used in statistics and machine learning, to tackle practical tasks where structure must be found in data (see, e.g., [Gao, Li, and Vitányi 2000; Quinlan and Rivest 1989; Kontkanen, Myllymäki, Buntine, Rissanen, and Tirri 2005]).
- "Ideal MDL" [Vitányi and Li 2000] can be justified on theoretical grounds as choosing explanations of data that give reliable predictions, given quite general assumptions.
- MDL is very closely related to Bayesian inference methods [Chater 1996; Vitányi and Li 2000] which are widely used in computational theories of perception [Knill and Richards 1996]
- MDL-based methods have been applied successfully in computational models of perception [Mumford 1996; Bienenstock, Geman, and Potter 1998]

These considerations suggest that the perceptual system would do well to use the MDL principle to analyze perceptual input. The question addressed here is: Does it? Or, more precisely, to what extent, and in what ways, does the MDL principle provide a productive framework for theorizing about perception?

This chapter has the following structure. The first section, *The power of simplicity: MDL-style theories of perception*, briefly outlines the MDL-type viewpoint on perception. We also describe the kinds of perceptual phenomena that this approach has been viewed as explaining. The second section, *Simplicity is not enough: Empirical challenges for MDL theories of perception*, we describe various classes of empirical phenomena that appear to pose difficulties for an MDL approaches to perception. The third section, *Where next? Prospects for an MDL approach to perception*, considers how these challenges might be addressed within an MDL framework.

15.2 The Power of Simplicity: MDL-Style Theories of Perception

The physicist and philosopher Ernst Mach [1959] proposed that the goal of perception, and for that matter the goal of science, is to provide the most economical explanation of sensory data. Thus, for example, the postulation of rigid, opaque objects may provide an economical explanation of the flow of sensory input as an observer moves through the environment. As the observer moves, the sensory input does not change in a random way—rather it unfolds predictably, given the principles of geometry and optics. The postulation of a specific three-dimensional (3D) environment provides a concise summary of these perceptual inputs. Indeed, this 3D structure also simultaneously explains other features of sensory inputs, most notably the patterns of relative disparity between the two eyes of the locations of the same objects. Common sense suggests that the world has a 3D structure, and that this gives rise to these and many other"depth cues" in the sensory input. Mach's perspective inverts this logic : the reason that we perceive a 3D world of depth at all is that this interpretation provides such an economical summary of many aspects of sensory input.

A very different example is that ofcolor. The spectrum of light that is reflected from a surface depends on the reflectance function of the surface and the spectrum of the incident light. So the spectral properties of light bouncing off a particular green apple will differ enormously depending on the illuminant: whether the apple is in sunlight, shade, or artificial light. Yet its perceived color under these widely varying circumstances is roughly constant [Shepard 1992]. Crudely, perceived color captures aspects of the surface reflectance function of the individual surfaces and objects in the scene and this is independent of the illuminant. According to the principle of economy, the invariance of color is justified because it serves to provide an economical description of the spectra across a whole scene. The spectrum of the illuminant, combined with illuminant-invariant reflectance functions for different kinds of object or surface in the scene, provides a brief explanation of the apparently chaotic spectral variation across different parts of a scene, and over changes in illumination.

A third class of examples are the"laws of form" studied by the Gestalt psychologists (e.g., Koffka [1935]). Consider, for example, the law of common fate—that a collection of items that move coherently are seen as parts of a single object or surface. This provides an economical explanation of the movements of all the items involved—because rather than having to describe the movement of each item separately, the motion of the whole group is specified at once. In the same way, grouping by proximity, and by similarity, provides an economical way of encoding the shape, color, or other properties of many items simultaneously.

To state the same point more generally, one goal of perception is to find patterns in sensory input, and the MDL principle provides an elegant way of deciding between alternative patterns—by favoring patterns that support economical descriptions of sensory input.

In informal terms, the MDL approach to perception was proposed, as we have noted, by Mach. Mach's viewpoint was an example of positivism—roughly, the idea that the goal of science is to capture regularities in sensory experience, with the goal of making predictions about future sensory experience. A common sense version of this viewpoint is to claim that by finding the simplest explanation of past sensory experience, the perceiver or scientist is also likely to gain insight into how the real world works: what objects it contains, the nature of their causal powers, and so on. Mach took the more radical view that such apparent insight is illusory. Indeed, for Mach, that the idea that the goal of perception or science might be to find out about the nature of the world is conceptually incoherent. Unusually, for a leading physicist working well into the first part of the twentieth century, he did not believe in the reality of theoretical postulates of physics, such as atoms-these were merely convenient fictions for calculating relations between sensory experiences.

Mach's type of antirealist viewpoint, emphasizing the epistemological primacy of sensation, has not become philosophically unpopular, partly because it seems impossible to reconstruct a world of everyday objects, or scientific theoretical terms, from the language of sensation; and partly because the psychological notion of sensation has come to seem epistemologically problematic as an indubitable foundation for knowledge, given that psychological research has revealed how poorly people can access their own phenomenal experience [Dennett 1991; Laming 1997; O'Regan and Noe 2001]. This does not imply that Mach's view of the goal of perception and science, as providing economical descriptions of sensory experience, is misguided. The last twenty years has seen a substantial revival of interest in Bayesian models in the philosophy of science [Earman 1992; Horwich 1982; Howson and Urbach 1993]; and we shall see in the next section that the simplicity principle and Bayesian inference are closely related. Moreover, there may be advantages to adopting a simplicity-based view of scientific inference, over and above the strengths of the general Bayesian framework [Chater and Vitányi 2003]. Overall, I suggest that to the extent that there is an analogy between scientific inference and perception, as Mach assumed, this analogy may strengthen, rather than undermine, the case for a simplicity principle in perception.

After Mach's discussion, the case for a simplicity principle in perception was developed further by the Gestalt psychologists, in particular in relation to their formulation of laws of form, mentioned briefly above. More formal attempts to understand perception in terms of economy of representation were put forward with the development of information theory, and the developmental of formal theories of visual representation. For example, some theorists attempted to test directly the assumption that the perceptual system prefers short codes, where these are measured by optimal code lengths, given assumptions about the probability distributions over sensory inputs [Attneave 1959; Attneave and Frost 1969; Garner 1962; Garner 1974; Hochberg and McAlister 1953]. The theoretical starting point of this approach is communication theory [Shannon 1948]. This theory specifies, among other things, the code lengths that produce the briefest expected code length required to encode the output of a probability distribution. But measuring complexity relative to a probability distribution is problematic—because it ignores the instrinsic complexity of the objects being perceived. For example, suppose that in an experimental context, a participant may be presented either with a black square on a white background or a high-definition forest scene. If these have the same probability (of one half), then from an information-theoretic standpoint they have the same complexity. But this does not capture the very different perceived complexity of the two stimuli (see [Miller 1967] for more general reflections on difficulties with applying standard information theory to cognition). Moreover, outside the laboratory, we face the problem that the probability distribution over natural images is unknown—indeed, to define it would require knowledge of the full range of regularities that govern the visual world. Hence it is unclear that the information-theoretic approach is applicable, even in principle—because it bases codes on probabilities, and probabilities seem either irrelevant (as in our experimental example) or unknown.

