Predictive processing simplified: The infotropic machine (Penultimate draft)

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

April 7, 2016

Abstract

On a traditional view of cognition, we see the agent acquiring stimuli, interpreting these in some way, and producing behavior in response. An increasingly popular alternative is the predictive processing framework. This sees the agent as continually generating predictions about the world, and responding productively to any errors made. Partly because of its heritage in the Bayesian brain theory, predictive processing has generally been seen as an inherently Bayesian process. The 'hierarchical prediction machine' which mediates it is envisaged to be a specifically Bayesian device. But as this paper shows, a specification for this machine can also be derived directly from information theory, using the metric of predictive payoff as an organizing concept. Hierarchical prediction machines can be built along purely information-theoretic lines, without referencing Bayesian theory in any way; this simplifies the account to some degree. The present paper describes what is involved and presents a series of working models. An experiment involving the conversion of a Braitenberg vehicle to use a controller of this type is also described.

Keywords: predictive processing, predictive coding, hierarchical prediction machine, Bayesian brain, information theory, cognitive informatics

1 Introduction

On a traditional view of cognition, we see the agent acquiring stimuli, interpreting these in some way, and producing behavior in response. An increasingly

popular alternative is the predictive processing (also known as predictive coding) framework. This sees the agent as continually generating predictions about the world, and responding productively to any errors made (Rao and Ballard, 1999; Lee and Mumford, 2003; Rao and Ballard, 2004; Knill and Pouget, 2004; Friston, 2005; Hohwy et al., 2008; Jehee and Ballard, 2009; Friston, 2010; Huang and Rao, 2011; Brown et al., 2011; Clark, 2016). Clark characterizes this as 'the emerging unifying vision of the brain as an organ of prediction using a hierarchy of generative models' (Clark, 2013a, p. 185).[1] Granting that we can view actions as predictions put into a behavioral form, the proposal has the effect of unifying interpretive and behavioral functionality (Brown et al., 2011; Friston et al., 2009).[2] The model is also well positioned to use information theory (Shannon, 1948; Shannon and Weaver, 1949) as a way of explaining what is achieved. By improving performance in prediction, the agent renders the world less surprising, effectively gaining information (Cover and Thomas, 2006; Friston et al., 2012). The process can be seen as characteristically infotropic in this way (Thornton, 2014).

Partly because of its heritage in the Bayesian brain theory (Doya, 2007), predictive processing has generally been seen as an inherently Bayesian process. The 'hierarchical prediction machine' that mediates it is seen to be a specifically Bayesian mechanism. Processing is considered to be accomplished by inferential calculations. Backwards inference (i.e., application of Bayes' rule) is seen to be the means by which probabilities travel 'up' hierarchical structures, and forwards inference is the means by which they travel 'down.' Out of this bidirectional process, all functionalities of the brain are assumed to grow,[3] with the predictions of the machine being encapsulated in the conditional probabilities that connect one level of the hierarchy to another.

What the present paper draws attention to is an alternative way of specifying a machine of this type. In addition to the Bayesian formulation, there is an information-theoretic model, which is simpler in some respects. Key to this alternative is the metric of predictive payoff. Using basic principles of information theory, it is possible to measure the informational value of a prediction, provided we know the value of the outcome predicted and whether or nor it occurs. We can measure the informational 'payoff' with respect to an event of known value. This metric then gives rise to a way of building prediction machines. Any network of inter-predicting outcomes in which evaluations are kept up-to-date propagates information between outcomes in a machine-like way. The general

[1] The claim is part of a tradition emphasizing the role of prediction in perception and cognition, however (e.g. von Helmholtz, 1860/1962; James, 1890/1950; Tolman, 1948; Lashley, 1951; Mackay, 1956).

[2] The assumption underlying this is that 'the best ways of interpreting incoming information via perception, are deeply the same as the best ways of controlling outgoing information via motor action' (Eliasmith, 2007, p. 7).

[3] The 'pulling down' of priors is considered particularly significant (Hohwy, 2013, p. 33). As Clark comments, 'The beauty of the bidirectional hierarchical structure is that it allows the system to infer its own priors (the prior beliefs essential to the guessing routines) as it goes along. It does this by using its best current model—at one level—as the source of the priors for the level below' (Clark, 2013a, p. 3).

effect is that the machine transitions towards informational value. The network behaves infotropically, in a way that replicates the inferential activities of a Bayesian hierarchical prediction machine. The idea of predictive processing can thus be framed in a purely information-theoretic way, without using Bayesian theory.

The remainder of the paper sets out this alternative formulation in detail. Section 2 introduces the metric of predictive payoff, and examines its relationship to other measures from the Shannon framework. Section 3 shows how the metric provides the basis for building an information-theoretic version of the hierarchical prediction machine. Section 4 then demonstrates the behavior of some sample machines, including one deployed as the control system for a Braitenberg robot. Section 6 discusses neurophysiological issues, and Section 7 offers some concluding remarks.

2 Measuring predictive payoff

The theoretical foundation for the present proposal is Shannon information theory (Shannon, 1948; Shannon and Weaver, 1949). At the heart of this framework is the observation that certain events are well-behaved from the informational point of view. Given a strict choice of outcomes (i.e., a set of events out of which precisely one occurs), the informational value of the outcome that does occur can be defined as

$$-\log p(x)$$

where x is the outcome in question, and $p(x)$ is its probability. As Shannon notes, measuring the value in this way can be justified on a number of grounds. For one thing, it ensures that more improbable outcomes have higher informational value, as intuition suggests they must. For another, the value then corresponds to the quantity of data needed to signal the outcome. If we take logs to base 2 and round the value up to an integer, it is also the number of binary digits needed to signal what occurs.[4] For this reason, the value is often said to be measured in 'bits' (a contraction of BInary digiTS).[5] More formally, the quantity is termed the *surprisal* of the outcome (Tribus, 1961). Weather events are a convenient way to illustrate use of the measure. If everyday it rains with probability 0.25, but is fine otherwise, the informational value of the outcome of rain is $-\log_2 0.25 = 2$ bits.

Given this way of measuring the informational value of individual outcomes, it is straightforward to derive an average. Assuming we know the probability for all outcomes within the choice, the average information gained from discovering the result is

[4] For example, if event x has probability 0.25, we expect it to be drawn from a choice of $\frac{1}{0.25} = 4$ alternatives, for which we will need $-\log_2 0.25 = 2$ binary digits to signal the outcome.

[5] The term is original due to John Tukey.

$$-\sum_x p(x) \log_2 p(x)$$

This formula defines the information gained on average from discovering the outcome. We can also see it as the information that is *expected* to be gained from discovering the outcome. More generally, we can see the quantity as the uncertainty that exists with respect to the choice. Shannon notes this average plays an important role in statistical mechanics, where it is termed entropy. Accordingly, Shannon uses the term entropy as a description. Average information may thus be termed entropy, expected surprisal, average surprisal, expected information or uncertainty (Cover and Thomas, 2006; Mackay, 2003).[6] The weather illustration can be extended to show how entropy measurement is applied: if everyday it rains with probability 0.2, snows with probability 0.1, and is fine otherwise, the average informational value of an outcome is

$$-(0.2 \log_2 0.2 + 0.1 \log_2 0.1 + 0.7 \log_2 0.7) \approx 1.15 \text{ bits}$$

One difficulty with the framework is the status of the probabilities taken into account. Whether they are objective (defined by the world), or subjective (defined by a model possessed by the observer) is not specified.[7] In practice, either interpretation can be applied, and theorists tend to adopt whichever is appropriate for their purposes. Where entropy is seen as quantifying uncertainty, probabilities are likely to be seen as subjective. Where the formula is seen as quantifying generated information, they are likely to be seen as objective.[8]

Problems then arise if there is any difference between the two distributions. To give a concrete example, imagine that every day it rains with probability 0.2, but that an observer predicts rain with probability 0.4. The observer's prediction gives rain a higher probability than it really has. Plugging the objective probability into the formula, we find that the outcome generates a little over 0.7 bits of information. Using the subjective probability, the figure is nearly 1 bit. Without a distinction being made between subjective and objective probabilities, the evaluation is ambiguous.

One way of dealing with this situation is simply to disallow it. The position can be taken that the Shannon framework does not accommodate any deviation between subjective and objective probabilities. More productively, we can view the subjective distribution as a predictive model. On this basis, the predictions that arise can be seen (and evaluated) as ways of acquiring the informational

[6] In developing the framework, Shannon was particularly concerned with problems of telecommunication (Shannon, 1956). Events are conceptualized as messages sent from a sender to a receiver by means of a communication channel. Theoretical results of the framework then relate to fundamental limits on channel capacity, and the way statistical noise can be eliminated by introduction of redundancy.

