Infotropism as the underlying principle of perceptual organization

Chris Thornton
Centre for Research in Cognitive Science
University of Sussex
Brighton
BN1 9QJ
UK
c.thornton@sussex.ac.uk

December 7, 2014

Abstract

Whether perceptual organization favors the simplest or most likely interpretation of a distal stimulus has long been debated. An unbridgeable gulf has seemed to separate these, the Gestalt and Helmholtzian viewpoints. But in recent decades, the proposal that likelihood and simplicity are two sides of the same coin has been gaining ground, to the extent that their equivalence is now widely assumed. What then arises is a desire to know whether the two principles can be reduced to one. Applying Occam's Razor in this way is particularly desirable given that, as things stand, an account referencing one principle alone cannot be completely satisfactory. The present paper argues that unification of the two principles is possible, and that it can be achieved in terms of an incremental notion of 'information seeking' (infotropism). Perceptual processing that is infotropic can be shown to target both simplicity and likelihood. The ability to see perceptual organization as governed by either objective can then be explained in terms of it being an infotropic process. Infotropism can be identified as the principle which underlies, and thus generalizes the principles of likelihood and simplicity.

Keywords: perceptual organization, likelihood, simplicity, information theory, Occam's razor

1 Introduction

Processes of perceptual organization provide a fascinating insight into the interpretive faculties of the human mind. While we generally have no conscious awareness of any separation between cognitive and perceptual levels of interpretation, visual stimuli can be devised that show this to exist. The two levels can operate quite independently in some situations. The classic demonstration is the Necker cube. This is a wire-frame image of a cube, drawn so as to eliminate the foreshortening that would normally suggest a particular orientation (see Figure 1). On viewing this image, people usually recognize it to represent a 3-dimensional cube, but report a perception that flips slowly back and forth between two orientations. This reveals the degree to which perceptual organization can operate as an independent process of interpretation.

The Necker cube is far from unique in this respect. Hundreds of examples now exist of visual stimuli that prompt unstable, surprising, impossible, unpredictable, humorous or erroneous interpretations. Such 'visual illusions' demonstrate the remarkable extent to which perceptual processes can march to their own beat (see Figures 2 and 3). At the same time, they beg the question of principles. The more we recognize the potential decoupling of perceptual interpretation, the more we would like to know the principles that apply. If perceptual organization proceeds according to specific rules, it is important to know what these are.

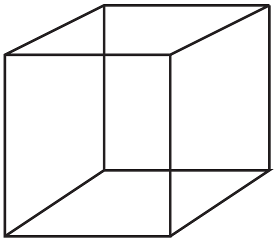

Figure 1: The Necker cube.

The attempt to understand the principles of perceptual organization has been underway for well over a century. In the early years of the 20th Century, theorists of the Gestalt school tried to explain perceptual organization in terms of some 114 laws, such as *good continuation*, *symmetry* and *closure* (Pomerantz and Kubovy, 1986; Wagemans et al., 2012) From this set, Boring extracted 14 condensed laws, including *naturalness of form* and *persistence of form* (Boring, 1942, pp. 253-254). It is a key to the Gestalt approach, however, that such laws grow out of the general principle of prägnanz. This asserts, roughly, that stimuli are organized in the way that most simplifies their global structure. The Gestalt approach has thus come to be associated with the proposition that *simplicity* is the governing principle of perceptual organization, i.e., that it is the simplest interpretation of a distal stimulus that is preferred (Hochberg and McAlister, 1953; Attneave, 1954).

Contrasting with this is the Helmholtzian view of perceptual organization. Stemming from the work of von Helmholtz (1860/1962), this also has complex and diverse origins. What is key for present purposes is the commitment made to the *likelihood principle* (Gregory, 1974; Brunswick, 1956). This states that

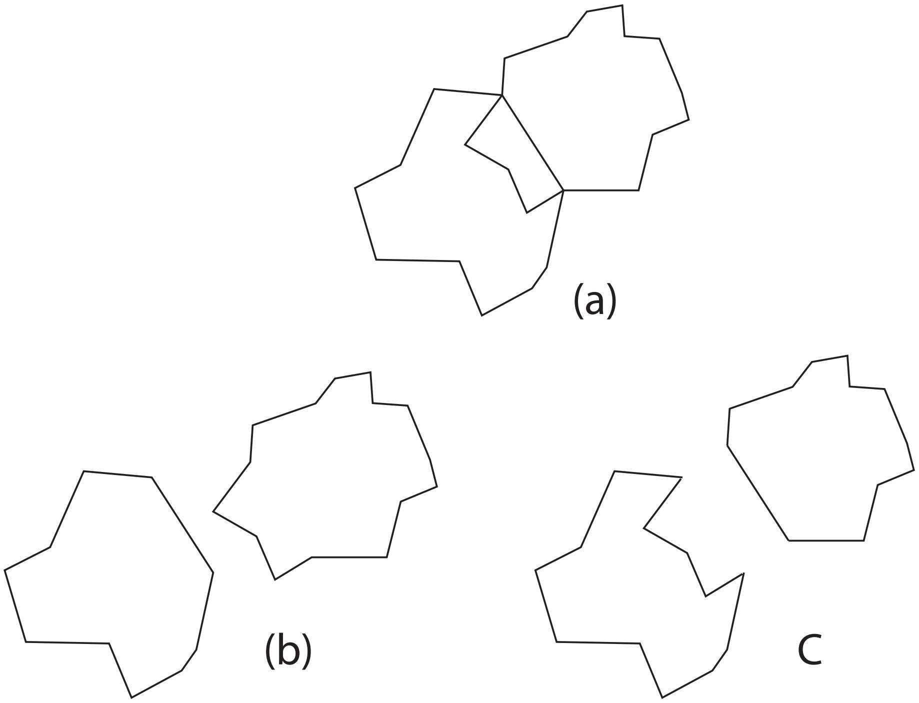

Figure 2: The pattern in (a) tends to be organized as two superimposed figures, as in (b), rather than (c).