An attractive alternative approach is to take coding, rather than probability, as the basic notion. This is the idea behind an MDL approach to perception explored in this chapter. This approach immediately deals with the case of equiprobable stimuli: the black square can be coded briefly; the forest scene will require a code of enormous complexity. The primacy of coding, rather than probability, also underpins a psychological research tradition that developed in parallel to the mathematical formulation of MDL: structural information theory [Buffart, Leeuwenberg, and Restle 1981; Restle 1970; Restle 1979; Van der Helm and Leeuwenberg 1996; Van der Helm 2000]. The aim of structural information theory is to develop a theory of representation for simple visual inputs, according to which the visual system systematically prefers interpretations that correspond to short descriptions. This approach has been concerned with a particular, and deliberately quite limited, set of coding languages and visual stimuli, so that concrete predictions can be made. In this chapter, I am concerned instead with the general claim that the perceptual system chooses the simplest representation, without commitment to a particular coding language for perceptual input. The key idea, then, is that we consider the possibility that the perceptual system aims to find the shortest representation, given some coding system for perceptual input.

15.2.1 Simplicity and Bayesian Inference

In the literature on perceptual organization [Pomerantz and Kubovy 1986], the simplicity principle is typically contrasted with the *likelihood* principle, that the perceptual system should choose the organization that corresponds to the environmental structure that has the highest probability, given the data. This type of viewpoint was proposed by Helmholtz [1925], and has been developed in parallel with the simplicity-based approach, being advocated by vision scientists [Knill and Richards 1996], as well as researchers on computational theories of human and machine vision [Geman and Geman 1984; Hinton, Ghahramani and Teh 2000; Mumford 1996].

Recent formulations of the likelihood approach use a Bayesian framework. They take the goal of perception to be to choose the perceptual hypothesis or explanation, H, which maximizes $P(H|D)$ for perceptual data, D. Using an elementary rule of probability theory, Bayes' theorem, this is equivalent to choosing the H which maximizes $P(D|H)P(H)$. This means that hypotheses about the organization of the stimulus are favored to the extent that they predict the data well (i.e., $P(D|H)$ is high), and that also have high prior probability ($P(D)$ is high).

The hypothesis, H, that maximizes this term is, of course, the same H that maximizes the logarithm of this term; and this is also the H that *minimizes* the negative of the logarithm of this term, which can be written

$$-\log_2 P(H) - \log_2 P(D|H).$$

Now by Shannon's coding theorem from information theory [Cover and Thomas 1991], we know that an optimal (binary) code length for a probability, p, is $\log_2 p$. Hence we can interpret the above terms as code lengths:

$$\text{codelength}(H) + \text{codelength}(D|H)$$

Putting all this together, we can conclude that the hypothesis, H, that maximizes likelihood can also be viewed as minimizing the code length that the hypothesis provides for the data, where that hypothesis is expressed as a two-part code. That is, the hypothesis that satisfies the likelihood principle also satisfies the simplicity principle, where simplicity is interpreted in terms of code length. This heuristic argument suggests that the simplicity and likelihood principles can be viewed as equivalent [Chater 1996; Mumford 1996]. Indeed, applying theoretical results on the relationship between a version of "ideal" MDL and Bayesian inference [Vitányi and Li 2000], it is possible to prove a "mimicry" theorem [Chater, Vitányi, Olivers, and Watson 2005]. It turns out that, for any "computable" coding language (where this is quite a mild restriction) there is a "dual" probability distribution. This dual probability distribution has the property that, for any data whatever, the hypothesis which assigns that data the shortest length using that specific coding language is also the hypothesis with the highest probability, in the dual probability distribution. Again subject to mild restrictions, the converse result holds. This mimicry theorem implies that any specific version of the simplicity principle can always be reconceptualized as a specific version of the likelihood principle, and vice versa. From this perspective, it is not surprising that the century-long debate between the simplicity and likelihood principles has not been resolved successfully by empirical evidence—because any evidence that can be explained by an account using one principle can be recast by an account using the other. [1].

The close relationship between Bayesian and MDL viewpoints in perception is particularly interesting, given the large amount of theoretical [Knill and Richards 1996], computational [Geman and Geman 1984], neuroscientific [Blakemore 1990; Rieke, Warland, de Ruyter van Steveninck, and Bialek 1997], and psychological [Weiss 1997] work that takes seriously the possibility that perception operates by Bayesian inference. Below, we shall see relations between issues that face the MDL approach to perception with related issues that have been considered in the probabilistic framework.

15.3 Simplicity is Not Enough: Empirical Challenges for MDL Theories of Perception

We have considered the prima facie attractions of an MDL approach to perception. But to understand both the principle and the challenges that it faces as a theoretical framework for perceptual theory requires facing some of the areas in which the approach does not presently seem to be adequate. In this section, we consider three challenges in turn: *the principle of least commitment*, *causality and independence*, and *representation and process constraints.*

15.3.1 The Principle of Least Commitment

In perceiving a stimulus, the visual system delivers a range of representations that abstract away, to varying degrees, from the details of the specific stimulus that is presented. For example, suppose that we see a somewhat noisy picture of a black square on a white background (Figure 15.1a). The perceptual system presumably represents this as a black square; and, indeed, more abstractly, as a square (Figure 15.1b). Accordingly, it will be classified as highly similar to other noisy, or less noisy, stimuli depicting squares (e.g., Figure 15.1c). Moreover, the orientation and size of the square may be separately abstracted (or perhaps thrown away) — although if the square were aligned such that its adjacent corners, rather than

1. There are still some possible controversies surrounding the traditional debate between simplicity and likelihood, however [Van der Helm 2000; Chater et al. 2005]. At an algorithmic level, one or other framework might appear more natural—for example, if it turns out that the brain carries out empirical sampling from probability distributions, for example, using Markov chain Monte Carlo, Gibbs sampling, or related methods, as is embodied in many neural network models, then arguably a probabilistic formulation might seem more natural. Moreover, in statistics, the relationship between MDL-type techniques and Bayesian methods is much debated. Some theorists treat these methods as identical [Wallace and Freeman 1987]; others treat MDL as fundamentally different from, and logically prior to, the Bayesian approach [Rissanen 1986, 1989].

its sides, are aligned with the perceptually salient frame of reference, then it may not be categorized as a square at all, but as a diamond (Figure 15.1d). On the other hand, of course, one could imagine instead that the perceptual system might use much less abstract codes—specifically, it might have a shape template for the precise, although stupendously arbitrary, details of that particular square. And then specifying a particular perceptual input, such as Figure 15.1a, might be merely a matter of specifying a size, rotation, translational location, and luminance,[2] for this extraordinarily precise template. Going to the opposite extreme, one could imagine that the perceptual system might choose instead an overgeneral model of its perceptual image. For example, it might use a "null" model that would assign equal probability to all possible gray-level image configurations, whether they depict black squares, faces, or pure random noise.

