Predictive processing: Does it compute?

Chris Thornton
University of Sussex
Brighton
UK
c.thornton@sussex.ac.uk

June 2, 2020

Abstract

In what is called the predictive processing framework, the brain is viewed as a multi-layered prediction engine, whose task is to anticipate incoming flows of sensory information. Each layer of the engine is seen to express a generative model, in an arrangement that involves higher layers sending predictions to lower layers, and lower layers passing prediction errors upward. Minimizing these errors is assumed to turn the structure into a largely veridical model of the world. The scheme is advocated as a way of explaining processing in the brain. But what is its status from the computational point of view? What calculations are implied? Over what data do they operate? What effects are achieved? This paper considers predictive processing from a computational/engineering perspective, and identifies a number of technical problems in the scheme. How these can be eliminated is also considered.

1 Introduction

There is mounting enthusiasm for what Clark calls 'the emerging unifying vision of the brain as an organ of prediction using a hierarchy of generative models' (Clark, 2013, p. 185). Part of a long tradition emphasizing the role of prediction in perception (von Helmholtz, 1860/1962; James, 1890/1950; Tolman, 1948; Lashley, 1951; Mackay, 1956), this approach is now advancing on a broad range of fronts (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; Williams, 2018; Yon et al., 2019). Given the principle 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), the proposal can also be seen as a way of unifying interpretive and behavioral functionality (Brown et al., 2011; Friston et al., 2009). The implication then becomes that 'perceiving and acting are but two different ways of doing the same thing' (Hohwy, 2013, p. 76).

Clark's proposal (Clark, 2013, 2016) characterizes function organized in this way as predictive processing (PP). In his view, the brain is an inner engine of probabilistic prediction that is 'constantly trying to guess at the structure and shape of the incoming sensory array' (Clark, 2016, p. 3). Each layer of the engine is seen to express a generative model, in an arrangement that involves higher layers sending predictions to lower layers, and lower layers passing prediction errors upward. Minimizing these errors, it is presumed, will turn the structure into a largely veridical model of the world.

The proposal continues to gain support (Williams, 2018; Yon et al., 2019). But there are questions about what is implied computationally. The operations involved are not generally specified in precise detail, as Clark acknowledges. He describes his own characterization of predictive processing (Clark, 2016) as 'relatively abstract' (p. 298), and no more than a 'mid-level organizational sketch' (p. 2). But with the fine-details left open in this way, questions of functionality inevitably arise. By means of what calculations do higher layers send predictions downward? How are the prediction errors computed? What is the mechanism for transmitting these upward through the hierarchy?

A key question relates to layer coordination. In Clark's description of the scheme, individual layers in the hierarchy are considered to be largely independent. Each layer is seen to predict '... the response profiles at the layer below' (Clark, 2016, p. 93), while also reducing any error reported back. Predictions make up the downward flow of information within the hierarchy, while error signals make up the upward flow. There is no other mechanism of communication or coordination between layers. As Clark emphasizes, it is a distinctive characteristic of the proposal that 'it depicts the forward flow of information as solely conveying error, and the backward flow as solely conveying predictions' (Clark, 2016, p. 38).

The difficulty is to see how this arrangement would have the effect of improving prediction of sensory input overall. The onus on each layer is to reduce its error in predicting the state of the layer below.[1] But it is only the state at the lowest layer of the hierarchy which represents sensory data. Without layers being coordinated in some way, there is no reason why prediction at the sensory layer should be improved by reducing error higher in the hierarchy. Representational coordination of layers would seem to be pre-requisite. In the case of data compression by predictive coding, often cited as an inspiration for PP, coordination of predictive sources is ensured by specification of the algorithm (e.g. Kobayashi, 1974; Pensiri and Auwatanamongkol, 2012).