[7] The present paper makes no distinction between a subjective probability and a Bayesian 'degree of belief'; whether there is a valid distinction to be made is unclear (cf. Ramsay, 1990).

[8] For example, for purposes of analyzing perceptual organization, van der Helm (2011) takes probabilities to be inherently objective. For purposes of analyzing musical creativity, Temperley (2007) takes them to be inherently subjective.

value of an outcome *before* it occurs. The calculation is made as follows. A predictive model must give rise to particular predictions. Given the informational value of a correct prediction must be the informational value of the correctly predicted outcome, we can calculate the expected informational value of predictions with respect to an outcome that does occur. We can find out, in other words, how much of the outcome's informational value is obtained in advance, by application of the predictive model.

Consider the following case. Imagine we are dealing with a choice of two outcomes, α and β. Let α' denote a prediction of outcome α, and β' a prediction of β. If the two events are objectively equiprobable, the informational value of each is $-\log_2 \frac{1}{2} = 1$ bit. If the predictive model gives rise to α' alone, and α is the outcome, we then have

$$I(\alpha') = I(\alpha) = 1 \text{ bit}$$

The value of the predictive model is 1 bit. Similarly, if the model gives rise to β' and β is the outcome, we have

$$I(\beta') = I(\beta) = 1 \text{ bit}$$

Again the model is worth 1 bit. If the model gives rise to both predictions together, its informational value is zero by definition. Predicting both outcomes is equivalent to making no prediction at all—the prediction merely recapitulates the choice. Thus

$$I(\alpha' \text{ and } \beta') = 0$$

Since the informational value of predicting two events together must be equal to the summed value of predicting them individually, it follows that

$$I(\alpha') + I(\beta') = 0$$

From this we can deduce that

$$I(\alpha') = -I(\beta')$$

The informational value of a predicted outcome that does *not* occur is, in this case, the negative of the value it acquires if it does occur. If the two outcomes are objectively equiprobable, the evaluations are 1 and -1 bits respectively. The informational value of a predicted outcome—occurring or non-occurring—can then be defined as

$$I(x') = \begin{cases} 1 & \text{if } x \text{ occurs} \\ \\ -1 & \text{otherwise} \end{cases} \tag{1}$$

Given $I(x')$ is the informational value of the prediction of x, we can then calculate the expected value of a subjective distribution treated as a predictive model. If Q is the distribution in question, and $Q(x)$ is the asserted probability

of outcome x, $Q(x)$ can also be seen as the probability of the observer predicting outcome x. The expected informational value of model Q with respect to the outcome is then an average, in which the informational value of each predicted outcome is weighted by the probability of being predicted:

$$I(Q) = \sum_x Q(x)I(x') \tag{2}$$

This is the informational value of distribution Q (treated as a prediction generator) with respect to the outcome that occurs. It is the informational revenue acquired implicitly by the model, in advance of the outcome's occurrence. This quantity is termed *predictive payoff*.

Notice the maximum predictive payoff is what we would expect—it is the full informational value of the occurring event. This satisfies the requirement that a veridical prediction acquires the predicted event's value. Less intuitive is the minimum payoff: this is the corresponding negative. Unexpectedly, the payoff of a predictive model can be either a gain or loss.

The potential for a prediction to be loss-making seems counter-intuitive at first. The process has an unexpected sting in the tail. But this turns out to be perfectly in line with the way predictions are assessed in day-to-day life. Weather forecasts again offer a good illustration. Say that a forecast gives only a 20% chance (probability 0.2) of rain on a particular day, but that in the event it *does* rain. This forecast would be judged positively misinformative, i.e., worse than one giving a 50/50 chance of rain. The mathematical evaluation explains why. Assuming just two equiprobable outcomes, the payoff is

$$0.2 \times 1 + 0.8 \times -1 = -0.6 \text{ bits}$$

Given the actual outcome is rain, the forecast generates a loss of 0.6 bits. Examining alternative scenarios confirms that the calculation always reflects informativeness in this way. If the forecast gives a 70% chance of rain and there is rain, the judgement is 'fairly informative' and the payoff is a 0.4 bit gain. If there is rain following a forecast giving only a 10% chance, the judgement is 'highly misinformative' and the payoff is a 0.8 bit loss. If the forecast gives a 50% chance of rain, the judgement is 'completely uninformative' and the payoff is neither gain nor loss. Evaluations range from positive → zero → negative just as judgements range from 'informative' → 'uninformative' → 'misinformative.'[9]

2.1 The general measure

The definition of predictive payoff set out above deals with the simple case of a choice encompassing two equiprobable outcomes. In general, a choice can have any number of outcomes, and their probabilities may vary. We need to define

[9]Key to this match is isolation of the boundary between informative and misinformative cases. As Gibson notes, the 'line between the pickup of misinformation and the failure to pick up information is hard to draw' (Gibson, 1979, p. 244).

the informational value of events (occurring and non-occurring) in a way that takes account of this. The following revision achieves the required effect.

$$I(x') = \begin{cases} -\log_2 P(x) & \text{if } x = x^\star \\ \\ \frac{P(x)}{1-P(x^\star)} \log_2 P(x) & \text{otherwise} \end{cases} \tag{3}$$

Here, $P(x)$ denotes the objective probability of outcome x, and $x^\star$ is the outcome that occurs. In the case of there being two equiprobable outcomes, the formula reduces to Eq. 1. It allows there to be any number of outcomes, however, and for these to have varying probability. The definition uses the relative probability of a non-occurring event,

$$\frac{P(x)}{1 - P(x^\star)}$$

in order to normalize the non-occurring event's negative contribution. This ensures the summed negative contribution is equal to the positive contribution, guaranteeing that a predictive model which distributes probability uniformly over outcomes will have an informational value of zero, as logic requires.

Using this generalized metric, we can evaluate predictive payoff in arbitrarily complex situations. Examining cases more widely confirms that, regardless of the number of outcomes, informativeness and predictive payoff always remain in step. If a model has positive payoff, it is judged informative. If the payoff is negative, the model is judged misinformative. If the payoff is zero, the model is seen to be completely uninformative (i.e. neither informative nor misinformative).

Consider, for example, a doctor who gives a 75% chance of a certain test producing a negative result, a 15% chance of it producing a positive result, and a 10% chance of it producing no result. Here, the indicated probabilities are 0.75, 0.15 and 0.1 respectively. Assuming the objective probabilities are uniform, the informational value of an outcome is $\log_2 3 \approx 1.58$ bits. If the outcome is a positive result, the predictive payoff of the doctor's forecast is then

$$0.15 \times 1.58 - \frac{0.75}{0.85} \times 1.58 - \frac{0.1}{0.85} \times 1.58 \approx -0.43$$

Given the doctor's strong prediction of a negative result, the outcome of a *positive* result would lead us to judge the prediction as fairly misinformative. The evaluation corroborates the judgement. The prediction is found to produce a loss of around 0.43 bits in relation to the outcome.

Evaluation in the face of a four-way choice can also be illustrated. Imagine a housing agent who 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. The implied probabilities are then 0.6, 0.1, 0.2 and 0.1 respectively. Given the objective probabilities are uniform, the informational value of each outcome is $\log_2 4 = 2$ bits. In the case of a sale above the asking price, we would judge the

forecast to be fairly informative—this is the outcome most strongly predicted. Again, mathematical evaluation corroborates the judgement. Since

$$0.6 \times 2 - \frac{0.1}{0.4} \times 2 - \frac{0.2}{0.4} \times 2 - \frac{0.1}{0.4} \times 2 \approx 0.93$$

The predictive payoff is found to be a gain of approximately 0.93 bits.

2.2 Relation to KL-divergence and other metrics

How does predictive payoff fit into the Shannon framework more generally? What is the connection with quantities such as mutual information and conditional entropy? The simplest assessment is to say that this metric bears no relation to any existing measure. No existing measure allows a distinction to be made between objective and subjective probabilities. And none gives rise to negative values. Digging a little deeper, however, some general connections can be made.