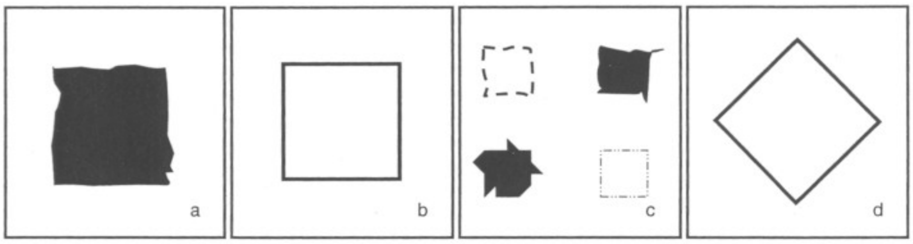

Figure 15.1 Levels of abstraction in processing a simple perceptual stimulus. (a) shows a noisy black square, which the perceptual system may, among other things, represent simple as a square (b), abstracting away from color, irregularities, and so on. (c) shows various other noisy or imperfect squares that may be classified with (a), at this level of abstraction. On the other hand, a change of orientation in relation to the natural frame of reference may lead an apparently very similar figure to be classified, instead, as a diamond (d).

From the point of view of understanding perceptual representations from the point of view of MDL, it is important that we have some way of explaining why "sensible" perceptual representations are to be preferred over these inappropriate types of overspecific or overgeneral representation.

Overgeneral models can straightforwardly be eliminated by MDL—because they do not embody all the constraints that the stimulus actually conforms with, they require excessive code length. To take a trivial example, if an image consists of uniform black, then it will have a near-zero description length in terms of a sensible model that expresses this uniformity. But in terms of the null model that we mentioned above, each "pixel" must be encoded separately, so the resulting bit-string will be enormously large (and no shorter than the code for a white noise, which would be perceived, of course, very differently).

2. We ignore color here to avoid complications.

Overspecific models, by contrast, are more problematic. Suppose that we have some particular set of data (in this case, visual input). Consider the ultimately overspecific model that predicts precisely that these, and no other, possible data can occur. This model provides a maximally efficient representation of the data—no information at all is required to encode the data, given the model. But this is only achieved by putting all the complexity into the specification of the model itself. But doing this does not increase the overall complexity of the encoding of the data—because, after all, encoding this hyperspecific model just *is* a matter of encoding the data, and hence the minimal code for specifying the model will be a minimal code for the data. More generally, there will typically be many overspecific models that attain minimal code length, by including essentially arbitrary features of the data as part of the model itself, and yet still function as minimal codes for the data. Intuitively, this family of models is such that the data are *typical*, and hence uncompressible, given these models. This constraint holds because if the data are not typical, then the data contain patterns that the model does not describe, and hence the models provide an inefficient representation of the data. That is, these models must extract all the *structure* from the data—but the problem of overgeneral models is that models can also encode additional, essentially arbitrary, information about the data.[3]

In statistical theory, a statistic that captures all the structure in the data is called a *sufficient* statistic. Following a suggestion of Kolmogorov, there has been recent interest in a generalization of this idea, which views models as Kolmogorov sufficient statistics (KSSs), a notion that can be formally captured [Cover and Thomas 1991; Gács, Tromp, and Vitányi 2001]. There is frequently a number of KSSs for any given data.[4] Intuition suggests that the KSS of most interest is that which makes the most modest claims about the data (i.e., where the model has the shortest code length, such that the data remain typical with respect to that model). This is the Kolmogorov *minimal* sufficient statistic (KMSS). Perhaps, then, an interesting idealization of perception is as searching not merely for short codes for visual input, but for the KMSS for that input. Thus we have an interesting principle of least commitment for perceptual theory[5]: to derive only that structure from the stimulus that is necessary to explain the data, and no more.

Consider, for example, Figure 15.2. The figure in (a) is perceived as a wire-frame cube with rigid joints. On the other hand, the joints of the figure in (b), which

3. In practical MDL, the focus is often on the code length associated with specifying the parameters in a model to an appropriate level of precision from which to predict the data, rather than on the code lengths of the type of model itself. If used without caution, this approach will favor overspecific models. In practice, this is rarely a problem, as overspecific models, unless chosen post hoc, will almost always be overspecific in the wrong ways, and hence encode the data poorly.

4. A recent important result by Vereshchagin and Vitanyi [2002] shows, remarkably, that the number of KSSs for any data has some, presumably very large, finite bound.

5. This term is used in a different, but perhaps related, context by Marr [1982].

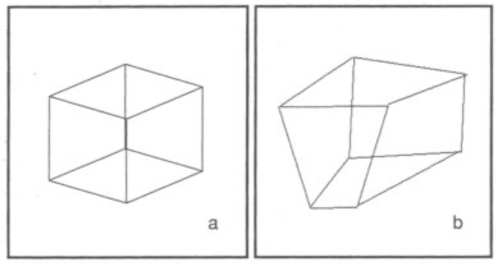

Figure 15.2 Illustration of the principle of least commitment. In (a) the highly regular 2D projection can be explained on the assumption that we are viewing a rigid cube. If the structure is nonrigid (e.g., a frame of rigid rods, attached by loose, flexible joints), then it is a causally inexplicable co-incidence that the rods happen to align to produce a 2D projection consistent with the projection of a rigid cube. The perceptual system disfavors interpretations that involve such causally inexplicable coincidences. In (b), by contrast, a nonrigid interpretation is quite natural (although joints can also be interpreted as rigid).

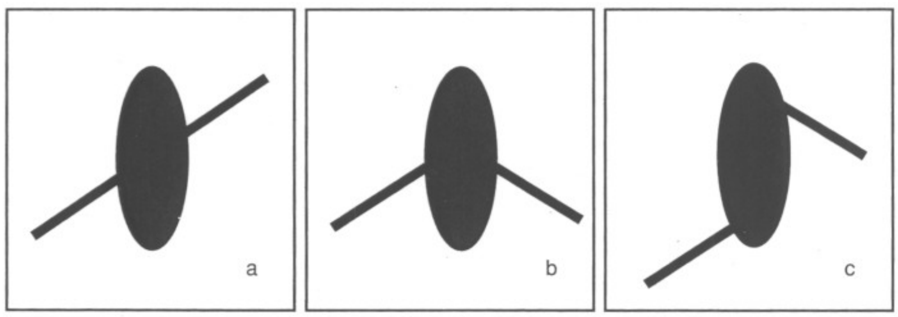

Figure 15.3 Simplicity and rigidity constraints. In (a), the alignment between the two straight lines is interpreted by the perceptual system as significant. That is, the lines are viewed as part of a single rod, occluded by the ellipse. Thus, for example, it seems natural to presume that if one rod were moved by force, the other would also move. That is, the lines are not perceived as merely aligned, but causally connected. This connection is weaker in (b), where the rods are still nonaccidentally related, by symmetry; and is weakened still further in (c) where there is no simple spatial relationship between the two rods.

is perceived as irregular, are perceived as potentially flexible. In the former case, the rigidity assumption is necessary, because if joints are flexible (and hence the whole figure is "collapsible," then it is pure coincidence that the shape happens to be aligned as a cube. The cube alignment is hence not typical with respect to a nonrigid model, which can be rejected. By contrast, in the latter case, the commitment to rigid joints is unnecessary, and hence, by the principle of least commitment we have described, it is not made. Similarly, Figure 15.3 shows how the strength of the interpretation of *attachment* between two linear bars depends on their alignment. However misaligned they are, they *could* be attached (any kind of shape linking them could be present, but occluded by the oval). But the perception of attachment is only present to the degree that this provides a short encoding of the data; if the bars were specified independently, their alignment would be mere coincidence.[6]

We have considered how wildly overgeneral and overspecific models may be eliminated. But this still leaves a huge space of intermediate models. Ultimately, one might hope that from studying the properties of natural images to which the visual system is exposed might provide predictions concerning the nature of the hierarchy of representational levels that are revealed by neuroscientific and psychological research (although, of course, it is unlikely that MDL-type constraints are the only constraints on the perceptual system—one would expect, for example, that the nature of the specific biological hardware will favor some kinds of representation over others [Chater and Oaksford 1990]).