The lack of specificity about critical computational details is also problematic. The calculations that are assumed to convey the upward and downward flows of information are not precisely specified and, as will be seen, it is not

[1] See Clark's assertion that each layer must be '... capable of predicting the response profiles at the layer below' (Clark, 2016, p. 93); see also Williams' observation that 'the data for every level of the hierarchy—with the exception of the first—consists of representations at the level below' (Williams, 2018, p. 151).

always obvious what is implied. There is a need to develop a specification that resolves these ambiguities. This is the initial aim of the present paper. The intention is to establish what predictive processing involves at a detailed, computational level of description.

The calculations that mediate upward and downward flows in predictive processing are potentially defined in terms of the inferential operations of Bayesian probability theory (Berger, 1985; Howson and Urbach, 1989; Jaynes, 2003). Use of the framework of information theory (Shannon, 1948; Shannon and Weaver, 1949) is also a possibility. Which of these two approaches leads to a more coherent computational implementation will be carefully examined. The degree to which adopting a specific implementation addresses the question of layer coordination will also be assessed.

The paper is divided into four main sections. Sections 2 and 3 examine contrasting implementations of predictive processing. Section 2 evaluates an implementation based on use of Bayesian calculations; Section 3 examines an implementation based on use of information-theoretic calculations. The degree to which the approach addresses the requirement for layer coordination is explored. Section 4 then presents a general discussion and some concluding comments.

2 An inferential implementation

Although the calculations that mediate predictive processing are not generally specified in detail, it is often assumed they must be Bayesian in nature (Hohwy, 2013). They are typically taken to be acts of probabilistic inference. This acknowledges the degree to which the framework is founded in the Bayesian-brain hypothesis, which proposes that 'the brain codes and computes weighted probabilities' (Clark, 2016, p. 41), and the general assumption that neural processing can be understood as a form of Bayesian inference (cf. Doya et al., 2007; Pouget et al., 2013).

There are two ways in which an inferential implementation of PP can be developed, however. A Bayesian probability linking two outcomes allows an unconditional probability to be inferred for either the conditioning outcome, or the conditioned outcome. There are two forms of inference available—one forward, one backward—for mediating the flows of information, and two forms of flow—upward and downward—to be implemented. A particular implementation can be arrived at by mapping inferential forms to flows in a particular way, therefore.

A convenient approach links the downward flow to forward inference. This has the unfortunate consequence of creating a terminological clash, as the downward flow is termed 'backward' neurologically. The advantage is that it has the effect of placing conditioned outcomes (i.e., Bayesian evidence) below conditioning outcomes (i.e., hypotheses) in the hierarchy. Outcomes representing sensory evidence are also placed in the lowest layer. Mediation of the upward flow is then by means of Bayes' rule—i.e., by derivation of posterior probability.