'sensory elements will be organized into the most probable object or event (distal stimulus) in the environment consistent with the sensory data (the proximal stimulus)' (Pomerantz and Kubovy, 1986, p. 36-9). The claim, in this case, is that preference is given not to the simplest, but to the most likely interpretation of the distal stimulus.

To some degree, the attempt to explain perceptual organization has come to be framed as a debate between those who take perceptual organization to be governed by simplicity, and those who take it to be governed by likelihood (Hatfield and Epstein, 1985; Leeuwenberg and Boselie, 1988; van der Helm, 2000). But the possibility of these principles being intimately related is also considered plausible. Brunswick argues that they might co-exist (Brunswick, 1956). Mach proposes that the visual sense operates in conformity with both (Mach, 1906/1959). Attneave (1982) suggests they may be two sides of the same coin. Chater (1996) argues they are undifferentiable. He makes the point that a more probable perceptual representation can always use default assumptions to a greater degree, thereby increasing apparent simplicity. He also sets out a mathematical proof which demonstrates that simplicity and likelihood are formally equivalent.[1]

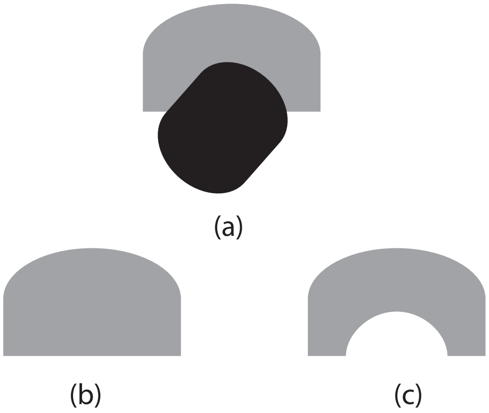

Figure 3: A perceptual preference relating to an occluded shape. The dark area in (a) tends to be perceived as a shape occluding (b) rather than (c).

If simplicity and likelihood are two sides of the same coin, the disagreement between the Gestalt and Helmholtzian positions looks set to dissolve. But there then arises a strong desire to know if the two principles can be reduced to one. Is a unification feasible? Applying Occam's Razor in this way seems particularly desirable given that, as things stand, an account that references one principle alone cannot be completely satisfactory. If it invokes the simplicity principle alone, it misses the point about likelihood. If it invokes the likelihood principle alone, it misses the point about simplicity.

The present paper takes the position that the two principles can indeed be reduced to one. The proposal, specifically, is that the equivalence of simplicity and likelihood can be explained in terms of *infotropism*, which is the label the present paper gives to information seeking. It is shown that processing which behaves in this way acts in conformity with both principles. Any explanation in terms of simplicity or likelihood can thus be reduced to an explanation in terms of infotropic processing.

The argument is developed over four main sections. The section immediately to follow (Section 2) introduces inverse information measurement, which is key for characterizing infotropic processing. Section 3 demonstrates the sense in which infotropic processing of representations has the effect of targeting both simplicity and likelihood. Section 4 discusses related work, and examines the

[1] Chater uses Kolmogorov complexity theory (Chaitin, 1966; Kolmogorov, 1965; Li and Vitanyi, 1997; Solomonoff, 1964) for this. The proof is not uncontroversial, however. In the view of Feldman (2009, p. 885), Chater shows in a 'a very general but also somewhat abstract sense' that the two principles are identical. But van der Helm argues the proof is invalid due to mistaking 'the Bayesian formulation of the Occamian simplicity principle for the Helmholtzian likelihood principle' (van der Helm, 2011, p. 340; see also van der Helm, 2000).

more general implications of the proposal. Information theory is the main reference throughout. But the information-theoretic concepts used are at the simpler end of the mathematical spectrum, and they are all explained informally within the text. A full understanding of Shannon information theory is not required. Two technical points are also worth mentioning at the outset. Many of the examples in the text cite information values. These are rounded to two decimal places where necessary. Logarithmic values are calculated to base 2 in all cases.

2 Inverse information measurement

It is almost a century since R. L. Hartley first proposed the principle of logarithmic measurement on which information theory is based. Hartley noted the informational value of an outcome drawn from a choice of N is naturally measured using a logarithm of N (Hartley, 1928). Assuming logs are calculated to an integer base, this is the number of digits (in that base) required to identify or communicate the outcome. Given use of base 2, information values can then be expressed in terms of binary digits or 'bits'. The informational value of an outcome drawn from a choice of N is said to be $\log_2 N$ bits. This is just the number of binary digits required to communicate a 1-in-N outcome. The metric can be applied to any situation that presents a well-defined choice of outcomes. There is no requirement for a communications context. Consider, for example, a party cracker that contains either a paper hat or a gift. Here, there are just two outcomes. Given $N = 2$, the informational value of an outcome is $\log_2 2 = 1$ bit. The outcome obtained by pulling the cracker has an informational value of 1 bit.