While no existing constituent of the Shannon framework distinguishes subjective from objective probabilities, there are several which quantify relationships between probability distributions. These offer a way of accommodating the distinction. Mutual information is a case in point. This quantifies the informational relationship between two random variables, taking into account their individual distributions (Cover and Thomas, 2006). The measure quantifies how much one distribution tells us about the other. Unfortunately, it also references the joint distribution, which plays no part in the calculation of predictive payoff. This distribution is not assumed to be known. Mutual information and predictive payoff are incommensurable for this reason. The same applies to conditional entropy and cross-entropy. The former is defined in terms of a conditional distribution and the latter in terms of a set of observations. Neither figure in the calculation of predictive payoff.

One measure that can be compared is Kullback-Leibler (KL) divergence. This quantifies the relationship between two distributions without referring to any additional data. Given probability distributions P and Q, the KL divergence of P from Q is the information lost when Q is used to approximate P (Kullback and Leibler, 1951).[10] The KL divergence of distributions P and Q is

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_i \ln\left(\frac{P(i)}{Q(i)}\right) P(i)$$

Distributions that are identical have a KL divergence of zero, and the value rises to infinity as they diverge. A relationship with predictive payoff can then be discerned. Taking the true distribution to be one which places all probability on the occurring outcome, it will be seen that predictive payoff always decreases as the predicted and true distributions diverge. KL divergence varies in the opposite direction, decreasing as predictive quality increases.

[10] The measure has an intimate relationship with log loss. The log loss sustained by mispredicting a binary outcome is also the KL divergence of the suggested distribution from a distribution which gives all probability to the realized outcome (Mackay, 2003).

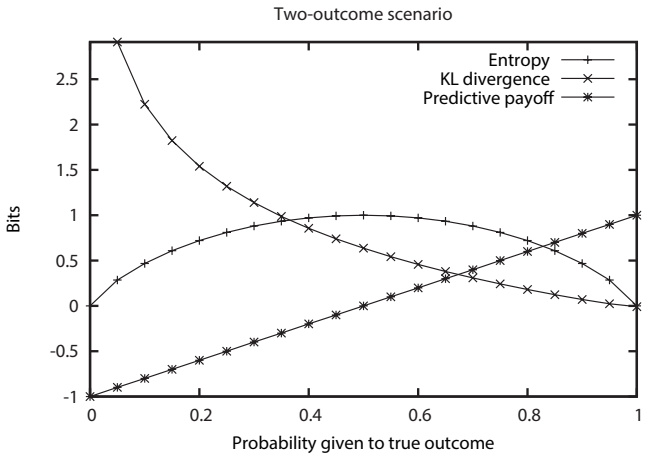


Figure 1: KL divergence v. predictive payoff.

Predictive payoff has as its maximum the informational value of the occurring event, and as its minimum the corresponding negative. A value of zero also identifies the special case of a neutral (uninformative) prediction. KL divergence maps this range into zero to infinity, and inverts it. For purposes of measuring predictive payoff, it has several drawbacks therefore. Values do not increase with predictive quality as we would expect. The qualitative distinction between good, bad and neutral predictions is not made. Critically, the quantity identified is *not* the informational payoff attained. A relationship exists, but there are significant differences. The graph of Figure 1 shows how values of KL divergence and predictive payoff compare in the two-outcome scenario. The entropy of the predicted distribution is also shown for comparison.

Beyond the Shannon framework, predictive payoff can be related to scoring functions in decision theory.[11] These are also a way of evaluating probabilistic forecasts, and they behave much like KL divergence. Consider the situation where a weather forecaster gives an 80% chance of rain, but there is no rain. Since predictive payoff reflects the probability given to the true outcome, its value in this case would be negative. We can also apply a scoring rule to evaluate the forecast with respect to this outcome. We might use the Brier rule (Brier, 1950) for example. This is defined by

[11] Potential links with psychological theories, such as (Dretske, 1981, 1983), are ignored, here, partly in recognition of the degree to which they have fallen out of favour in recent decades (Luce, 2003). As Haber noted some three decades ago, 'Ten years ago, I briefly reviewed the demise of information theory in psychology (R. N. Haber '74), concluding that it died for empirical reasons—it just did not work. Specifically, while it was generally easy to calculate the amount of information in a stimulus or in a response, such calculations did not correlate with any interesting or relevant behaviors of real perceivers, rememberers, or thinkers' (Haber, 1983, p. 71).

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_t - o_t)^2$$

where N is the number of forecasts made, p_t is the probability forecast at time t, and o_t is 1 if the forecasted outcome occurs and 0 otherwise. It will be seen that, in the case of a single forecast, the Brier score also varies monotonically with the probability given to the true outcome, with a score of 0 being awarded in the best case (all probability allocated to the outcome that occurs) and a score of 1 in the worst case (all probability given to the outcome that does not occur). Again, the effect is to map predictive payoff onto a non-negative, inversely varying quantity. Figure 2 shows how the two measures compare in the two-outcome scenario.[12]

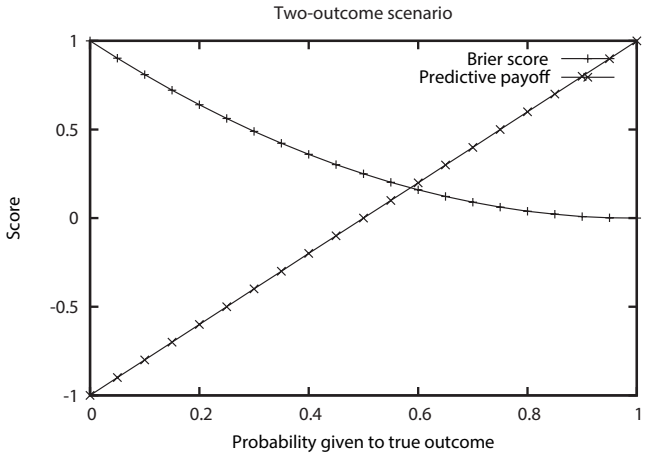


Figure 2: Brier score v. predictive payoff.

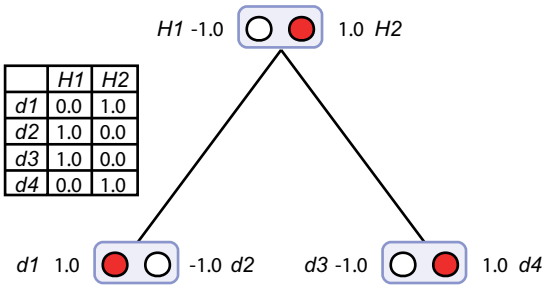
3 Infotropic machines

It is possible to return now to predictive processing, and to the 'hierarchical prediction machine' which is assumed to be its medium. The theoretical content of the previous section relates to this in a specific way. The mathematics of predictive payoff gives rise to a model of predictive processing that is independent of Bayesian theory. Given a network of outcomes in which connections are predictive probabilities, and evaluations are kept up-to-date, information given to one

[12] In practice, scoring functions are used to evaluate a series of forecasts with respect to an observed probability distribution over events. A scoring rule is termed 'proper' if it is maximized when forecasted probabilities are equal to true probabilities, and 'locally proper' if this maximum is unique.

outcome 'propagates' to others in accordance with predictive relationships (i.e., from predicted to predicting outcome, and vice versa). Outcomes that predict better acquire more informational value, ensuring information flows towards the most fecund sources of prediction. Inferences then arise implicitly. At the same time, a kind of error-correction is accomplished. Given outcomes are organized in choices, any outcome that predicts poorly acquires a negative value, implicitly 'switching' the host choice to an outcome that predicts better. The general effect is to replicate the behaviors of the hierarchical prediction machine in a way that meets the predictive-processing mandate.

In more detail the scheme is as follows. Let a predictive network be a set of outcomes in which any one outcome may designate a predictive model over others, and in which evaluations are always kept up-to-date (i.e., any evaluation that can be made, is made). A simple illustration is provided by Figure 3. This depicts a predictive network of six outcomes, organized into three choices. Each rounded rectangle represents a choice, with the enclosed circles being the choice's outcomes. Every choice has just two outcomes in this case. The value immediately adjacent to an outcome (e.g., 1.0) is the outcome's present informational value, and the label adjacent to that (e.g., *d1*) is the outcome's name. The table shows the predictive relationships that define the structure of the network, with predicted outcomes tabulated vertically, and predicting outcomes horizontally. The top two cells of the first column show the predictive model that outcome *H1* designates for the choice *d1*/*d2*. Notice this places all probability on *d2*. The outcome in each choice with the highest informational value (and the one to which the choice is implicitly resolved) is also filled for emphasis.



	H1	*H2*
d1	0.0	1.0
d2	1.0	0.0
d3	1.0	0.0
d4	0.0	1.0

Figure 3: A simple predictive network. See text for details.