We have also considered only the case where the perceiver has access to a single perceptual input (although this input might be of arbitrary size and temporal duration). But frequently, we do not want to treat the perceptual input in this unanalyzed way—instead, it frequently seems appropriate to view the input as consisting of several independent perceptual inputs, which must be encoded separately— for example, several different views of the same object, from different vantage points; or views of several different objects from the same category (e.g., views of various different dogs). To the extent that we can analyze the perceptual input in this way, we can further resolve the problem of overgeneral models. This is because an overgeneral model fails to explain why multiple, discrete chunks of data all fall within a smaller category. That is, the overgeneral category *animal* may be eliminated as a hypothesis explaining what one is likely to see when visiting a dog show, because it fails to explain why all the animals encountered have so much more in common than would be expected of a typical selection of animals—and this leads to excessively long codes for the representation of the specific animals seen. By contrast, the category *dog* captures these regularities neatly. But we can also readily deal with overspecific models—because multiple data points (multiple examples of dogs) will have a low probability of fitting an overspecific category. If the first dog I encounter when visiting the dog show is a dachshund, I may impute an overspecific hypothesis. But once I see other dogs, it is highly likely that I will see non-dachshunds and hence that this overspecific hypothesis can be eliminated. More generally, similar considerations may help explain how the interpretation of a particular item can be modified by the presence of other items. For example, an apparently arbitrary orientation of a square will suddenly seem significant in the presence of other items with this same orientation; an apparently arbitrary pattern of markings may suddenly become salient if this is repeated (just as an artist's scrawled signature may become salient, upon seeing it in several paintings). That is, some of the ambiguity

6. Strictly, this point relates to our next topic—that MDL can only exploit *causally* explicable regularities.

in determining how to interpret a specific object may be resolved by comparing it with other objects. If, as some theorists suggest, specific instances or exemplar information about past perceptual inputs are stored in memory [Hintzman 1986; Nosofsky 1991], there may be considerable scope for effects of this kind.

This point raises an interesting theoretical question concerning the granularity at which we consider sensory input in analyzing perception. At one extreme, the entire history of sensory stimulation might be treated as a single input, to be compressed; at the other extreme, very spatially and temporally local chunks of sensory information may be considered separately. This question of granularity remains an interesting issue for future research (see also Hochberg [1982] for related discussion). The possibility of breaking sensory input into discrete chunks, which must be encoded separately, also raises the possibility of applying the simplicity principle in perception to categorization—the considerations that we have just described form the basis for a theory of how categories might be learned from samples of their instances, using a simplicity principle. I do not develop these ideas here, but note that simplicity-based ideas have been pursued in building psychological models of both supervised [Feldman 2000] and unsupervised [Pothos and Chater 2002] categorization.

In this subsection, I have argued that merely finding short descriptions of sensory input is not enough. The perceptual system must also be able to separate genuine structure from mere "noise." At a technical level, the Kolmogorov sufficient statistic appears to be a valuable tool for understanding how this separation might occur; and indeed, the hierarchy of Kolmogorov sufficient statistics for sensory input might even relate to the hierarchy of representational levels computed by the perceptual system.

15.3.2 Causality and Independence

Consider the three projections of a cube, shown in Figure 15.4. The rightmost projection (c) is readily interpreted as a cube. The leftmost projection (a) is most naturally perceived as a 'pinwheel'—although it is the projection of a cube viewed precisely along an axis through diagonally opposite vertices. The middle figure, (b), is viewed from a slightly skewed version of this viewing angle—but so that one pair of axes is still aligned. This is slightly easier to interpret as a cube than (a), but is most naturally interpreted in two dimensions as two connected diamonds.

A traditional puzzle in the interpretation of 2D line drawings is that each line drawing could be generated by an infinite number of 3D structures. How does the perceptual system decide on a particular interpretation out of this infinite array of options? We assume that MDL can answer this puzzle: the preference for a cube interpretation, in, say, Figure 15.4, arises because the cube is simpler than

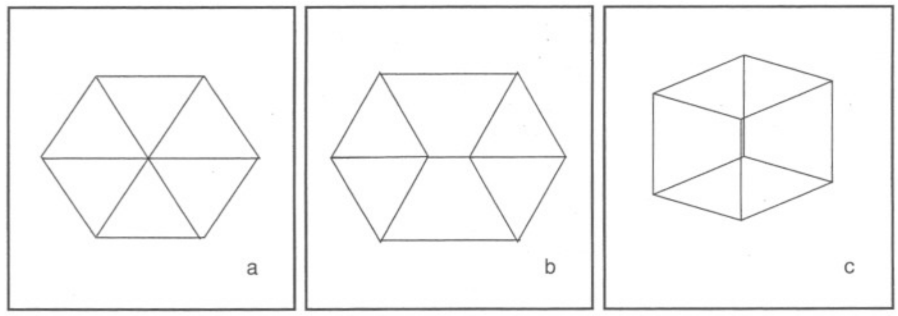

Figure 15.4 Viewpoint dependence and the perception of projections of a cube. This figure shows three projections of a wire-framed cube. The first (a) is seen from a viewpoint according to which two diagonally opposite vertices are in the same line of sight. This is a highly coincidental viewpoint, and the perceptual system disfavors such coincidences. Instead, (a) is perceived as a flat pinwheel. (b) is also projection of a cube from a coincidental viewpoint: two adjacent sides of the cube lie on the same line of sight. (b) is also not readily viewed as a cube, but as a flat structure composed of two diamonds, joined by horizontal lines. A slight misalignment between adjacent sides (c) reduces the coincidence; and also eliminates any simple alternative description in terms of a 2D pattern. Thus, (c) is readily viewed as a projection of a cube.

alternative, less symmetric, polyhedra.[7] But Figure 15.4 illustrates a more difficult problem—that some 2D projections of a cube are more readily interpreted, not as projections of a cube, but as 2D patterns.

This is prima facie puzzling for the MDL approach, because a cube seems to be a rather efficient way of specifying, say, the pinwheel pattern. We need merely specify that the cube is viewed so that two adjacent corners are precisely aligned with the axis of view to obtain the appropriate figure.[8] Though this presumably corresponds to a short code, it relies on a striking coincidence—in this case, concerning the precise alignment of the cube in relation to the viewer. There is a large class of cases of this kind, and theorists in computer vision and perception have proposed that the perceptual system dislikes interpretations that depend on coincidences of viewing angle. Constraints of this kind are known as *generic viewpoint* constraints [Biederman 1985; Binford 1981; Hoffman 1998; Koenderink and van Doorn 1979]. But from an MDL viewpoint, constraints that are viewer-dependent can, nonetheless, lead to short descriptions, and therefore it does not seem that the MDL approach alone can readily capture the perceptual system's tendency to avoid interpretations that rely on precisely specifying a viewpoint.