In the probabilistic generalization that Shannon developed twenty years after Hartley's publication, the information measure is redefined as $-\log_2 p$ where p is the probability of the outcome (Shannon, 1948; Shannon and Weaver, 1949). This accommodates the case where the outcome is drawn from a choice of N, since then $p = \frac{1}{N}$ and $-\log p = \log N$. But Shannon's formulation is more general, since it makes no assumptions about how probabilities are derived. Averaging is also facilitated. Given a distribution over outcomes, the average informational value of an outcome is just a weighted sum:

$$\sum_i (-\log_2 p_i) p_i$$

Noting that this formula also defines physical entropy, Shannon adopted the term 'entropy' for average information, and put the formula into its usual arrangement:

$$-\sum_i p_i \log_2 p_i$$

Average information is generally termed entropy as a result. However, where $-\log_2 p$ is called the 'surprisal' the term 'average surprisal' is a valid alternative (Tribus, 1961). In this paper, 'surprisal' is used throughout. But average information may be termed entropy or average surprisal depending on context.

The enormous impact that information theory has had on neuroscience and psychology stems from this formula. The cognitive interpretation that arises is seen as particularly significant. The amount of information that is expected with regard to some set of possible outcomes can be seen as the level of *uncertainty* that exists regarding those outcomes. Entropy can be seen as measuring uncertainty in this way. The effect can be illustrated using the party cracker example. Say initially we know that party crackers contain hats with probability 0.7. The average value of an outcome is $(-\log 0.7 \times 0.7) + (-\log 0.3 \times 0.3) =$ 0.88 bits.[2] We have 0.88 bits of uncertainty about about whether a cracker will contain a hat or a gift. Subsequently, we discover that crackers contain hats with probability 0.9. The average value of an outcome then decreases to $(-\log 0.9 \times 0.9) + (-\log 0.1 \times 0.1) = 0.49$ bits. Our level of uncertainty is reduced to 0.49 bits. If crackers are eventually discovered to contain hats with probability 1.0 (i.e., to never contain gifts), the entropy falls to zero bits, reflecting the completely elimination of uncertainty from the scenario.

Entropy measures uncertainty in this way, and for theorists interested in cognition, the relationship is of great interest. Especially of note are the implications for sensory processing. In general, the purposes of a sensory mechanism are better served if sensory data have *lower* entropy, since where there is less entropy, there must be less uncertainty about what is sensed. On this basis, the objective of a sensory mechanism can be framed as the attempt to minimize the entropy of sensory signals. This is the fundamental premise of the 'efficient coding' paradigm of neuroscience, advocates of which include Barlow (1959, 1985, 2001), Watanabe (1960, 1969), Atick (1992), Field (1994), Olshausen and Field (1996), Gross (2002) and Quiroga (2005).[3] But sensory processing is far from the only functionality that can be modeled this way. Researchers have found numerous ways in which knowledge construction, and cognitive function more generally, can be modeled as entropy minimization (e.g. Pearce et al., 2008; Pothos, 2010; Friston, 2010, 2013).

Entropy minimization is one type of 'information seeking'. By reducing uncertainty, we implicitly increase the information possessed in advance. But it is not the type with which the present paper is concerned. The present goal is to posit information seeking as the underlying principle of perceptual organization. This is problematic if we equate information-seeking with entropy minimization. In order to model functionality F as entropy minimization, we need to know the true probabilities of data used by F, and we need to be able to view F as a single act of minimization. In the case of perceptual organization, both requirements are difficult to satisfy. Nobody knows the true probabilities of the sensory signals underlying human perception; and it is hard to see how

[2] Real values are rounded to two decimal places throughout.

[3] The roots of the tradition predate the establishment of information theory, however (e.g. Craik, 1943). Theorists of the nineteenth Century, such as Mach (1896/1959), Pearson (1892) and von Helmholtz (1860/1962), also emphasized the way efficient coding of experience may underpin knowledge of the world.

perceptual organization could be accomplished by a single act of minimization.

Proceeding in this way is not an impossibility, however. Chater shows something along these lines as part of his proof that simplicity and likelihood are equivalent (Chater, 1996). In the argument made from information theory, Chater equates perceptual interpretations with optimal codes. This has the effect of making simplicity and likelihood equivalent by definition. The connection follows directly from the fact that optimally short codes are also maximally probable. But the probability distributions on which optimality is decided remain unspecified, and it is not obvious how a single entropy minimization could mediate the organizational process.[4]

Modeling perceptual organization directly as entropy minimization faces significant difficulties, then. But the situation for information seeking more generally is more complex. In the standard use of the Shannon paradigm, we measure the informational value of data (e.g., signals) given an applicable probability distribution. It is possible to invert the measure, however, so as to derive the informational value of a probability distribution given observed data. This inversion of the standard metric is what the present paper terms *inverse information measurement*. The method is of particular interest for present purposes as it offers a way of seeking information that avoids the need for entropy minimization.

Consider the cracker scenario again. Imagine a cracker that has a label specifying the chance of getting a gift versus the chance of getting a hat, e.g., '30% chance of a gift, 70% chance of a hat'. The real probabilities are unknown, but when the cracker is pulled, what emerges is a hat. What is the informational value of the suggested distribution in light of this outcome? To calculate the answer, we need to determine the informational value of the suggested distribution given the observed outcome.

The immediate obstacle for this form of measurement is that it cannot use the suggested distribution as a basis for calculating informational value. Being only 'suggested', the distribution does not define true probabilities. What can be used instead is the principle of maximum entropy (Jaynes, 1957).[5] This asserts that where there is no knowledge of a particular set of outcomes, a distribution of maximum entropy should be assumed. Bringing this principle into play, it is possible to calculate the expected informational value of a suggested distribution in the light of a particular outcome.