Given that evaluations are kept up-to-date, a network of this form has dynamic behavior of a particular type. If we give an outcome the value it has when observed, any models which predict this outcome can then be re-evaluated. Other outcomes may then acquire new informational values, giving rise to further calculations in an ongoing way. Information given to one outcome propagates through the network, travelling towards relevant sources of prediction. But this processing is not under the control of any supervisory apparatus. The sole driving force is evaluation of predictive payoff. With evaluations made

whenever possible, the network becomes a kind of machine, whose behavior is an ongoing transition 'towards' informational value. A network of this type is an infotropic machine in this sense.

As an illustration of the processing that can be accomplished, consider the effect of cueing outcomes *d1* and *d4*. In this context, 'cueing' an outcome means distributing informational value in a way that implicitly resolves the host choice to the outcome in question. In this case, both choices have two outcomes. The effect is thus achieved by giving *d1* and *d4* the relevant positive value (1.0 bit), and *d3* and *d4* the corresponding negative (-1.0 bit). Following this input, there is the potential to evaluate all models that predict the cued outcomes. Given the (summed) predictive payoff for the models designated by *H1* and *H2*, the latter acquires a value 2.0 bits, and the former a value of -2.0 bits. The choice is implicitly resolved to *H2*. At this point, no further evaluations can be made and processing terminates.

The Bayesian interpretation of this processing can then be brought to light. What is accomplished is essentially two applications of Bayes' rule. More specifically, it is an act of maximum *a posteriori* (MAP) inference. Viewing choices as discrete variables, predicting outcomes as hypotheses, informational values as (unnormalized) priors, and predicted probabilities as conditional probabilities, the cueing of *H2* can be seen to select the conditioning state with the highest posterior. Implicitly, the machine performs MAP inference with respect to the observation of *d1* and *d4*. Given that the definition of predictive payoff is essentially Bayes' rule with information (rather than probability) used as currency, this result is not unexpected. With uniform priors, the machine identifies *H2* as the best predicting entity for the same reason MAP inference does: this is the outcome/hypothesis which places maximum probability on the observed data. The difference is that, with information used as a currency, the combined value of two predictions is a summation, not a multiplication. An infotropic machine of this type thus avoids the problem that plagues Bayesian calculation, namely the descent towards vanishingly small posteriors (Chater et al., 2006).[13]

This machine also replicates other characteristically Bayesian calculations. The schematics of Figure 5 give some account of the behaviors that can be elicited. Each of the four figures represents the state of the machine after a particular processing sequence. Probabilities and outcome labels are omitted to avoid clutter, but evaluations are shown. The diagram also uses arrows to show the patterns of propagation that arise. Schematic (A) shows the effects of cueing outcome *H2*. For both outcomes predicted by *H2*, it is then possible to derive an expected value, based on the informational value of *H2* and the predicted probability of the outcome. Information propagates top-down to *d1* and *d4*. These outcomes are cued in result. This replicates forwards Bayesian inference, i.e., derivation of conditioned probabilities from relevant conditionals and priors. The machine identifies *d1* and *d4* as the outcomes predicted by *H2*.[14]

[13] This is one of the reasons Knill and Pouget remark that 'unconstrained Bayesian inference is not a viable solution for computation in the brain' (Knill and Pouget, 2004, p. 718).

[14] Formally, top-down propagation is propagation from a predicting to a predicted outcome,

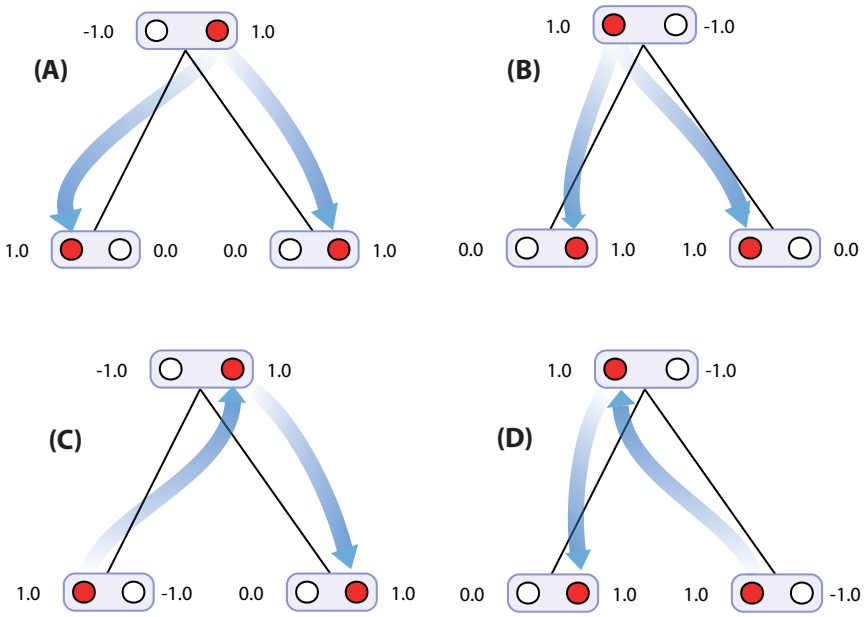


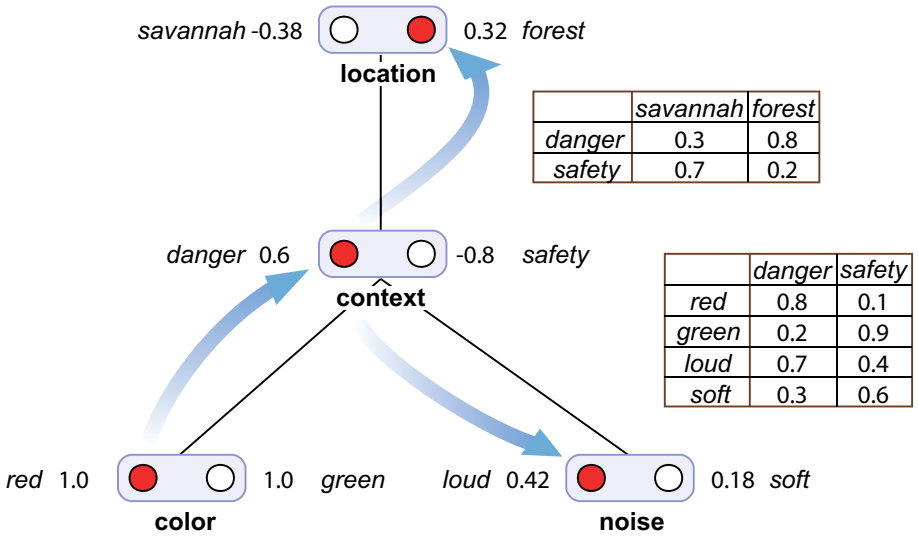
Figure 4: Propagational behaviors of a 3-choice infotropic machine.

Schematic (B) shows the complementary case involving the cueing of *H1*. The knock-on effect of this is to cue *d2* and *d3* rather than *d1* and *d4*. Schematic (C) shows a more complex behavior involving one phase of bottom-up propagation followed by one phase of top-down propagation. Initially *d1* is cued. Bottom-up propagation then cues *H2*, which gives rise to top-down propagation that cues *d4*. This can be thought of as replicating a combination of forwards and backwards inference. Alternatively, it can be thought of as implementing a kind of schema completion. Cueing *d1* cues the model that predicts *d1*, which has the effect of cueing the other outcome that this model predicts. Schematic (D) shows the corresponding case where *d3* is cued initially. The knock-on effect is to cue *H1*, and then *d2*.

What is found is that a suitably configured predictive network can reproduce the basic inferential operations we would expect of a hierarchical prediction machine. The predictive processing that is accomplished can be given a simplified, information-theoretic explanation accordingly. Rather than viewing the behavior as the work of a Bayesian inferential apparatus, it can be seen to grow out of information theory and the definition of predictive payoff. The Bayesian account of predictive processing can be reduced to an information-theoretic ac-

while bottom-up propagation is propagation from a predicted to a predicting outcome. However, since all diagrams situate predicting outcomes above predicted outcomes, the terms can also be interpreted as meaning 'downwards' and 'upwards'.

count, with fewer explanatory entities brought into play.



	savannah	forest
danger	0.3	0.8
safety	0.7	0.2

	danger	safety
red	0.8	0.1
green	0.2	0.9
loud	0.7	0.4
soft	0.3	0.6

Figure 5: Bidirectional propagations.