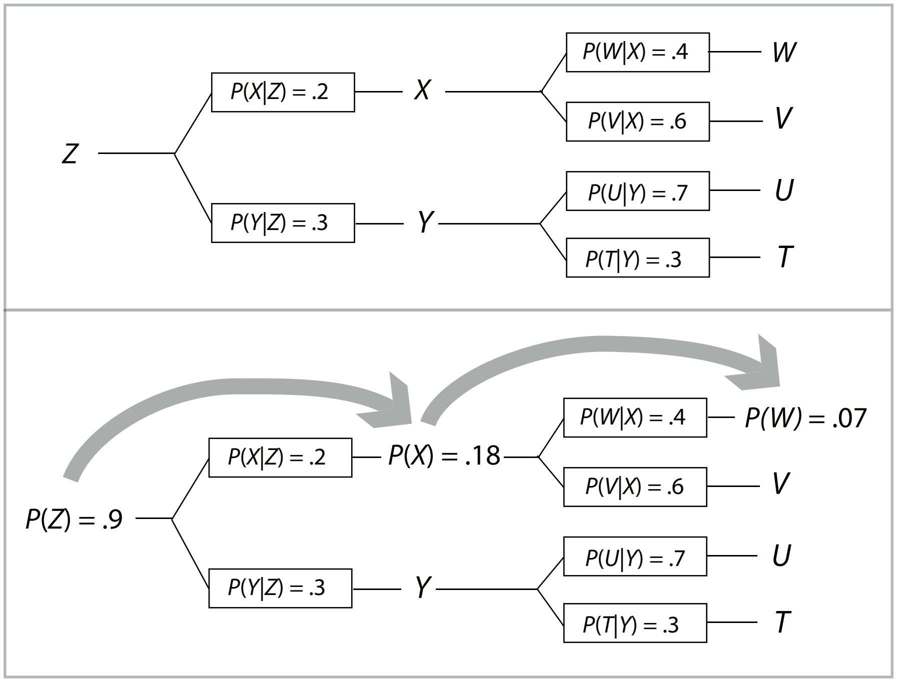

Figure 1: Downward flow through Bayesian inference.

This general scheme is recognized by Clark as one way of realizing predictive processing in a specifically Bayesian way (Clark, 2016, pp. 172-175).

A simple illustration appears in the upper panel of Figure 1. This shows a single Bayesian hierarchy involving the outcomes T, U, V, W, X, Y, Z. The hierarchy is drawn on its side, with the root on the left. Each between-layer connection is defined by a likelihood, as shown. This states the conditional probability of an outcome in the layer below, given an outcome in the layer above. For example, $P(X|Z) = .2$ defines the top-left connection. (Here and elsewhere, probabilities are approximate.) Outcomes at the lowest layer are considered to be sensory in nature. Outcomes W, V, U and T are sensory, then, while outcomes X, Y and Z are internal.

The lower panel of Figure 1 illustrates processing involved in the downward flow. The sequence is initiated by the assertion of an unconditional (i.e., prior) probability for the root outcome Z. With $P(Z)$ given the value .9, unconditional probabilities can then be inferred for outcomes at lower layers. $P(X) = P(X|Z)P(Z) \approx .18$ can be derived, followed by $P(W) = P(W|X)P(X) \approx .07$. The probability given to Z eventually yields a probability for W in this way.

Under the depicted arrangement, priors at higher layers of the hierarchy are seen to 'cascade downwards' by probabilistic expectation. A point in favour of this scheme is that it echoes the way theorists have assumed the downward flow must operate. Both Clark (2013) and Hohwy (2013) describe the downward flow in just these terms. Hohwy refers to the importance of what he calls the 'pulling down' of priors (Hohwy, 2013, p. 33). Clark sees downward flow as the way in which a system can '... infer its own priors (the prior beliefs essential to the guessing routines) as it goes along' (Clark, 2013, p. 3). The proposed implementation, in which the downward flow is exclusively in this form, may go beyond what these theorists intend, however. The role played by lateral (within layer) connectivity is also particularly emphasized in (Clark, 2016), for example.[2]

Figure 2 illustrates the upward flow. The upper panel shows the (approximate) posterior probabilities that can be inferred for X and Z, after W is awarded an unconditional probability of .5. All outcomes are considered to have a default prior of 1, and each inferential step represents an application of Bayes' rule. X's probability, for example, is the posterior

$$P(X|W) = \frac{P(W|X)P(X)}{P(W)} \approx .8$$

The lower panel shows the upward flow after all the sensory outcomes are awarded unconditional probabilities, and posteriors are combined appropriately. As will be seen, X comes to acquire an (approximate) unconditional probability of .6 rather than .8, due to the influence of $P(V|X) = .6$ and $P(V) = .8$.

The results achieved by implementing the upward flow in this inferential way are less satisfactory. What is then conveyed upward is posterior probability, whereas what should be conveyed upward is prediction error. The two quantities are related, so it is not unreasonable to ask whether one might represent the other. It is certainly the case that the posterior probability of an outcome increases with the degree to which it is predicted by the relevant likelihood and conditional prior. Treating posterior probability as a an *inverse* measure of prediction error might be considered an option on this basis.