7. Symmetric figures have shorter codes because they have fewer degrees of freedom.
8. Aside from the obvious further parameters concerning size, orientation, location, type of lines used; but these must be specified, whatever underlying model is used for the geometric structure of the stimulus.

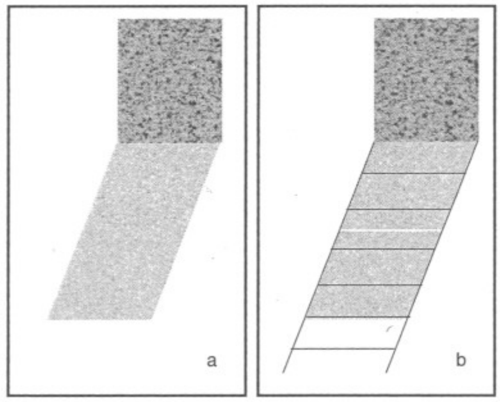

Figure 15.5 Co-incidences and the interpretation of shadows. This figure illustrates how the interpretation of a light bar as the shadow of a darker tower (a) is over-ridden when there is a patterned "path with paving slabs" that precisely coincides with this light bar (b). Intuitively, the perceptual system is biased against co-incidences; if the light bar is a shadow, then it is a striking co-incidence that it happens to line up precisely with the black outlines of the path. Therefore, in (b), the shadow interpretation is disfavored.

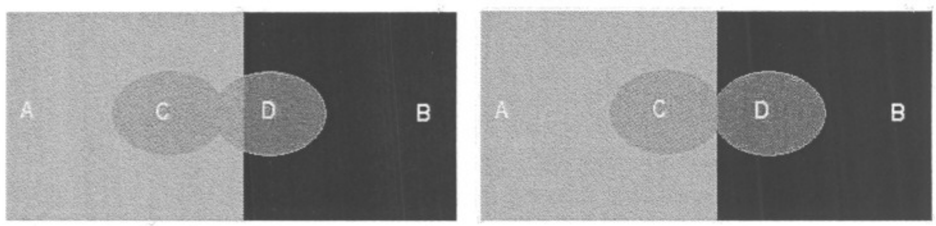

Figure 15.6 Coincidence and transparency. In the left box, the simplest interpretation is that a transparent "spectacle-shaped" object (consisting of patches C and D) lies over a pair of squares, one light gray and one black (A and B). In the right box, this transparency interpretation is impeded, because it involves a coincidental alignment between the overlaid transparent pattern and the squares, so that the indented points in the spectacle-shaped object are precisely aligned with the line between the two squares. There seems to be no causal explanation for this coincidence. Hence, the perceptual system prefers an interpretation in which A, B, C, and D are simply different colored regions, in the same plane.

The perceptual system also appears to be averse to other kinds of co-incidences. For example, Figure 15.5 illustrates how the interpretation of a light bar as the shadow of a tower (Figure 15.5a) is overridden when a patterned "path" coincides precisely with this light bar (Figure 15.5b). Intuitively, the perceptual system is, as before, biased against co-incidences; if the light bar is a shadow, then it is a striking co-incidence that it happens to line up precisely with the path. Thus, instead, the light bar in Figure 15.5b is perceived merely as a path of gray paving slabs, rather than as a shadow. Figure 15.6 shows a further example, using transparency. When the axis of symmetry of the "spectacles-shaped" object (taken from Hoffman [1998]) is precisely aligned with the division between the two background squares, then the transparency interpretation, that is otherwise so strong, disappears. The reason is that, on the transparency interpretation, there are two overlaid surfaces; and there is no apparent causal constraint that would align the patterns on these surfaces. From a causal point of view, this is mere coincidence. Notice that this kind of case is not a matter of viewpoint, and hence not a case to which the generic viewpoint constraint can directly apply. The point here is that the alignment of different parts of the figure is not arbitrary, or "generic," with respect to the causal constraints that appear to be operative.

These cases can all be viewed as indicating that the perceptual system disfavors interpretations according to which there are unexplained, or "coincidental" regularities in the input. That is, perception disfavors interpretation in which there is no *causal* explanation of why, for example, a figure has a particular alignment with the viewer, why shadows happen to line up, or why a spectacle-shaped transparent surface happens to line up precisely with the square-pattern of the surface upon which it is overlaid.

This indicates that we need to add constraints on the kinds of regularities that an MDL model can exploit— that these regularities must be causally natural. After all, the perceptual input is typically the end result of an enormously complex set of causes. The structure of the input is, of course, a product of the laws that govern this causal story; and the goal of the perceptual system is, to some extent at least, to uncover this causal story from the perceptual input. Thus, we have the constraint that the codes derived by an MDL analysis should not be merely short but must correspond to causally viable explanations of the data; codes that rely on short descriptions that are pure 'co-incidence' from a causal point of view (e.g., unexplained alignments) should be rejected.

The addition of this constraint may seem disappointing from the perspective of Mach's positivist program, in perception and in science. For Mach, the apparently somewhat opaque and metaphysical notion of causality should be avoided by science (and hence, presumably, by the cognitive system). Instead, the goal of science is to find the most economical descriptions of the available data. A positivist viewpoint can, perhaps, reconceptualize the current suggestion, without the concept of causality. If we focus not just on a single visual stimulus but over the entire history of stimuli which the perceptual system receives, cases in which co-incidences with no causal explanation will, of necessity, be rare, and hence, these regularities should be assigned long codes, and be disfavored by the MDL principle. Regularities that arise systematically and consistently must, presumably, arise from causal regularities in the perceptual input. These will be frequent, and hence will have short codes, and be preferred by MDL. Despite the possibility of this kind of hard-line MDL approach, in which causal constraints in interpreting any given input are derived from an MDL analysis over the history of perceptual input, it is also entirely possible, of course, that information about causal constraints might be derived from other sources. Thus, causal constraints may be innate [Sperber, Premack, and Premack 1995] or derived from active perceptual-motor experimentation (e.g., moving the head, adjusting a light source, or sliding a transparent sheet, which may disturb co-incidental regularities). Wherever causal constraints are derived from, however, it seems crucially important to incorporate them into a properly constrained MDL account of perception.

We have argued, in this subsection, that there appear to be substantial causal constraints on perceptual interpretations, and these constraints do not immediately arise by favoring the simplest explanation of the perceptual input. It is worth noting, too, that causal constraints are not merely important in predicting which interpretations the perceptual system favors. Rather, extracting a causal model of the external world is a critical goal of perception. This is because the ultimate utility of perception resides, of course, in its ability to drive action; and the effectiveness of action in bringing about the perceiver's objectives will critically depend on correctly perceiving which regularities in the external world are co-incidental and which are causal. Unless the perceiver can distinguish causal structure from perceptual input, it is as likely to attempt to lift a saucer by picking up the cup, as it is to attempt to lift a cup by picking up the saucer. How far the simplicity principle could support the learning of *causal* structure merely by finding the shortest description in observed data is an open question.