In the case of the envisaged cracker, the label says '30% chance of a hat, 70% chance of a gift'. The label divides probability between the observed outcome (a hat) and the unobserved outcome (a gift). Probability assigned to the observed outcome can also be seen as the probability of gaining the value of this outcome by correctly predicting it. The complementary probability can be seen as the probability of losing the value of the observed outcome by *failing* to predict it. Averaging these gains and losses then establishes the informational value of the

[4] Chater himself describes the proof from information theory as 'unsatisfactory in two ways' (Chater, 1996, p. 572). However, this is on the basis of a different analysis of the problems arising.

[5] This is sometimes called the principle of 'not being over-confident'.

distribution as a prediction of the observed outcome.

Given there are only two outcomes, the maximum entropy in this scenario is $\log_2 2 = 1$ bit. The informational value of the suggested distribution (P), given the observed outcome (x) is thus

$$I(P, x) = \sum_{y \in D} P(y) \begin{cases} 1 & \text{if } y = x \\ \\ -1 & \text{otherwise} \end{cases} \tag{1}$$

This defines the informational value of distribution P as a prediction of outcome x, given two possible outcomes. On this basis, the maximum evaluation is just the full value of an outcome (i.e., the maximum entropy). This is achieved if all probability is assigned to the correct outcome. The minimum evaluation is the negative of the maximum. This is achieved if no probability is assigned to the correct outcome. With N outcomes, it is then the case that

$$\log_2 N \geq I(P, x) \geq -\log_2 N$$

Using this way of dealing with the two-outcome scenario, it is possible to compare evaluations of Eq. 1 against intuitive judgements. Let the assumption continue to be that pulling the cracker produces a hat. This is a situation in which a label giving more probability to hats is most informative. But the degree of emphasis is relevant too. Given the outcome is a hat, a label specifying a 90% chance of a hat seems more informative than one specifying a 75% chance of a hat, which seems more informative than one specifying a 60% chance, and so on. A label giving a 50/50 chance of either outcome would seem to have no informational value whatsoever. But what of a label that gives a 75% chance of a gift? Given the outcome is a hat, this seems even less informative than the 50/50 label. It is positively *misinformative.* How can this be, if the value of the 50/50 label is itself zero?

Close examination of Eq. 1 resolves the puzzle. The formula faithfully tracks our informal judgements. Given the outcome is a hat, the '90% chance of a hat, 10% chance of a gift' label has a value of $0.9 - 0.1 = 0.8$ bits[6], whereas the '75% chance of a hat, 25% chance of a gift' label has the lower value of $0.75 - 0.25 = 0.5$ bits. The 50/50 label has a value of zero bits, as expected. The '25% chance of a hat, 75% chance of a gift' label, on the other hand, has a value of $0.25 - 0.75 = -0.5$ bits. Here, the value is less than zero due to being negative. Negative evaluations are obtained whenever probability is distributed away from the observed outcome. The expected loss then exceeds the expected gain, producing a net loss. If probability is distributed towards the observed outcome, the expected gain exceeds the expected loss, producing a net gain. The label's informational value can be either positive or negative in this way. In the two-outcome scenario, it increases linearly (from -1 to 1 bit) with the probability assigned to the correct outcome, as in the graph of Figure 4.

[6] Calculations in this simplified form are possible whenever the maximum entropy value is 1.

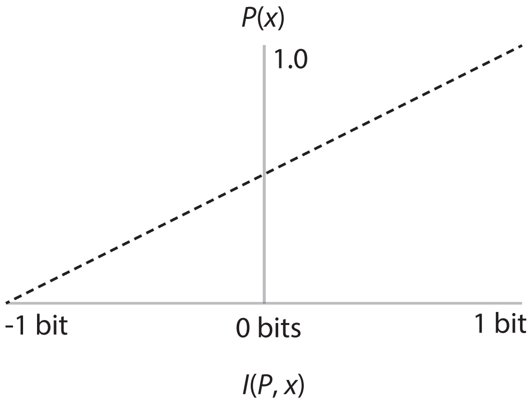

Figure 4: Inverse information measurement with two outcomes.

An obviously unsatisfactory feature of Eq. 1 is its assumption of there being two possible outcomes. In general, there may be any number. If there are more than two, the maximum entropy is greater than 1 bit, and the number of possible outcomes *not* arising in each case is also greater than 1. The definition needs to be restated to link the informational value of a choice to the appropriate entropy. We also need to ensure commensurability between gains and losses by normalizing the latter with respect to the total number of unobserved outcomes. These requirements are met in the following generalization:

$$I(P,x) = \sum_{y \in D} P(y) \begin{cases} \log_2 |D| & \text{if } x = y \\ -\frac{\log_2 |D|}{|D|-1} & \text{otherwise} \end{cases} \tag{2}$$

This defines the informational value of distribution P as a prediction of outcome x, where D is the set of possible outcomes. Notice that set D may be of any size, allowing for application to a choice with arbitrarily many outcomes. If D contains two elements, the formula reduces to Eq. 1. The informational value is $\log_2 2$ as before, and the normalization is just a redundant division by 1. If D contains more than two elements, the loss is then normalized with respect to the number of false outcomes, as required.

Eq. 2 is a completely general way of measuring the informational value of a distribution as a prediction of a particular outcome. It can be used to evaluate any distribution with respect to an observed outcome to which the distribution applies.[7] It is important to keep a clear distinction between entropy

[7] Consider the following illustrations: (1) A doctor gives a 75% chance of a test producing a negative result, a 15% chance of it producing a positive result, and a 10% chance of it producing a non-result. If the outcome is a positive result, the suggested distribution has the negative value of -0.435 bits, and the doctor's analysis is found to be positively misinformative; (2) A weather forecaster gives a 20% chance of sun, a 20% chance of rain, and a 60% chance

measurement and inverse information measurement, however. The former is the standard metric defined by Shannon (1948; Shannon and Weaver, 1949). This is the appropriate measure for any conventional setting, i.e., whenever interest focuses on the informational value of data. The latter is an inversion of the standard metric, of use in the special case where interest relates to the informational value of predictions.