In practice, this arrangement fails. It can be shown that posteriors cannot represent prediction errors in certain cases. Imagine a situation in which the predicted probability of an outcome exceeds its observed probability. Rain might be observed to have a probability of .5, say, but be predicted on theoretical grounds to have a probability of .9. (This situation, in which the predicted probability of an outcome exceeds its observed probability, is typical for a weather forecast.) The prediction potentially gives rise to a prediction error. But an error of this kind cannot be dealt with by derivation of posteriors, even in principle.

Given the predicted probability derives from a likelihood of .9 and a prior of 1, the posterior probability of rain cannot be derived. The numerator in the Bayesian calculation is then greater than the denominator, implying an invalid posterior. Within the terms of Bayesian theory, the posterior has to be

[2] Clark notes that 'in the standard implementation of PP higher level "representation units" send predictive signals laterally (within level) and downwards (to the next level down) thus providing priors on activity at the subordinate level' (Clark, 2016, p. 143).

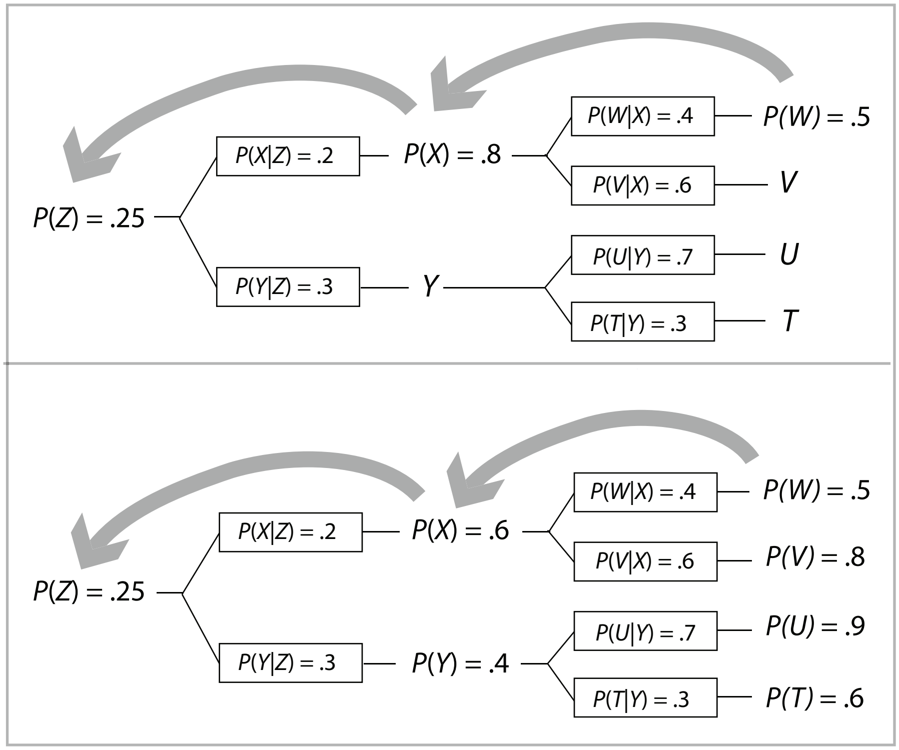

Figure 2: Upward flow by means of Bayesian inference.

considered undefined.[3] Posteriors cannot represent prediction errors in general, then. Implementing the upward flow in this way is ruled out.[4]

The general conclusion is that attempting to operationalize predictive processing in a strictly Bayesian way, using inferential calculations to mediate information flows, fails to produce a satisfactory result. Arguably, the effect is simply to beg further questions. If the upward flow of posteriors cannot serve to convey prediction error, how does it fit in to the processing otherwise performed? If transfer of error involves information flow that progresses up the hierarchy in some way, how does this interact with the upward flow that reflects ordinary Bayesian inference? Do the two flows proceed in parallel? Are they integrated in some way? The attempt to develop a strictly Bayesian implementation of

[3] Neutral predictions are problematic for the same reason. Imagine the predicted probability of rain is .5. Given the observed probability of rain is also .5, this prediction is entirely neutral. It is completely without value and, in that sense, maximally in error. The derived posterior, on the other hand, is now maximized, implying minimal prediction error.