3.1 Hierarchical machines

The capacity of a predictive network to reproduce the behaviors of a genuinely *hierarchical* prediction machine—one with more than one level of representation—is particular of interest. It is behaviors of this sort that are considered most characteristic of predictive processing (Clark, 2016, 2013b). As an illustration of the effects that can arise, consider the network of Figure 5. This uses the schematic conventions of the previous diagram, but with choices given labels (e.g., **context**). All choices have two outcomes as before, but here they are named in a way that reflects implicational relationships between properties of a simple 'forest' environment. For example, one of the conditional probabilities encapsulated by the network is

$$P(danger \mid savannah) = 0.3$$

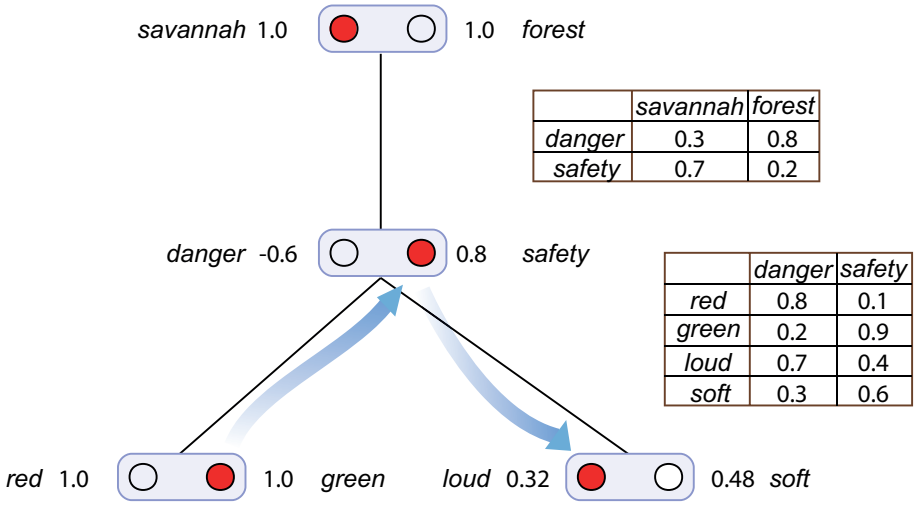
This can be seen as asserting that a situation of danger arises with probability 0.3, given a savannah location.[15] Similarly, the probability

$$P(red \mid danger) = 0.8$$

asserts that a red color arises with probability 0.8 given a situation of danger.

[15]Strictly speaking, what is encapsulated by the network is the probability of *danger* predicted by a model designated by *forest*.

The behaviors of this network can then be shown to reproduce the kinds of multi-level inference we expect a hierarchical prediction machine to deliver. Consider the state of Figure 5. This is the result reached after *red* is cued. The immediate effect of cueing this outcome is to enable the outcomes of the **context** choice to be evaluated. Information propagates bottom-up from **color** to **context**, giving *danger* and *safety* the values 0.6 and -0.8 bits respectively. The effect is to cue *danger*. Subsequently, outcomes in the other two choices can also be evaluated. In the case of **noise**, *loud* and *soft* are found to have the values 0.42 and 0.18 bits respectively. In the case of **location**, *savannah* and *forest* acquire the values the -0.38 and 0.32 bits respectively. The cued outcomes are then *loud* and *forest*. At this point, no source of information remains unacknowledged. No further evaluations can be made and processing terminates.



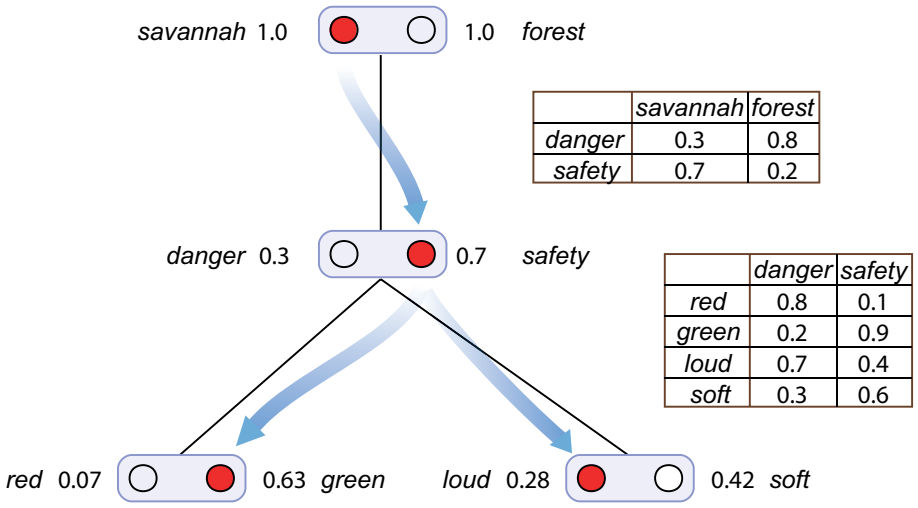
	savannah	forest
danger	0.3	0.8
safety	0.7	0.2

	danger	safety
red	0.8	0.1
green	0.2	0.9
loud	0.7	0.4
soft	0.3	0.6

Figure 6: Top-down/bottom-up conflict resolution.

A troubling situation arises if top-down and bottom-up propagations come together simultaneously at the same choice. In Bayesian terms, this is the problematic scenario where a 'pulled-up posterior' potentially contradicts a 'pulled-down prior.' An illustration using the machine of Figure 6 can be set up as follows. Assume that outcomes *green* and *savannah* are both cued. Outcomes of the **context** choice can then be evaluated either top-down (from *savannah*) or bottom-up (from *green*). In the latter case, the informational value of *safety* is found to be 0.8 bits; in the former it is 0.7 bits. The bottom-up propagation is thus preferable from the informational point of view. Subsequently, outcomes in the **noise** choice can be evaluated top-down, as illustrated in Figure 6. The values of *loud* and *soft* are found to be 0.32 and 0.48 bits respectively, making *soft* the cued outcome. This processing can be conceptualized as multi-step

inference. Alternatively, it can be seen to be a form of constraint satisfaction. What is obtained are the outcomes most consistent with the cues originally established.



	savannah	forest
danger	0.3	0.8
safety	0.7	0.2

	danger	safety
red	0.8	0.1
green	0.2	0.9
loud	0.7	0.4
soft	0.3	0.6

Figure 7: Top-down revision across two levels.

Given the predictive content of the hierarchy in this example, the results generated are most easily seen as inferences. But the quantitative aspect is also salient. Consider the case where *savannah* alone is cued (Figure 7). Revisions are then propagated top-down, cueing *green*. We can view this as the inference that a *green* color can be inferred from a *savannah* location. Contrast this with what happens when *safety* is cued. This also triggers top-down propagation, with *green* being the result. But in this case, the cued outcome is at a level immediately above that of *green*. The result is that *green* acquires the value of 0.9 bits, as opposed to the lower value of 0.63 bits that it acquires when *savannah* is cued.

Both *savannah* and *safety* produce *green* as an inference. But in the latter case, the inference is found to be have a higher informational value. The outcome *green* is more confidently inferred from *safety* than from *savannah*. This illustrates something resembling schema processing. Assertion of *safety* has the effect of invoking *savannah*, *green* and *soft*. If these four outcomes are viewed as comprising the *savannah* schema, the machine can be interpreted as having implicitly completed the schema in question. Given the way predictive relationships in this network represent implications in an idealized world, its behaviors can often be seen as inferential in character. But interpretations citing constraint satisfaction and schema completion are also possible.

3.2 Error correction

Inference is a vital element in the repertoire of the hierarchical prediction machine. The machine must be robust and consistent in the way it exploits the patterns of implication that can arise in a hierarchical structure of predictive models. Aside from this, the key functionality is *error-correction*. The machine needs to be able to correct what is predicted where this is found to be in error. This capacity is particularly important, as it is assumed to be the means of driving action. Bayesian theory itself provides no mechanism for correcting prediction error. Even in simple scenarios, the task is less than straightforward. But correcting predictions that stem from inferential processes operating in a multi-level hierarchy of predictive models is more challenging still. This is a credit-assignment problem of considerable complexity.