The case where we have multiple observed outcomes can also be dealt with. In this situation, the measurement of informational value is arrived at by averaging. The simplest situation is where we know the true probabilities with which outcomes arise. We might know the true probabilities with which particular objects are put into crackers, for example. We can then calculate the average over all outcomes as an expectation:

$$I(P_s, P_r) = \sum_{x \in D} P_r(x) I(P_s, x) \tag{3}$$

Here, P_s is the suggested distribution, and P_r is the real distribution. The value of $I(P_s, P_r)$ is the average informational value of the distribution for predicting outcomes defined by P_r. The more likely scenario is where the true probabilities are unknown, but we have access to a set of observations. In this case, the mean value is obtained in the usual way, by averaging across observations.

3 Infotropism, simplicity and likelihood

Having taken the inverse method of information measurement into account, it is possible to look again at the issue of 'information seeking'. As noted, this process is conventionally identified with entropy reduction (e.g. Watanabe, 1969; Barlow, 1959; Friston, 2013). Entropy models have been developed in a broad range of areas (e.g. Pearce et al., 2008; Pothos, 2010; Friston, 2010). In light of the potential for inverse information measurement, a shift of perspective can be made. Instead of equating information seeking with entropy minimization, we can equate it with model selection. Adoption of a probabilistic model can be considered to gain information just in case its (inversely measured) informational value with regard to relevant data is greater than otherwise achieved. This is also 'information seeking'; but a single act of entropy minimization is no longer assumed.

What is relevant for present purposes is the scope this offers for new forms of explanation. There is the opportunity to classify a perceptual system in a novel way. Any perceptual system which updates representational choices in accordance with a preference for informational value can be said to be seeking

of snow. Given the outcome is snow, the value of the suggested distribution is 0.633 bits; the forecast is found to be moderately informative; (3) An estate agent gives a 60% chance of selling a property for more than the asking price, a 10% chance of selling for the asking price, a 20% chance of selling for less than the asking price, and a 10% chance of not selling at all. Given the outcome is a sale at less than the asking price, the value of the distribution is -0.133 bits; the agent's analysis is found to be misinformative.

information in this way. Entropy minimization is neither assumed nor implied. The present paper introduces *infotropism* as a label for information seeking of this sort. By extension, seeking information by means of inverse information measurement is termed *infotropic*. A perceptual system which updates representational choices in accordance with a preference for informational value is said to exhibit infotropic behavior.

Implications for perceptual organization can then be examined. It has been seen that processes of perceptual organization can be explained in terms of adherence to the simplicity principle or, equivalently, adherence to the likelihood principle. What can now be shown is that perceptual processing that is infotropic must subserve both principles. Where observed data are given relatively more probability, the informational value of the distribution in question is increased. Resolving a representational choice infotropically thus ensures that observed data acquire greater probability. Infotropic processing of an interpretation ensures observed data acquire the highest probability given available representational choices. This is the sense in which infotropic processing upholds the likelihood principle.

At the same time, the simplicity principle is accommodated. Under inverse information measurement, the value of a distribution in a particular situation depends on the degree to which probability is concentrated on the observed outcome(s). Infotropic processing of interpretive representations ensures that extremal distributions (i.e., ones with lower entropy) are preferred. Probabilistic representations accommodating relatively fewer outcomes are a natural consequence of infotropic processing.[8] This is the sense in which infotropic processing upholds the simplicity principle. A perceptual system that is infotropic in character will appear to choose representational states in accordance with a preference for both simplicity and likelihood.

This has the effect of unifying the two principles of perceptual organization, and explaining why they are equivalent. But it also leads to a new way of characterizing the experience of perceptual organization. Instead of seeing it as a process in which we unconsciously seek out the simplest (or most likely) interpretation, we can characterize it as one in which we seek an interpretation that is maximally informative of sensory data, given representational choices. What we are seeking, on this account, is an interpretation which makes existing representations maximally informative with respect to sensory data. By definition, this interpretation is both maximally simple and maximally likely.

A potential criticism of the account relates to its subjectivist interpretation of the likelihood principle. As in (Chater, 1996), the proposal assumes that the most likely interpretation of the distal scene must be one that is represented by the perceptual agent as most probable. It is sometimes argued that this is invalid. In particular, van der Helm argues 'the Helmholtzian likelihood principle is about unconscious inference and holds that the visual system chooses the interpretation which objectively is most likely to be true, that is, not that

[8] A distribution that accommodates relatively fewer outcomes is one which places negligible probability on relatively more.

it chooses the one which persons subjectively believe is most likely to be true' (van der Helm, 2011, p. 338). The present approach proceeds on the basis that, in the absence of magical powers of perception, a genuinely objective choice of this sort is a logical impossibility. A subjective interpretation of the likelihood principle is taken to be inevitable.[9]

4 Discussion

Given the degree to which the likelihood and simplicity principles have been pitted against each other in explanations of perceptual organization, their putative equivalence places theorists in an odd position. What were once seen as competing accounts must now be seen as explanations that are, strictly speaking, *both* false. More positively, we might say both are true up to a point. Either way, the situation calls for remedial action. There is a need to discover if the two principles can be reduced to one. Given the assumption of equivalence, it is natural to assume there must be some underlying principle out of which both principles grow. The critical question from the scientific point of view is: What is this principle?