[4] An additional problem with this implementation is that fails to respect the stipulated architectural requirements. It is seen as a key part of the scheme that there should be a '... functional separation between encodings of prediction and prediction error' (Clark, 2016, p. 39).

predictive processing begs a number of questions.

3 An informational implementation

The information flows in predictive processing can also be calculated using operations drawn from the framework of information theory (Shannon, 1948; Shannon and Weaver, 1949). Like the Bayesian framework, this deals with assignments of probability. But whereas the Bayesian framework is concerned with how probabilities can be updated from relevant priors and likelihoods, information theory focuses on what happens when an outcome of given probability *occurs*. A way of quantifying the information that is then generated is the framework's key contribution. Being focused on outcomes in this way, information theory is well-suited to deal with prediction of outcomes, and hence with predictive processing.

Key to information theory is the principle that an outcome has an informational value that is inversely related to its probability (Shannon, 1948; Shannon and Weaver, 1949). Specifically, the informational value of some outcome W is defined as $-\log_2 P(W)$ bits.[5] If $P(W) = .5$, for example, the informational value of W is $-\log_2 .5 = 1$ bit. This quantity is termed the outcome's surprisal (Tribus, 1961). For present purposes, it is prediction of outcomes that is of interest, and this is naturally modeled in terms of conditional probability. Consider $P(W|X)$. This denotes the conditional probability of outcome W given outcome X. Equivalently, it can be seen to denote the probability of outcome W that outcome X predicts. The assertion $P(W|X) = .3$ can be considered to assert that outcome X predicts outcome W with probability .3, for example.

A relatively simple PP implementation can then be envisaged, which is essentially a direct translation of the Bayesian implementation. In it, the conditional probabilities which define the structure of the model are considered to express predictions. Upward and downward flows are progressed by probabilistic expectation as before, but with the values derived now being quantities of information rather than probabilities. Figure 3 illustrates the processing then obtained. The upper panel depicts the downward flow commencing from the assignment $P(Z) = .25$. This yields a surprisal value—here denoted as $\inf(Z)$ — of 2 bits. The lower panel depicts the upward flow commencing from the assignment by $P(W) = .5$. This yields an information value of .2 bits for outcome Z.

Unfortunately, this direct translation of the Bayesian implementation lacks generality. It accommodates the various ways in which outcomes can be predicted, but not the possibility that they might not occur. This needs to be taken into account. The informational value of an outcome is its surprisal, but only if the outcome occurs. A predicted outcome has the potential *not* to occur. The outcome may be mispredicted. Where there is prediction of an outcome that fails to occur, how is the expected value of the predicting outcome to be derived? To obtain a fully general specification for the upward flow, this issue needs to be resolved.

[5] The quantity is expressed in bits just in case logs are taken to base 2.