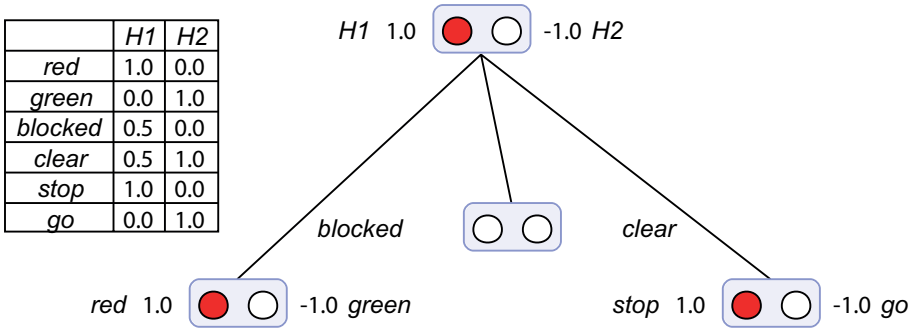


Figure 8: Action from error-correction in an infotropic machine.

This is where the infotropic approach pays an additional dividend. An active predictive network corrects its own errors automatically, without any additional mechanism being involved. If the predictions of a cued outcome are inconsistent with what is observed, evaluation will have the effect of cueing an outcome that does a better job. Informational evaluation *is* an error-correction process in this sense. A simple illustration of the effect is provided by the network of Figure 8. This can be viewed as a machine which, by process of error-correction, prompts an appropriate behavioral response in a simplified 'traffic light' scenario.

Outcome *H2* predicts *green*, *clear* and *go*, while outcome *H1* predicts *red* and *stop*. Assume *green*, *clear*, *red* and *blocked* are all perceptual outcomes, while *stop* and *go* are behavioral outcomes. The machine can be seen as a controller producing a *go* action if *green* and *clear* are perceived, and a *stop* otherwise. Imagine that *H2* is the presently cued outcome. This predicts and thus elicits a *go* action. As soon as *red* is perceived, the predictions stemming from *H2* are in error. The correction needed is then accomplished by process of evaluation. Evaluation of *H1/H2* awards greater value to *H1*. This is the outcome which predicts and thus cues a *stop*. 'Correction' of error gives rise to appropriate 'action' in this way. The capacity of the hierarchical prediction machine to

drive behavior through error-correction can also be reproduced by a predictive network.

3.3 Simulating a Braitenberg behavior

Use of error-correction to drive behavior can also be illustrated in a more concrete way. This subsection describes an experiment involving a robot taken from the Braitenberg collection, namely Vehicle 3a (Braitenberg, 1984, pp. 10-12). The behavior of this robot is 'slowing approach', i.e., moving towards a source of stimulation at decreasing speed, and stopping in front of it. This is the behavior Braitenberg characterizes as 'love' (Braitenberg, 1984, p. 10). The experiment carried out involved configuring the robot to use a predictive network as its controller. (A video of the behavior achieved is available at `www.sussex.ac.uk/Users/christ/demos/bpp.mp4`.)

Vehicle 3a is a two-wheeled robot equipped with two light sensors, as illustrated in Figure 9. The robot is represented by the large rectangular shape. The smaller rectangles represent the robot's two wheels, and the cup shapes represent the sensors. The circle represents the sole stimulus in the environment. This is assumed to be a light of some kind. In the original design, the robot used direct sensor-motor connections with inhibitory damping. Configured with 'uncrossed' connections, the robot was shown to move smoothly towards the sensed stimulus at decreasing speed, eventually halting in front of it.

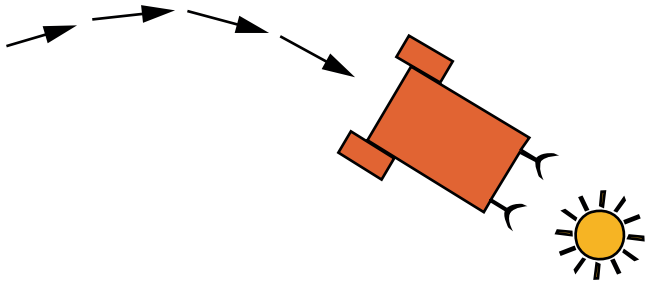
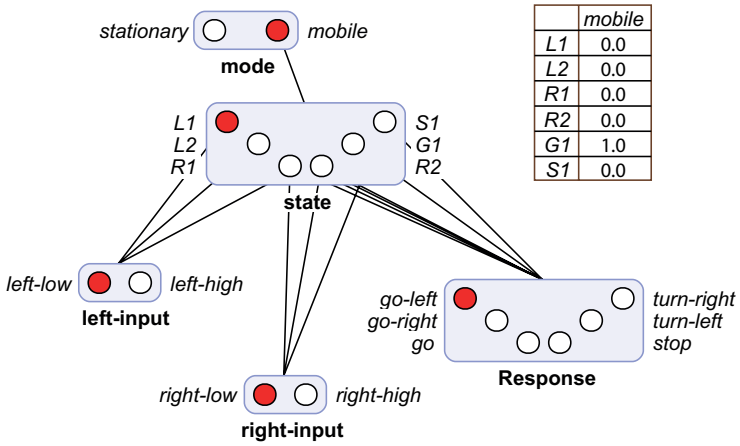


Figure 9: Vehicle 3a from the Braitenberg series (Braitenberg, 1984).

To illustrate the way implicit error-correction can drive behavior, the original control system of the robot was replaced with an infotropic machine, as illustrated in Figure 10. The general design of the machine was along the lines of Figure 8. Each sensor was configured to generate either a *high* or *low* outcome depending whether stimulation from the light source was high or low. Specifically, the machine was given a **left-input** choice, with outcomes *left-low* and *left-high*, and a **right-input** choice, with outcomes *right-low* and *right-high*. The output side of the machine then took the form of a **response** choice, in which the outcomes were five actions: *stop*, *go*, *turn-left*, *turn-right*, *go-left* and *go-right*.

To obtain the desired mapping from input to output, a **state** choice was introduced, with the outcomes *S1*, *L2*, *R1*, *S1*, *G1* and *R2*. Each of these was configured to predict a particular pattern of sensory input, and the corresponding behavioral response. Also added was a **mode** choice whose single outcome (*mobile*) predicts *G1* and thereby the *go* response. With this controlling apparatus in place, the robot was found to produce a reasonable approximation of the slowing-approach behavior. In the absence of sensory stimulation, the prediction of *G1* by *mobile* prompts a *go* action, causing the robot to move forward. Any sensory stimulation then has the effect of correcting the **state** choice (i.e., cueing a different outcome), prompting the appropriate *go* or *turn* action. On encountering high stimulation on both sensors at once, the cued outcome of the **state** choice reverts to *S1*, prompting a *stop* response.



	mobile
L1	0.0
L2	0.0
R1	0.0
R2	0.0
G1	1.0
S1	0.0

	L1	*L2*	*R1*	*R2*	*G1*	*S1*
left-low	1.0	0.0				0.0
left-high	0.0	1.0				1.0
go-left	1.0	0.0	0.0	0.0	0.0	0.0
go-right	0.0	0.0	1.0	0.0	0.0	0.0
go	0.0	0.0	0.0	0.0	1.0	0.0
stop	0.0	0.0	0.0	0.0	0.0	1.0
turn-left	0.0	1.0	0.0	0.0	0.0	0.0
turn-right	0.0	0.0	0.0	1.0	0.0	0.0
right-low			1.0	0.0		0.0
right-high			0.0	1.0		1.0

Figure 10: Infotropic controller for Vehicle 3a.

As the video (see `www.sussex.ac.uk/Users/christ/demos/bpp.mp4`) shows, the behavior of the rewired robot was somewhat less smooth than in Braitenberg's original example. The robot moves towards the stimulus following a zig-zag path. This is due to the fact that turns generally have the effect of eliminating

stimulus from the side which produced the response, while replacing it with stimulus on the side which did not. This problem can be addressed by use of a more fine-grained sensory-motor system, e.g., one accommodating *high*, *medium* and *low* sensory outcomes. By expanding the number of sensory/motor outcomes, an arbitrarily smooth reproduction of the original behavior can be obtained.

4 Discussion

The idea of predictive processing lies at the heart of what Clark describes as 'the emerging unifying vision of the brain as an organ of prediction using a hierarchy of generative models' (Clark, 2013a, p. 185). But the notion of the brain as a hierarchical prediction machine also has roots in the Bayesian brain theory—the claim that the brain computes by means of probability distributions (Doya, 2007).[16] This heritage is one reason theorists have generally conceptualized the hierarchical prediction machine in specifically Bayesian terms.