The present paper has argued that the underlying principle is infotropism, where this is defined as information-seeking using inverse information measurement. On this account, perceptual organization seeks neither the simplest nor the most likely interpretation of the distal scene. Rather, it targets an interpretation that is maximally informative of sensory data, given representational choices. The effect is identical since this interpretation *is* both maximally simple and maximally likely. The proposal can be seen as reflecting the observation that simpler representations must award higher probability to represented data (e.g. Mackay, 2003). This is sometimes called the 'Bayesian Occam factor' (Duda et al., 2001; Tenenbaum and Griffiths, 2001; Feldman, 2009). In the closely related minimum-description-length (MDL) approach, Rissanen argues that under broad assumptions the least complex hypothesis has the highest Bayesian posterior (Rissanen, 1978).

The proposal is within the general category of theories that deploy a notion of information preference. A large number of these exist in a broad range of areas. But the notion used here is not the usual one. The present account relies on the idea of inverse information measurement, whereas models of this type generally invoke the entropy calculation. (They are often termed 'entropy models' as a result, e.g. Pothos, 2010). Under the standard assumption, information is seen to be sought by a single act of entropy reduction. On the proposed view, it is sought incrementally, by acts of representation change. As an illustration of the difference, consider Linsker's *infomax* principle (Linsker, 1988). This also uses a notion of information preference. The principle states

[9] The main objection that van der Helm raises against Chater's (1996) equivalence proof also revolves around this point. He argues that the proof is invalid due to mistaking 'the Bayesian formulation of the Occamian simplicity principle for the Helmholtzian likelihood principle' (van der Helm, 2011, p. 340; see also van der Helm, 2000).

that a function which maps a set of input values to a set of output values should be configured so as to maximize the mutual information between the two sets (Linsker, 1989). This is on the basis of entropy minimization using non-inverse information measurement, however. Infotropism and infomax differ in this way.

The present proposal also has connections with coding theory (Hochberg and McAlister, 1953; Restle, 1979; Simon, 1972), and with the approach of structural information theory which stems from it (Leeuwenberg, 1968, 1971; Buffart and Leeuwenberg, 1981; Leeuwenberg and Boselie, 1988; van der Helm and Leeuwenberg, 1996). These are brought to light by examining infotropic processing from the efficient-coding perspective. In this view, the objective of the process is to seek out the simplest probabilistic model (of observed data) that can be constructed in terms of representational choices. Since an outcome can be the way another choice is resolved, a representation of this form can have arbitrarily many levels of structure. The general effect is then not unlike what is envisaged in structural information theory, in which the objective is the simplest structural representation that can be assembled in terms of perceptually realistic codes.[10]

In coding theory, however, the driving force is the intrinsic desirability of representational simplicity rather than infotropism. As Buffart et al. put it, 'The law of simplicity, within coding theory, is this: The perceptual system ... will arrive at the interpretation having the lowest information load. This, in a natural sense, is the interpretation having the simplest code and can therefore be thought of as the simplest interpretation.' (Buffart et al., 1981, p. 250-251). Convergence towards the most likely perceptual interpretation is then explained as a side-effect: van der Helm (2000) uses the *precisal* concept for this. The precisal of a stimulus is characterized as its minimum description length C converted into artificial probability $p = 2^{-C}$. On this basis, simpler interpretations are more likely by definition. This then explains 'where the perceptual system gets its probabilities from' (van der Helm, 2000, p. 772).

Also of interest are connections with Feldman's maximum-depth framework (Feldman, 2009, 1997, 2003). Feldman focuses on representations in the form of category hierarchies. More specifically, he envisages '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' (Feldman, 2009, p. 875). Feldman notes that where perceptual organization is mediated by representations of this type, the best interpretation is obtained by 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' (Feldman, 2009, p. 875). Under reasonable assumptions,

[10] It is argued that structural information theory faces a considerable challenge in deciding what these are (Wagemans, 1999; Wagemans et al., 2012). Chater comments that one of the problems dogging approaches based on coding theory is that 'the predictions of the theory depend on the description language chosen, and [the fact] that there is no (direct) empirical means of deciding between putative languages' (Chater, 1999, p. 571). In a similar vein, Olivers et al. note that the 'fundamental issue in building any coding language is which regularities to include and which not to include' (2004, p. 244).

this choice is the one that maximizes likelihood. Utilization of classes with the lowest position in hierarchical structure is then found to be a way of maximizing both likelihood and simplicity.

The present proposal can be related to Feldman's framework in the following way. Feldman shows that an underlying simplicity objective can lead to the emergence of representations that favor likelihood. The implication is the familiar one—simplicity and likelihood are equivalent. But, here, simplicity and likelihood are connected via a notion of processing specifically. The relationship with the present proposal is made a little closer. For a perfect fit, we would need to see Feldman's maximum-depth procedure as infotropic processing. Is this possible? One way to conceive a hierarchy of model classes is as a structure of choices in which outcomes represent classes, and distributions define subsumption relations. In such a hierarchical structure, we would expect informational values to decrease as choices are resolved (infotropically) bottom-up, following presentation of primitive data. On this basis, the lowest outcomes in the hierarchy that subsume *all* relevant data will tend to have the highest informational values. Given the way we are viewing the hierarchy, this set is also the maximum-depth interpretation. The maximum-depth procedure can be seen as implicitly infotropic in this way, and the relationship with Feldman's framework is thus quite close. But the present proposal differs in having a specifically information-theoretic constitution, and in reducing simplicity and likelihood to a single principle.