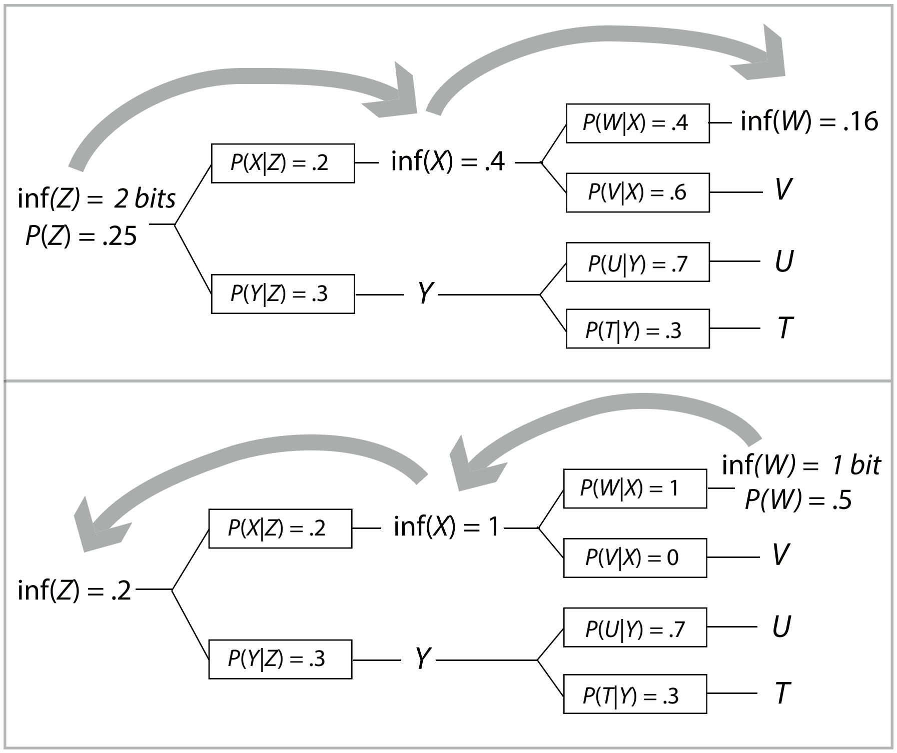

Figure 3: Upward and downward flows using informational values.

Shannon's framework (Shannon, 1948; Shannon and Weaver, 1949) does not explicitly address the case of non-occurring outcomes. It can be deduced that a prediction of a non-occurring outcome must always have a negative value in this context, however. This is best demonstrated using a concrete example. Imagine a weather forecast that gives a 100% chance of rain during the day. This prediction can be modeled as the assertion $P(W|X) = 1$, where W is occurrence of rain, and X represent whatever cue(s) the forecast derives from. Assuming rain is observed to occur on 50% of all days, we have $P(W) = .5$, and hence $\inf(W) = \log_2 .5 = 1$ bit. Given these assignments, what is the informational value of the prediction of rain?

If the rain occurs, the evaluation is straightforwardly derived. The forecast has the effect of bringing forward the outcome in question, supplying its information content in advance. Acceptance of the forecast increases possession of information by the value of the outcome predicted. The value of the forecast is 1 bit. If the rain does not occur, the forecast is then a misprediction and, as we might expect, the evaluation turns negative. To see this, consider a forecast that gives a 50% chance of rain. Implicitly, this also gives a 50% chance of *no rain* — it gives the two possible outcomes equal probability. The forecast is completely neutral then and, given the observed probability of rain is itself .5, entirely without value.

On this basis, the evaluation of the original forecast in the case of the rain failing to occur can be deduced. Assuming the value of a successful prediction must be positive, it follows that the value of its unsuccessful counterpart must be correspondingly negative. This is what the zero evaluation of the neutral forecast entails. With correct and incorrect predictions given equal probability, the forecast's zero evaluation requires that the positive value of the correct prediction is precisely offset by the negative value of the incorrect prediction. The evaluations of the prediction and misprediction must be equal and opposite.

Combining this with the observation that the value of a correct prediction is the surprisal of the outcome predicted, the value of a misprediction can then be defined as the negative of the outcome's surprisal. This has the effect of ensuring that a misprediction is exactly as costly as its counterpart is beneficial. The relationship can be stated formally as

$$I_P(e) = \begin{cases} -\log_2 P(e) & \text{if } e = e' \\ \\ \log_2 P(e) & \text{if } e \neq e' \end{cases} \tag{1}$$

where $P(e)$ is outcome e's observed probability and e' is the outcome that occurs. $I_P(e)$ is then the informational value of predicting outcome e. (Notice that the upper value is positive, and the lower negative).