15.3.3 Representation and Process Constraints

A strong interpretation of an MDL approach to perception would be that the cognitive system chooses the shortest possible description of the perceptual stimulus. But this strong interpretation cannot be right, at least in general, because the search problem of finding shortest descriptions is typically intractable.[9] For some restricted codes, such as structural information theory, the search problem can be made tractable [Van der Helm and Leeuwenberg 1991]. But in most cases this is not possible. Difficult search problems arise even in the optimization calculations associated with very simple models of pixel-level correlational structure in images [Weiss 1997]; and in training Bayesian or neural networks to find structure in images [Hinton et al. 2000]. So an MDL approach to perception must make a more modest claim: that the perceptual system chooses the shortest code for perceptual data that it *can find*. Thus, the claim is that simplicity is used to choose between perceptual interpretations that are considered by the perceptual system.

Note that the space of interpretations that the perceptual system considers will be highly constrained. There will be an enormous variety of regularities that, when

9. Where the class of code is universal, for example, has equivalence in power to a universal programming language, then finding the shortest code is, in general, uncomputable[Li and Vitányi 1997].

encoded in an image and presented to the perceptual system, will be treated as mere random noise. To choose any of a myriad of possible examples, an "outsize" checkerboard whose black and white squares provide a binary encoding of, say, the initial 10,000 digits of π will have a short code, but will be perceptually indistinguishable from a random checkerboard.

Indeed, a critical question for an MDL model of perception is to draw on existing results from psychology and neuroscience to create realistic hypotheses about which regularities the coding language for perceptual stimuli can capture. Defining a representation or coding language makes, from this point of view, a rich set of hypotheses about the relative strengths of different perceptual interpretations; *of the hypotheses considered*, the shortest should be favored (subject to the caveats concerning causality and least commitment that we have described in the last two subsections). A second critical question is to define search processes over representations, to provide a mechanism for explaining how perception finds interpretations with short description length.[10]

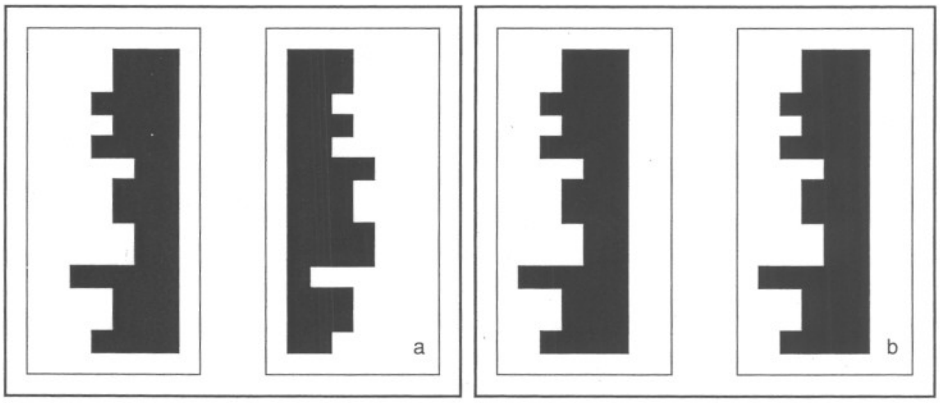

Figure 15.7 How representational format impacts the detection of simple patterns. This figure, loosely patterned on stimuli by Bayliss and Driver [1994, 1995], illustrates how a regularity that allows a stimulus to be encoded briefly may be more or less noticeable, depending on the representational format applied by the perceptual system. In both (a) and (b), there is an identity between the irregular boundaries. But in (a) the boundary is difficult to perceive because the figure/ground relation is reversed (the black shape is the figure, the white forms the background). Hoffman & Richards (1984) point out that boundaries are encoded in terms of convex protrusions of the figure into the ground. Therefore the codes that will be generated for the two irregular boundaries in (a) are very different, because what is coded as a protrusion in one boundary will be viewed as an indentation in the other (and hence will not be coded at all). By contrast, the identity of the boundaries in (b) is readily perceived.

10. Structural information theory is one approach to this problem, though not one we examine here.

To illustrate, consider Figure 15.7, based on stimuli used by Bayliss and Driver [1994, 1995]. Note how the identity between "boundaries" is easily perceived when they have the same figure/ground pattern—but is hard to perceive when they do not. Bayliss and Driver [1994, 1995] explain this difference by adverting to the claim of Hoffman and Richards [1984] that the edges of a figure are parsed in terms of protrusions of the figure into the ground; indentations of the figure are not encoded directly at all. This means that a boundary will be interpreted very differently depending which side of the boundary is viewed as the figure—because this will flip protrusions to indentations and vice versa. Hence, if the same boundary is present in a stimulus, but with a different figure/ground interpretation, then the codes for each will be very different, and the match between them cannot be exploited.

Figure 15.8 *Representation, figure/ground, and symmetry.* Considering each half of the figure separately, there is potential ambiguity concerning whether the black or white stripes are figure or ground. Note that the interpretation according to which the foreground stripes have a constant width is favored. Thus, we perceive white stripes in the left half of the figure, and black stripes in the right half of the figure. This makes sense, from the standpoint of the simplicity principle, only with the auxiliary assumption that the identity between the two parallel contours can only be captured if these contours are assigned to the same object (i.e., the same stripe). On this assumption, the identity between contours assigned to different stripes (in other words, between contours that are perceived as bounding a background region) cannot be captured. According to this line of reasoning, failing to capture this regularity causes the code length, with this representation, to be unnecessarily long, and hence this interpretation is disfavored by the perceptual system. Note, finally, that the symmetry in the stimulus is rather difficult to perceive. This is because perceiving this symmetry requires simultaneously representing the different figure/ground representations (white stripes on black in the left half; black stripes on white in the right half). Thus, representational and processing constraints play a substantial role in explaining which regularities the perceptual system can use to provide a short code for the stimulus.

A related example is shown in Figure 15.8. Note, first, that in the left-hand part of the figure we see white stripes against a black background, and the reverse pattern on the right-hand part of the image. A natural explanation for these figure/ground preferences is that the perceptual system prefers to postulate simpler, rather than more complex, objects. The fore-grounded stripes are simpler than the back-grounded stripes because they have parallel edges. Note, of course, that a pure MDL account would not give this preference for stripes with parallel edges—because, with the reverse segmentation, all that is required is shared contours between adjacent stripes (the left contour of one object would be shared with the right contour of an adjacent object, and so on). But the perceptual code presumably cannot capture this regularity, with the implication that this coding is not favored. Finally, note that the symmetry between the two halves of the pattern is rather difficult to exploit. One explanation for this is that symmetry can only be detected a fter segmentation of both halves of the stimulus has occurred; and simultaneously segmenting both halves of the image presents a difficult processing challenge (e.g., because the segmentations are somewhat rivalrous).