The other main area of related work has already been mentioned. Neuroscientific work in the efficient coding paradigm holds information seeking to be fundamental (e.g. Barlow, 1959; Watanabe, 1960; Atick, 1992; Field, 1994; Gross, 2002; Quiroga et al., 2005). The objective in this paradigm is to explain sensory processing in terms of use (or development) of informationally efficient codes. This is not unlike the present objective. But again, there are differences. Theorists in the efficient-coding tradition base their work on surprisal evaluation exclusively. Inverse information measurement, and the possibilities of infotropic processing play no part. Perceptual organization also tends to be a peripheral concern.

More generally, the over-arching relevance of Occam's Razor should be acknowledged. This principle states that simpler models are preferable in general. A widely adopted heuristic of science, the principle has been intensively debated in the modern era (e.g. Thorburn, 1918; Blumer et al., 1987). Much of what has been written on the topic bears on the present proposal in some way, and an exhaustive review is out of the question. But what is envisaged here is not inconsistent with the general spirit of the principle. Acquiring knowledge of phenomena is assumed to entail developing simpler models. Targeting simplicity in modeling should have the effect of enhancing knowledge as a result. The present proposal casts this idea into an information-theoretic form. Targeting simpler models is understood to be the process of finding a more *informative* representation of data, given representational choices. The present proposal can be seen as giving Occam's Razor an information-theoretic foundation in this way.

5 Concluding comment

Viewing any complex scene, we effortlessly arrive at an interpretation which resolves ambiguities in a particular way. There are then two questions to be answered: (1) Why is this particular interpretation chosen in this particular case? (2) By what mechanism is the interpretation obtained? A natural way of proceeding is to look for an account that explains the effect in perceptual terms. Taking the Helmholtzian path, we say that we arrive at a particular interpretation because we perceive it to be the simplest. Taking the Gestalt path, we say that we arrive at the interpretation because we perceive it to be the most likely. The drawback with both explanations is that they answer one question while begging the other. One perceptual result is explained in terms of another, but without light being shed on the mechanisms involved.

The present proposal pursues a lower-level approach. It offers an account of perceptual organization that is expressed in terms that are informational rather than perceptual. The perceptual disposition towards simpler/more likely interpretations is seen to result from infotropic processing of representations. This leads to a new way of characterizing the experience of perceptual organization. The interpretation we adopt in a particular case is recognized to be, neither the simplest that can be obtained, nor the most likely. Rather it is the interpretation that is most informative of sensory data, given representational choices.

References

Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *Network*, *3* (pp. 213-251).

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

Attneave, E. (1982). Pragnanz and Soap Bubble Systems. In Beck (Ed.), *Organization and Representation in Perception* (pp. 11-29), Hillsdale, NJ: Erlbaum.

Barlow, H. B. (1959). Sensory mechanisms, the reduction of redundancy, and intelligence. *The Mechanisation of Thought Processes* (pp. 535-539), London: Her Majesty's Stationery Office.

Barlow, H. B. (1985). Perception: What Quantitative Laws Govern the Acquisition of Knowledge from Senses? In Coen (Ed.), *Functions of the Brain*, Oxford: Clarendon.

Barlow, H. (2001). The Exploitation of Regularities in the Environment by the Brain. *Behav. Brain. Sci*, *24* (pp. 602607).

Blumer, A., Ehrenfeucht, A., Haussler, D. and Warmuth, M. (1987). Occam's Razor. *Information Processing Letters*, *24* (pp. 377-380).

Boring, E. G. (1942). *Sensation and perception in the history of experimental psychology*, New York: Appleton-Century-Crofts.

Brunswick, E. (1956). *Perception and the representative design of psychological experiments (2nd ed.)*, Berkeley: University of California Press.

Buffart, H. and Leeuwenberg, E. (1981). Structural Information Theory. In Geissler, Leeuwenberg, Link and Sarris (Eds.), *Modern Issues in Perception*, Berlin: Erlbaum.

Buffart, H., Leeuwenberg, E. and Restle, F. (1981). Coding Theory of Visual Pattern Completion. *Journal of Experimental Psychology: Human Perception and Performance*, *7* (pp. 241-274).

Chaitin, G. J. (1966). On the Length of Programs for Computing Finite Binary Sequences. *Journal of The Association of Computing Machinery*, *13* (pp. 547-569).

Chater, N. (1996). Reconciling Simplicity and Likelihood Principles in Perceptual Organization. *Psychological Review*, *103* (pp. 566-591).

Chater, N. (1999). The Search for Simplicity: A Fundamental Cognitive Principle? *The Quarterly Journal of Experimental Psychology,*, *52A*, No. 2 (pp. 273-302).

Craik, K. J. W. (1943). *The Nature of Explanation*, Cambridge: Cambridge University Press.

Duda, R. O., Hart, P. E. and Stork, D. G. (2001). *Pattern Classification*, New York, NY: Wiley.

Feldman, J. (1997). The Structure of Perceptual Categories. *Journal of Mathematical Psychology*, *41* (pp. 145170).

Feldman, J. (2003). The Simplicity Principle in Human Concept Learning. *Current Directions in Psychological Science*, *6* (pp. 227-232).

Feldman, J. (2009). Bayes and the Simplicity Principle in Perception. *Psychological Review*, *116*, No. 4 (pp. 875-887).

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation*, *6* (pp. 559-601).

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*, No. 2 (pp. 127-138).

Friston, K. (2013). Active inference and free energy. *Behavioral and Brain Sciences*, *36* (pp. 212-213).

Gregory, R. L. (1974). Choosing a Paradigm for Perception. In Cartarette and Friedman (Eds.), *Handbook of Perception. Volume 1: Historical and Philosophical Roots of Perception*, New York: Academic.

Gross, C. G. (2002). Genealogy of the “grandmother cell”. *Neuroscientist*, *8*, No. 5 (pp. 512-8).