As outcomes may be predicted with any probability in general, this should be allowed for. The overall evaluation then becomes a weighted average of the gains and losses produced by the individual parts of the forecast. It is the expected informational revenue of the predicted distribution. Let $Q(e)$ be the probability with which outcome e is predicted.[6] The informational value of the predictions expressed by distribution Q is then the average $I_{Q:P}$:

$$I_{Q:P} = \sum_e Q(e) I_P(e) \tag{2}$$

The measure can also be generalized for situations involving more than two outcomes. This requires use of a normalization. With more than two outcomes, we have more than one incorrect outcome, and consequently, more than one negative value in the summation of Eq. 2. To ensure commensurability between positive and negative contributions, the latter must be discounted by $n-1$, where n is the number of outcomes. The modified equation then becomes

$$I_P(e) = \begin{cases} -\log_2 P(e) & \text{if } e = e' \\ \\ \dfrac{\log_2 P(e)}{n-1} & \text{if } e \neq e' \end{cases} \tag{3}$$

Situations involving any number of outcomes can then be dealt with. The informational value of a prediction, assessed in this way, is termed its *predictive*

[6] We might have $Q(\text{rain}) = .3$ and $Q(\text{no rain}) = .7$, for example.

payoff (Thornton, 2017).[7]

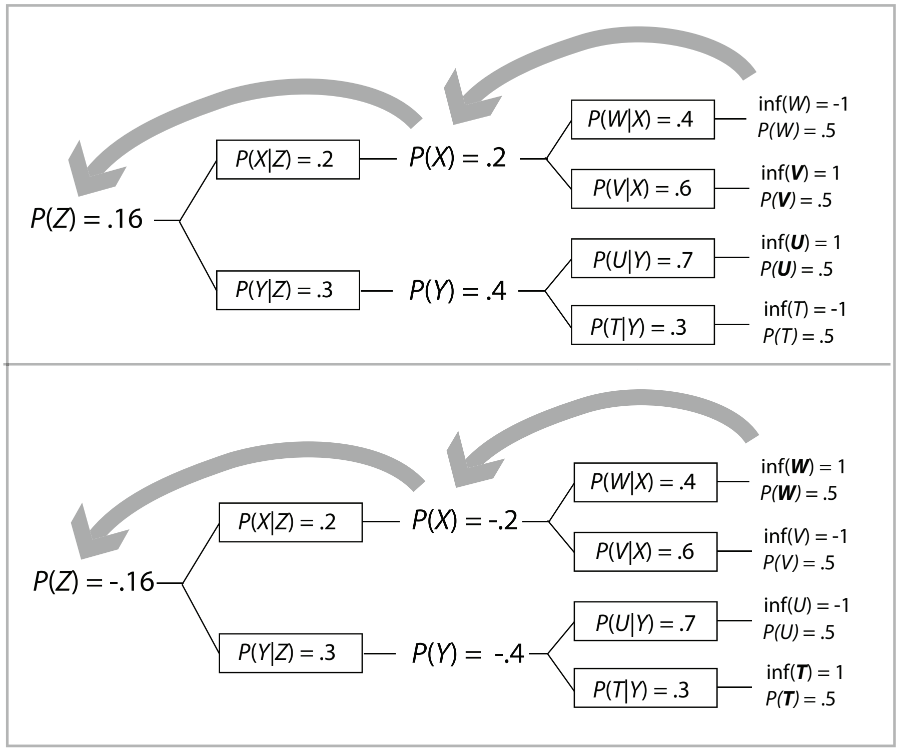

Figure 4: Upward flow by information-theoretic calculation.

How the upward flow is to be implemented in the face of outcomes that may or may not occur can then be resolved. The evaluation that any higher outcome obtains should be its predictive payoff with respect to the outcomes that it conditionally awards probability to. Upward flow should be progressed by derivation of predictive payoff, in other words. The upper and lower panels of Figure 4 illustrate the patterns of processing that are obtained on this basis.