The upshot of examples such as these is that a simplicity approach to perception cannot be considered *purely* in the abstract, in the absence of constraints concerning the representations used by the perceptual system and the cognitive processes that can operate over them. The perceptual system may, if the simplicity principle is correct, aim to find the shortest code for sensory input that it can. But the perceptual system's measure of code length will be influenced by the representational units that it has available, and by the processes that allow (or do not allow) particular regularities to be exploited using those units. Moreover, the perceptual system is limited in which short codes it is able to find—some codes might be highly preferred if they could be found, but they may not typically be found (or may only be found after unusually prolonged viewing, as is frequently the experience of viewers of autostereograms). Now, it seems reasonable to suggest that the specific representations and processes that the perceptual system has evolved will be well suited to the natural environment. And, more speculatively, one might propose that, in the present framework, this might mean that the perceptual system may have representations and processes that are particularly well suited to finding short codes for regularities that occur in the natural environment. Accordingly, we should not, perhaps, be surprised that binary expansions of π are not readily detected by the perceptual system, because presumably there are no natural processes that generate this pattern; and the tendency to specify shapes in terms of convex protrusions may be adapted to the geometry of real objects [Hoffman and Richards 1984]. Indeed, it seems natural to view processes of adaptation as a process of learning over phylogenetic time; this process of learning gradually reduces the code length with which the perceptual system can represent the natural world—because perceptual systems that provide excessively long codes are less well suited to representing real perceptual input, and tend to be eliminated by natural selection. Equivalently, in a Bayesian framework, it seems natural to view one role of natural selection being to select perceptual systems with priors that are as aligned as closely as possible with the actual probability distributions in the natural world.

Overall, the MDL approach to perception must integrate with existing experimental and neuroscientific knowledge of perception. The value of the MDL approach will ultimately depend on whether this integration is smooth and productive, or forced and unhelpful.

15.4 Where Next? Prospects for an MDL Approach to Perception

We mentioned at the outset that one justification for an MDL approach to perception is that perception is a kind of inductive inference, and MDL is a normatively attractive theory of how such inference should be conducted. But this argument is only enough, in reality, to establish that MDL should be given some initial credence. We discussed, too, a number of areas (stereovision, color perception, and laws of form) where some kind of MDL framework can be applied. But whether this framework proves to be scientifically productive depends on how far it can be integrated with the psychology and neurobiology of vision, to provide specific theoretical insights, and empirical predictions.

It is also possible that technical developments in MDL-type approaches to inductive inference might be usefully applicable to understanding aspects of perception.

One intriguing idea, suggested by Paul Vitányi (personal communication, December 2001) is that not merely the Kolmogorov minimal sufficient statistic, but the other nonminimal sufficient statistics might map onto perceptual representations. Perhaps, to some approximation, we might understand the sequence of levels of representation in the perceptual system as corresponding to different Kolmogorov Sufficient Statistics. For example, suppose that we take seriously the idea that very early visual processing filters the input to remove redundancy due to highly local statistical properties of images (e.g., that, in most image locations, luminance changes smoothly, although in a few locations it changes abruptly) [Barlow 1961]. This early filtering might correspond to a (rather long) Kolmogorov sufficient statistic for the image; it throws away mere noise from the image. A slightly more abstract level of representation (e.g., representing edges, blobs, bars) might correspond to another KSS. It throws aways further "noise" of no intrinsic perceptual interest. The Kolmogorov minimal sufficient statistic represents the limit of this process—the most abstract representation that does not throw away anything other than noise. This viewpoint is not intended, of course, to be any more than an idealization of how perceptual representations might be arranged (e.g., the calculations involved in finding minimal codes are provably intractable [Li and Vitányi 1997] and hence cannot, presumably, be computed by the brain). And it is likely that perception throws away a good deal of information— that is, it is much more selective than this analysis allows. Nonetheless, it may be interesting to pursue ideas of this type, in relation to understanding the general function of multiple levels of perceptual representation.

Another interesting line of thought concerns the degree to which we can explain why the output of perception presents the world to us as "modular," that is, of consisting of discrete objects and processes. Let us suppose that, at least at the macroscopic level relevant to much of perception, the regularities in the world are naturally "parsed" into discrete chunks (e.g., objects correspond to bundles of regularities: parts of the same object tend to be made of the same thing, to be attached to each other, to move coherently, to be at roughly the same depth, etc.); and the causal interactions between objects are usually also fairly local (e.g., between a cup, the water pouring from it, and its shadow on the table). To what extent can these sorts of modular structure be uncovered from purely minimizing description length? Current theory does not appear to give clear answers here. Plausibly, if a highly modular process has given rise to perceptual input, one might suppose that a code that replays that modular causal history to reconstruct the data might provide a brief encoding (although if the data sample is insufficiently rich, this may not be true). But there will also be codes of about the same length in which that modular structure is not apparent (just as the careful modular structure of a computer program may be lost from view when it is compiled into a different language). And there might also be short codes with a different, and perhaps contradictory, modular structure. Can MDL-related ideas help to explain how modular causal structures can be inferred reliably from data alone? This seems an important question for future theoretical research—not only, indeed, in helping to understand perception, but in any domain in which MDL may be applied. We want MDL not merely to give us short codes; we want those codes to be interpretable as descriptions that correspond to a local, causal description of the structure of the system that gives rise to the data.

Most crucial for the practical usefulness of the MDL approach to perception is, of course, the degree to which it can be integrated with existing and future research on the psychology and neuroscience of perception. To what extent can neural codes be viewed as compressing sensory input (Barlow [1974], but see also Olshausen and Field [1996])? How far can interpretations of perceptual stimuli be viewed as a choice of the briefest code that can reasonably be entertained, given what is known of the perceptual representations? And, at a theoretical level, to what extent does an MDL framework for perception have advantages over alternative theoretical approaches [Leeuwenberg and Boselie 1998]? The normative appeal and descriptive elegance of an MDL framework, even over the relatively limited territory that we have discussed here, seems sufficient justification for continuing to pursue Mach's dream that perception is driven by a search for economy, that perception is a search for minimal description length.

15.5 Acknowledgments

I thank Ulrike Hahn, Peter van der Helm, Boicho Kokinov, Mark Pitt, Paul Vitányi, an anonymous reviewer, participants at the NIPS workshop on MDL, 2001, and participants at the New Bulgarian University International Summer School in Cognitive Science, 2002, for discussions about these ideas. This work was partially supported by European Commission grant RTN-HPRN-CT-1999-00065, the Human Frontiers Program, the ESRC, the Leverhulme Trust, and Mercer Oliver Wyman.

References

Attneave, F. (1959). *Applications of Information Theory to Psychology.* New York: Holt, Rinehart & Winston.

Attneave, F., and R. Frost, (1969). The determination of perceived tridimensional orientation by minimum criteria. *Perception and Psychophysics, 6*, 391–396.

Barlow, H.B. (1961). The coding of sensory messages. In W.H. Thorpe and O.L. Zangwill (Eds). *Current Problems in Animal Behaviour*, (pp. 330–360). Cambridge, UK: Cambridge University Press.

Barlow, H.B. (1974). Inductive inference, coding, perception and language. *Perception, 3*, 123–134.

Bayliss, G.C. and J. Driver (1994). Parallel computation of symmetry but not repetition in single visual objects. *Visual Cognition, 1*, 377–400.

Bayliss, G.C. and J. Driver (1995). Obligatory edge assignment in vision: The role of figure and part segmentation in symmetry detection. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 1323–1342.

Biederman, I. (1985). Human image understanding: Recent research and a theory. *Computer Vision, Graphics, and Image Processing, 32*, 29–73.

Bienenstock, E., S. Geman, and D. Potter (1998). Compositionality, MDL priors, and object recognition. In M.C. Mozer, M.I. Jordan and T. Petsche (Eds.), *Advances in Neural Information Processing Systems 9.* Cambridge, MA: MIT Press.

Binford, T.O. (1981). Inferring surfaces from images. *Articial Intelligence, 17*, 205–244.