As has been seen, however, the Bayesian way of modeling the hierarchical prediction machine is not the only possibility. Taking the metric of predictive payoff into consideration, there is a purely information-theoretic alternative. A predictive network in which evaluations are kept up-to-date is a predictive mechanism with the functional characteristics we require. The specification is simpler in some respects, and has the advantage of making information the machine's internal currency, rather than probability. This avoids the need for a way of translating between probabilities and information-bearing signals. The prospect of the machine being a way of explaining 'perception and action and everything mental in between' (Hohwy, 2013, p. 1) is then made more compelling. Nevertheless, the infotropic prediction machine seems to be lacking some important features that the Bayesian version possesses. The present section works though a selection of these seemingly missing features, examining how, where and why they are implicitly realized.

4.1 Where are the precisions?

A natural place to begin the review is with 'precisions'. These are generally considered an important part of the predictive-processing mechanism. In Hohwy's view 'assessments of prediction error are hostage to how confident the system is that a prediction error is genuine and something worth minimizing' (Hohwy, 2013, p. 65). From the formal point of view, precisions are considered to be expected uncertainties (Friston et al., 2013; Yu and Dayan, 2005). But what does this mean in practice? One proposal states that precisions are implemented by regulating the *gain* on prediction error (e.g. Feldman and Friston, 2010). On this basis, a prediction with a higher precision has a higher gain, with the result

[16] More specifically, the assumption is that the brain represents 'information probabilistically, by coding and computing with probability density functions, or approximations to probability density functions' (Knill and Pouget, 2004, p 713).

that it is given more weight during processing. Each prediction error is endowed with an independent rating that establishes the confidence with which the error is obtained. This then dictates the level of emphasis given to minimizing the error.

The infotropic interpretation of predictive processing makes no mention of precisions. In this account, the need for an independent confidence-rating mechanism is eliminated. Prediction error is measured in informational units, a currency that is itself a measure of certainty. When prediction error is calculated this way, the values obtained are confidence ratings implicitly. A prediction error calculated with zero confidence translates into an informational value of zero bits, for example. The infotropic interpretation of predictive processing combines precision and prediction-error in a single measurement.

This highlights the way in which precision and uncertainty are connected. A mechanism able to mediate assertion of precision is also one which can handle acknowledgement of uncertainty. The need for the latter has long been recognized by theorists in the Bayesian brain tradition. As Knill and Pouget (2004, p. 718) comment, 'The real test of the Bayesian coding hypothesis is in whether the neural computations that result in perceptual judgements or motor behaviour take into account the uncertainty available at each stage of the processing.' The moral they draw is that 'neuroscientists must begin to test theories of how uncertainty could be represented in populations of neurons' (Knill and Pouget, 2004, p. 718). Given information is inherently a measure of uncertainty, the infotropic machine offers a candidate solution.

4.2 Where are the error units?

Also seemingly missing from the infotropic prediction machine are 'error units.' These are generally seen to be the medium by which prediction error is communicated upwards, from one level of the hierarchy to the next. They are the means of correcting error. This is considered significant from the explanatory point of view, as it leads us to expect a particular form of neurophysiology. We expect cells in the brain to be divided into two groups. There should be 'two functionally distinct sub-populations' (Friston, 2005, p. 829), one comprising the 'error-detecting neurons that signal the difference between an input and its prediction' (Rao and Ballard, 1999, p. 84), and the other comprising the representation units of the generative model.

Some theorists take the view that error units should not be taken too literally. Rasmussen and Eliasmith note that evidence for use of predictive processing does not necessarily imply 'a sharp division in the brain into these two different sub-populations' (Rasmussen and Eliasmith, 2013, p. 224). Functional integration is seen as equally viable, given 'units could be simultaneously sensitive to both error and representation, and still perform the relevant computations' (Rasmussen and Eliasmith, 2013, p. 224; see also Eliasmith and Anderson, 2003). More generally, it is acknowledged that predictive processing might be implemented in a way that departs from the envisaged 'duplex' architecture (Knill and Pouget, 2004; Doya et al., 2007). In Muckli's view, the 'claim

that the brain is a prediction machine might be true regardless of the precise implementation of predictive coding mechanism' (Muckli et al., 2013, p. 221).

Significantly, evidence for error cells in the brain remains inconclusive (Bubic et al., 2010). As Clark comments, 'direct, unambiguous neural evidence for these crucial functionally distinct sub-populations is still missing' (Clark, 2013a, p. 192). Further, Egner and Summerfield note that evidence supporting the concept of a duplex architecture is demonstrably lacking in the case of visual cortex. As they say, 'the proposition that there are simultaneous computations of prediction and prediction error signals carried out by distinct neural populations in visual cortex is presently only poorly substantiated' (Egner and Summerfield, 2013, p. 211; see also Summerfield and Egner, 2009). In Clark's view, this disconnect between the idea of a duplex brain and the neurophysiological evidence is 'potentially problematic' (Clark, 2013a, p. 188).

However this may be, the explanation for the lack of error units in an infotropic machine is simple enough. They are redundant in this context. Prediction error is considered to be indicated by negative informational value. Informational evaluations are implicitly measurements of error in this sense. Where one outcome acquires relatively greater informational value than another, an outcome that is relatively more laden with error is replaced with one that is relatively less laden. The effect is to 'correct' error. The need for an explicit signalling mechanism is avoided.

4.3 Where are the feedforward signals?

One of the attractions of the predictive-processing account is the way it explains forward signaling in the brain. As Clark notes, the 'information that needs to be communicated "upward" under all these [predictive processing] regimes is just the prediction error: the divergence from the expected signal' (Clark, 2013a, p. 183). Part of the appeal of this is that it tells us what feedforward signals mean. But there remains the difficulty that error-signaling cells have not been conclusively observed (see above). There is also the problem of semantics. If feedforward signals mean 'error', what do feedback signals mean? The general assumption is that they must represent predictions in some way (e.g. Clark, 2013a). Some theorists see this as too simplistic, however. Spratling comments that Clark is wrong to assume 'the feedforward flow of information solely conveys prediction error, while feedback only conveys predictions' (Spratling, 2013, p. 51). The way the infotropic account addresses the situation will already be apparent. On this view, both feedforward and feedback signals encode informational value. This is a language that can articulate both prediction and error.

4.4 Where is the Bayesian inference?

Some of the ways an infotropic machine replicates Bayesian inference have already been noted. If we look at the behavior of a machine such as the one

depicted in Figure 3, we see that it reproduces the effects of Bayesian MAP inference. If we look at the underlying mathematics, we see why. Predictive payoff is essentially Bayes' rule rewritten to use information as a currency. But there is also a more global sense in which an infotropic machine performs Bayesian inference. If global states of the hierarchy are considered to represent individual hypotheses, movement towards informational value is also movement towards the most globally predictive state. This can be seen as converging on the hypothesis that maximizes probability of the data. This is Bayesian inference of a more general type. An infotropic machine can be considered to perform Bayesian inference in this more abstract sense as well.

4.5 Where is the free energy?

A popular version of the predictive processing theory stems from the work of Friston and colleagues (e.g. Friston, 2005; Hohwy et al., 2008; Friston, 2010; Friston et al., 2012; Friston, 2013). In this 'free energy' variant, the mandate to reduce prediction error is seen to derive from a still more fundamental imperative: minimization of informational surprise (Friston, 2013; Friston et al., 2012). Minimization of surprise becomes the underlying objective—predictive processing is the way it is achieved. As Friston puts it, 'Predictive coding is a consequence of surprise minimization' (Friston, 2013, p. 32).[17]

On the assumption that informational surprise (entropy) cannot be minimized directly, the variational method of (Dayan et al., 1995; Hinton and Zemel, 1994) is brought into play. Free energy—an upper bound on entropy—is minimized by means of prediction-error observations, with the effect of minimizing informational surprise within the bounds of feasibility. The default behavior of an infotropic machine accomplishes something along these lines. The machine automatically transitions towards informational value, which implies a reduction of entropy. Accordingly, we might consider infotropic machines to be executing a kind of free-energy minimization. But the match is less than perfect, since different assumptions are made about how the informational value of a predictive model is established. In the free energy framework, it is established by minimizing free energy. In the infotropic account, it is established by Equations 2 and 3.

That being said, there is nothing to prevent making a connection at a more abstract level. Optimizing an infotropic machine with respect to a particular environment is certainly a possibility. The process would involve finding the hierarchical structure which, for the data generated by the environment, yields the greatest concentration of information, and thus the greatest predictive impact. An optimization process of this form is arguably the natural counterpart of free-energy minimization, albeit—confusingly—one which operates on an object that itself fulfils the predictive-processing mandate. There is a kinship between

[17] More generally, minimizing sensory entropy is seen to be a fundamental imperative of nature, counteracting the "thermodynamic forces" which increase physical entropy. The imperative to minimize sensory entropy is also seen to explain 'our curious (biological) ability to resist the second law of thermodynamics' (Friston, 2013, p. 213).

the present proposal and the free-energy account at some level. Tying them together in a detailed way is challenging.