Hartley, R. L. (1928). Transmission of Information. *Bell System Technical Journal* (pp. 535).

Hatfield, G. and Epstein, W. (1985). The Status of the Minimum Principle in the Theoretical Analysis of Visual Perception. *Psychological Bulletin*, *97* (pp. 155186).

Hochberg, J. E. and McAlister, E. (1953). A Qualitative Approach to Figural Goodness. *Journal of Experimental Psychology*, *76* (pp. 560-576).

Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Physical Review (Series II)*, *106*, No. 4 (pp. 620-630).

Kolmogorov, A. (1965). Three Approaches to the Quantitative Definition of Information. *Prob. Inf. Trans*, *1* (pp. 1-7).

Leeuwenberg, E. L. J. and Boselie, F. (1988). Against the Likelihood Principle in Visual Form Perception. *Psychological Review*, *95* (pp. 485491).

Leeuwenberg, E. L. J. (1968). *Structural Information of Visual Patterns: An Efficient Coding System in Perception*, The Hague: Mouton.

Leeuwenberg, E. L. J. A. (1971). A Perceptual Coding Language for Visual and Auditory Patterns. *American Journal of Psychology*, *84* (pp. 307-349).

Li, M. and Vitanyi, P. (1997). *An Introduction to Kolmogorov Complexity and Its Applications: Second Edition*, New York: Springer-Verlag.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, *21* (pp. 105-17).

Linsker, R. (1989). An Application of the Principle of Maximum Information Preservation to Linear Systems. In Touretzky (Ed.), *Avances in Neural Information Processing Systems vol 1* (pp. 186-94), San Mateo, CA: Morgan Kaufman.

Mach, E. (1896/1959). *The Analysis of Sensations, and the Relation of the Physical to the Psychical (Translation of the 1st, revised from the 5th, German Edition by S. Waterlow)*, New York: Dover.

Mach, E. (1906/1959). *The Analysis of Sensations and the Relation of the Physical to the Psychical*, New York: Dover Pupblications.

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

Olivers, C. N. L., Chater, N. and Watson, D. G. (2004). Holography Does Not Account for Goodness: A Critique of van der Helm and Leeuwenberg (1996). *Psychological Review*, *111*, No. 1 (pp. 242-260).

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

Pearce, M., Mullensiefen, D. and Wiggins, G. (2008). *An Information-dynamic Model of Melodic Segmentation*, Helsinki, Finland: International Workshop on Music and Machine Learning.

Pearson, K. (1892). *The Grammar of Science*, London: Walter Scott.

Pomerantz, J. R. and Kubovy, M. (1986). Theoretical Approaches to Perceptual organization: Simplicity and Likelihood principles. In Boff, Kaufman and Thomas (Eds.), *Handbook of Perception and Human Performance: Volume II Cognitive Processes and Performance* (pp. 36:1-45), New York: Wiley.

Pothos, E. M. (2010). An Entropy Model for Artificial Grammarl Learning. *Frontiers in Psychology*, *1*, No. 16.

Quiroga, R., Reddy, L., Kreiman, G., Koch, C. and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435* (pp. 1102-1107), 7045.

Restle, E. (1979). Coding Theory of the Perception of Motion Configurations. *Psychological Review*, *86* (pp. 1-24).

Rissanen, J. (1978). Modeling by the Shortest Data Description. *Automatica*, *14* (pp. 465-471).

Shannon, C. and Weaver, W. (1949). *The Mathematical Theory of Communication*, Urbana, Illinois: University of Illinois Press.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27* (pp. 379-423 and 623-656).

Simon, H. A. (1972). Complexity and the Representation of Patterned Sequences of Symbols. *Psychological Review*, *79* (pp. 369-382).

Solomonoff, R. (1964). A Formal Theory of Inductive Inference, Parts 1 and 2. *Information and Control*, *7*, No. 1 (pp. 1-22, 224-254).

Tenenbaum, J. and Griffiths, T. L. (2001). Generalization, Similarity, and Bayesian Inference. *Behavioural and Brain Sciencies*, *24* (pp. 629-641).

Thorburn, W. (1918). The Myth of Occam's Razor. *Mind*, *27* (pp. 345-353).

Tribus, M. (1961). *Thermodynamics and thermostatics: An introduction to energy, information and states of matter, with engineering applications*, D. Van Nostrand.

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., van der Helm, P. A. and van Leeuwen, C. (2012). *A Century of Gestalt Psychology in Visual Perception: II Conceptual and Theoretical Foundations*, Psychological Bulletin.

Wagemans, J. (1999). Toward a Better Approach to Goodness: Comments on van der Helm and Leeuwenberg (1996). *Psychological Review*, *106* (pp. 610621).

Watanabe, S. (1960). Information-theoretical aspects of Inductive and Deductive Inference. *I.B.M. Journal of Research and Development*, *4* (pp. 208-231).

Watanabe, S. (1969). *Knowing and Guessing: A Quantitative Study of Inference and Information*, New York: Wiley.

van der Helm, P. A. and Leeuwenberg, E. L. J. (1996). Goodness of Visual Regularities: A Nontransformational Approach. *Psychological Review*, *103* (pp. 429456).

van der Helm, P. A. (2000). Simplicity Versus Likelihood in Visual Perception: From Surprisals to Precisals. *Psychological Bulletin*, *126*, No. 5 (pp. 770-800).

van der Helm, P. A. (2011). Bayesian Confusions surrounding Simplicity and Likelihood in Perceptual Organization. *Acta Psychologica*, *138* (pp. 337-346).

von Helmholtz, H. (1860/1962). In Southall (Ed.), *Handbuch der physiologischen Optik, vol. 3*, Dover.