Blakemore, C. (Ed.) (1990). *Vision: Coding and efficiency.* Cambridge, UK: Cambridge University Press/

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

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

Chater, N., and M. Oaksford (1990). Autonomy, implementation and cognitive Architecture: A reply to Fodor and Pylyshyn. *Cognition, 34*, 93–107.

Chater, N. and Vitányi, P. (2003). Inference to the simplest description: An inference principle in science and cognition. Unpublished manuscript.

Chater, N., P. Vitányi, C. Olivers and D.G. Watson (under review). Simplicity and likelihood principles in perception: A mimicry theorem.

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

Dennett, D. C. (1991). *Consciousness Explained.* Harmondsworth, UK: Penguin.

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

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature 407*, 630-633.

Gács, P., J.T. Tromp, and P.M.B. Vitányi (2001). Algorithmic statistics. *IEEE Transactions in Information Theory*, 47, 2443–2463.

Gao, Q., M. Li, and P. Vitányi (2000). Applying MDL to learning best model granularity, *Artificial Intelligence 121*, 1–29.

Garner, W.R. (1962). *Uncertainty and Structure as Psychological Concepts.* New York: Wiley.

Garner, W.R. (1974). *The Processing of Information and Structure.* Potomac, MD: Erlbaum.

Geman, S., and D. Geman. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721–741.

Gregory, R.L. (1970). *The Intelligent Eye.* London: Weidenfeld & Nicolson.

Helm, P. van der, and E. Leeuwenberg (1991). Accessibility: A criterion for regularity and hierarchy in visual pattern code. *Journal of Mathematical Psychology, 35*, 151.

Helm, P. van der, and E. Leeuwenberg (1996). Goodness of visual regularities: A nontransformational approach. *Psychological Review, 103*, 429–456.

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

Helmholtz, H. von (1925). *Physiological Optics.* Volume 3. *The Theory of the Perceptions of Vision.* Translated from the 3rd German edition, 1910. Reprint, New York: Dover.

Hinton, G.E., Z. Ghahramani, and Y.W. Teh (2000). Learning to parse images. In S.A. Solla, T.K. Leen, and K.-R. Muller (Eds.), *Advances in Neural Information Processing Systems, 12*, pp. 463–469, Cambridge, MA: MIT Press.

Hintzman, D. H. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review 93*, 411–428.

Hochberg, J. (1982). How big is a stimulus? In J. Beck (Ed.), *Organization and Representation in Perception*, pp. 191–217, Hillsdale, NJ: Erlbaum.

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

Hoffman, D.D. (1998). *Visual Intelligence: How We Create What We See.* London: Norton.

Hoffman, D.D., and W.A. Richards (1984). Parts of recognition. *Cognition 18*, 65–96.

Horwich, P. (1982). Probability and Evidence. Cambridge, UK: Cambridge University Press.

Howson, C., and P. Urbach (1993). Scientific Reasoning: The Bayesian Approach. Chicago: Open Court.

Knill, D.C., and W.A. Richards (Eds.) (1996). Perception as Bayesian Inference. Cambridge, UK: Cambridge University Press.

Koenderink, J.J., and A.J. van Doorn. (1979). The internal representation of solid shape with respect to vision. *Biological Cybernetics 32*, 211–216.

Koffka, K. (1935). *Principles of Gestalt Psychology.* London: Routledge and Kegan Paul.

Kontkanen, P., P. Myllymäki, W. Buntine, J. Rissanen, and H. Tirri (2005). An MDL framework for data clustering. In P.D. Grünwald, I.J. Myung, and M.A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications.* Cambridge, MA: MIT Press.

Laming, D. (1997). *The Measurement of Sensation.* Oxford, UK: Oxford University Press.

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

Li, M., and P.M.B. Vitányi (1997). *An Introduction to Kolmogorov Complexity and Its Applications.* 2nd edition, New York: Springer-Verlag.

Mach, E. (1959). *The Analysis of Sensations* (Translated by C. M. Williams and S. Waterlow. Original edition, 1886). Reprint, New York: Dover.

Marr, D. (1982). *Vision.* San Francisco: Freeman.

Miller, G. (1967). *The Psychology of Communication.* London: Penguin.

Mumford, D. (1996). Pattern Theory: a Unifying Perspective. In D.C. Knill and W. Richards (Eds.), *Perception as Bayesian inference*, pp. 25–62. Cambridge, UK: Cambridge University Press.

Nosofsky, R.M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. Journal of Experimental Psychology: Human Perception and Performance, 17, 3–27.

Olshausen, B.A., and D.J. Field (1996). Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. *Nature, 381*, 607–609.

O'Regan, J.K., and A. Noe (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences 24.*

Pomerantz, J.R., and M. Kubovy. (1986). Theoretical approaches to perceptual organization: Simplicity and likelihood principles. In K.R. Boff, L. Kaufman, and J.P. Thomas (Eds.) Handbook of Perception and Human Performance. Volume 2, Cognitive Processes and Performance, pp.36-1–36-46. New York: Wiley.

Pothos, E., and Chater, N. (2002). A simplicity principle in unsupervised human categorization. Cognitive Science 26, 303-343.

Quinlan, J.R., and R.L. Rivest (1989). Inferring decision trees using the Minimum Description Length principle. *Information and Computation, 80*, 227–248.

Restle, F. (1970). Theory of serial pattern learning: Structural trees. *Psychological Review 77*, 481–495.

Restle, F. (1979). Coding theory of the perception of motion configurations. *Psychological Review 86*, 1–24.

Rieke, F., D. Warland, R. de Ruyter van Steveninck, and W. Bialek (1997). *Spikes: Exploring the Neural Code.* Cambridge, MA: MIT Press.

Rissanen, J.J. (1986). Stochastic complexity and modeling. *Annals of Statistics 14*, 1080–1100.

Rissanen, J.J. (1989). *Stochastic Complexity in Statistical Inquiry.* Singapore: World Scientific.

Rock, I. (1981). *The Logic of Perception.* Cambridge, MA: MIT Press.

Shannon, C. (1948). A mathematical theory of communication. *Bell Systems Technical Journal 27*, 379-423.

Shepard, R.N. (1992). The perceptual organization of colors: An adaptation to regularities of the terrestrial world? In J.H. Barkow, L. Cosmides, and J. Tooby (Eds.), The Adapted Mind: Evolutionary Psychology and the Generation of Culture. New York: Oxford University Press.

Sperber, D., D. Premack, and A.J. Premack (Eds.) (1995). *Causal Cognition: A Multidisciplinary Debate.* New York: Oxford University Press.

Vereshchagin, N., and P. Vitányi (2002). Kolmogorov's Structure functions with an application to the foundations of model selection. In *Proceedings of the Fourty-seventh IEEE Symposium on the Foundations of Computer Science*, pp. 751–760.

Vitányi, P., and M. Li (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory 46*, 446–464.

Wallace, C.S., and P.R. Freeman (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society, Series B, 49*, 240–251.

Weiss, Y. (1997). Interpreting images by propagating Bayesian beliefs. In M.C. Mozer, M.I. Jordan, and T. Petsche (Eds.), *Advances in Neural Information Processing Systems 9*, pp. 908–915. Cambridge, MA: MIT Press.