5 Concluding remarks

Bayesian probability theory has long been the framework of choice for work which views the brain as a hierarchy of generative models. It is well suited to this purpose. But information theory is an alternative worth considering. Once the metric of predictive payoff is taken into account, a prediction machine can be realized by a network of inter-predicting outcomes (subject to the rule that evaluations are kept up-to-date). A hierarchical prediction machine can then be just a predictive network with a hierarchical structure. This has the effect of simplifying the account, since it does away with the need to differentiate error units, error correction, precisions and Bayesian inference.

On the face of it, simplifying the machine in this way takes us away from a tried and trusted methodology. But digging deeper, the conversion is not quite as far-reaching as it seems. The driving force in the Bayesian machine is Bayes' rule. In the infotropic machine it is predictive payoff. But there is a close relationship between the two formulae, as noted. It is not unlikely, then, that the infotropic version of the hierarchical prediction machine will turn out to be functionally if not mathematically equivalent to the Bayesian version. The experiments described above seem to support this. More experimentation is needed before any firm conclusions can be drawn; it is hoped future work will make progress in this direction.

References

Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology*, London: The MIT Press.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78* (pp. 1-3).

Brown, H., Friston, K. and Bestamnn, S. (2011). Active inference, attention and motor preparation. *Frontiers in Psychology*, *2*, 218.

Bubic, A., von Cramon, D. Y. and Schubotz, R. I. (2010). Prediction, Cognition and the Brain. *Frontiers in Human Neuroscience*, *4*, No. 25 (pp. 1-15).

Chater, N., Tenenbaum, J. and Yuille, A. (2006). Probabilistic Models of Cognition: Conceptual Foundations. *Trends in Cognitive Sciences (Special issue on Probabilistic Models of Cognition)*, *10*, No. 7 (pp. 287-291).

Clark, A. (2013a). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36* (pp. 181-253).

Clark, A. (2013b). The many faces of precision (Replies to commentaries on "Whatever next? Neural prediction, situated agents, and the future of cognitive science"). *Frontiers in Psychology*, *4*, 270.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*, Oxford: Oxford University Press.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory 2nd Edition*, New York: John Wiley and Sons.

Dayan, P., Hinton, G. E., Neal, R. M. and Zemel, R. S. (1995). The Helmholtz Machine. *Neural Computation*, *7* (pp. 889-904).

Doya, K., Ishii, S., Rao, R. P. N. and Pouget, A. (eds.) (2007). *The Bayesian Brain: Probabilistic Approaches to Neural Coding*, MIT Press.

Doya, K. (2007). Preface. In Doya (Ed.), *Bayesian Brain: Probabilistic Approaches to Neural Coding* (pp. xi-xii), MIT Press.

Dretske, F. I. (1981). *Knowledge and the Flow of Information*, Oxford: Basil Blackwood.

Dretske, F. I. (1983). Précis of *Knowledge and the Flow of Information*. *Behavioral and Brain Sciences*, *6* (pp. 55-90).

Egner, T. and Summerfield, C. (2013). Grounding predictive coding models in empirical neuroscience research. *Behavioral and Brain Sciences*, *36* (pp. 210-211).

Eliasmith, C. and Anderson, C. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*, MIT Press.

Eliasmith, C. (2007). How to Build a Brain: from Function to Implementation. *Synthese*, *159* (pp. 373-388).

Feldman, H. and Friston, K. (2010). Attention, uncertainty and free-energy. *Frontiers in Human Neuroscience*, *4*, No. 215.

Friston, K. J., Daunizeau, J. and Kiebel, S. J. (2009). Reinforcement Learning or Active Inference. *PLoS One*, *4*, No. 7 (pp. 1-13).

Friston, K., Adams, R. A., Perrinet, L. and Breakspear, M. (2012b). Perceptions as Hypotheses: Saccades as Experiments. *Frontiers in Psychology*, *3*, No. 151.

Friston, K., Thornton, C. and Clark, A. (2012a). Free-energy Minimization and the Dark Room Problem. *Frontiers in Perception Science*.

Friston, K. J., Lawson, R. and Frith, C. D. (2013). On hyperpriors and hypopriors: comment on Pellicano and Burr. *Trends in Cognitive Sciences*, *17*, No. 1 (pp. 1).

Friston, K. (2005). A Theory of Cortical Responses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *360*, No. 1456 (pp. 815-836).

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).

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*, Boston: Houghton Mifflin.

Haber, R. N. (1983). Can Information be Objectivized? *Behavioral and Brain Sciences*, *6* (pp. 70-71).

Hinton, G. E. and Zemel, R. S. (1994). Autoencoders, minimum description length and Helmholtz free energy. In Cowan and Alspector (Eds.), *Advances in Neural Information Processing Systems 6*, Morgan Kaufmann.

Hohwy, J., Roepstorff, A. and Friston, K. (2008). Predictive Coding explains Binocular Rivalry: An Epistemological Review. *Cognition*, *108*, No. 3 (pp. 687-701).

Hohwy, J. (2013). *The Predictive Mind*, Oxford University Press.

Huang, Y. and Rao, R. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2* (pp. 580-93).

James, W. (1890/1950). *The Principles of Psychology (Vol. 1)*, New York: Dover.

Jehee, J. F. M. and Ballard, D. H. (2009). Predictive feedback can account for biphasic responses in the lateral geniculate nucleus. *PLoS (Public Library of Science) Computational Biology*, *5*, No. 5.

Knill, D. C. and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neuroscience*, *27*, No. 12 (pp. 712-19).

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22* (pp. 79-86).

Lashley, K. S. (1951). The Problem of Serial Order in Behavior. In Jeffries (Ed.), *Cerebral Mechanisms in Behavior* (pp. 112-136), New York, NY: John Wiley & Sons.

Lee, T. S. and Mumford, D. (2003). Hierarchical Bayesian Inference in the Visual Cortex. *Journal of Optical Society of America, A*, *20*, No. 7 (pp. 1434-1448).

Luce, R. D. (2003). Whatever Happened to Information Theory in Psychology. *Review of General Psychology*, *7*, No. 2 (pp. 183-188).

Mackay, D. (1956). Towards an information-flow model of human behaviour. *Br. J. Psychol*, *43* (pp. 30-43).

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

Muckli, L., Petro, L. S. and Smith, F. W. (2013). Backwards is the way forward: Feedback in the cortical hierarchy predicts the expected future. *Behavioral and Brain Sciences*, *36* (pp. 221).

Ramsay, F. P. (1990). Truth and probability. In Mellor (Ed.), *Philosophical Papers* (pp. 52-109), Cambridge University Press.

Rao, R. P. N. and Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*, No. 1 (pp. 79-87).

Rao, R. P. N. and Ballard, D. H. (2004). Probabilistic Models of Attention based on Iconic Representations and Predictive Coding. In Itti, Rees and Tsotsos (Eds.), *Neurobiology of Attention*, Academic Press.

Rasmussen, D. and Eliasmith, C. (2013). God, the devil, and the details: Fleshing out the predictive processing framework. *Behavioral and Brain Sciences*, *36* (pp. 223-224).

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).

Shannon, C. E. (1956). The Bandwagon. *Transactions on Information Theory*, No. 3, Institute of Electrical and Electronics Engineers.

Spratling, M. W. (2013). Distinguishing theory from implementation in predictive coding accounts of brain function. *Behavioral and Brain Sciences*, *36* (pp. 231-232).

Summerfield, C. and Egner, T. (2009). Expectation (and Attention) in Visual Cognition. *Trends in Cognitive Sciences*, *13* (pp. 403-409).

Temperley, D. (2007). *Music and Probability*, Cambridge, Massachusetts: The MIT Press.

Thornton, C. (2014). Infotropism as the underlying principle of perceptual organization. *Journal of Mathematical Psychology*, *61* (pp. 38-44).

Tolman, E. C. (1948). Cognitive Maps in Rats and Men. *Psychological Review*, *55* (pp. 189-208).

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

Yu, A. J. and Dayan, P. (2005). Uncertainty, neuromodulation and attention. *Neuron*, *46* (pp. 681-692).

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.