The upward flow of information to X, and from X to Z, is depicted in two different scenarios. The upper panel deals with the case where the sensory outcomes predicted by X with greatest probability (i.e., V and U) both occur. The lower panel deals with the case where they fail to occur. (Notice occurring outcomes are set in bold.) The upward flow is mediated by derivations of predictive payoff in all cases. Outcome X acquires a value of approximately .2 bits due to its attribution of probability to W and V, and their informational values. The payoff is

[7] The formulation presented in (Thornton, 2017) differs in the way it discounts disrewards in the case of there being more than one non-occurring outcome.

$$-1 \times .4 + 1 \times .6 \approx .2 \text{ bits}$$

An important feature of the upward flow, when implemented in this way, is the degree to which it both expresses prediction error, and brings about its reduction. This effect can be explained in terms of Figure 4. In the situation depicted in the upper panel, there are no mispredictions. All predictions are correct in the sense that occurring outcomes are given greater probability. (For example, $P(\boldsymbol{V}|X) > P(W|X)$.) Predictive payoff to both X and Y (and hence Z) is positive. In the situation depicted by the lower panel, the occurring outcome is T rather than U. This is now *mispredicted.* U is given greater probability than T, and the derived value of Y changes its sign in result. The prediction error leads to Y having a negative value.

The prediction error, in this case, leads to Y having a negative evaluation, and this then comes to serve as a kind of error signal. No particular significance attaches to the evaluation's sign. Its capacity to serve as a prediction error depends purely on its relational properties. The strength with which a model in this form expresses a particular prediction depends on the informational value of the outcome from which it derives. The effect produced by the upward flow is to concentrate predictive strength at outcomes that better predict. Information flows towards sources of prediction, then, and in proportion to their predictive efficacy. Information congregates as prediction originates. On this basis, better predictions are then naturally forthcoming. Prediction error that is implicitly conveyed upward is also implicitly reduced. The system predicts, and reduces error at the same time, without requiring any extraneous mechanism of error-reduction.

Unlike its predecessor, this informational implementation meets all requirements of the predictive processing scheme, then. Both information flows are handled appropriately. Error is conveyed upward in a way that ensures it is reduced at each layer. Predictions are conveyed downward in a way that meets the requirement for the model to be 'generative in nature' (cf. Clark, 2016, p. 93). An additional attraction is that the implementation respects Clark's description of error derivation. This requires the upward flow to originate in measurements of surprisal, specifically.[8] Upward flow in the proposed implementation originates in exactly this way. One drawback of the implementation, however, is that it fails to separate encodings of prediction and prediction error in the way that Clark (2016, p. 39) emphasizes is important. Under the proposed implementation, they are fully integrated.

[8] Clark states that prediction error should reflect '... the 'surprise' induced by a mismatch between the sensory signals encountered and those predicted. More formally—and to distinguish it from surprise in the normal, experientially loaded sense—this is known as surprisal' (Clark, 2016, p. 25).

4 Discussion

Regardless of how useful the predictive processing framework may be for explaining functionalities of the brain, there is a need to determine whether it makes sense computationally—whether it hangs together as a system of calculation. This is the main aim of the present paper. The result of the study is largely positive. It has been shown that a computationally precise interpretation of the scheme can be assembled. There are various reservations to be noted, however.

Layer coordination is a prominent concern. The original scheme envisages a hierarchical structure in which higher layers send predictions to lower layers, and lower layers pass prediction errors upward. On the proviso that each layer predicts the state at the layer below (and the lowest layer predicts sensory input),[9] it is assumed that minimizing error at any layer will have the effect of improving prediction of sensory input overall. This is problematic. If it is assumed that the task of each layer is to predict the state at the layer below, there is no reason why this effect should occur. Improving prediction of a state that itself predicts badly cannot be a way to make it predict better. The effect might easily be the reverse.

Designing the hierarchy in a way that allows each layer to predict sensory input directly would appear to resolve this difficulty. Any error passed upward is then the residual of a single quantity. It is what remains after lower-layer predictions have been taken into account. Minimizing errors of this kind can only improve prediction of sensory input overall. Unfortunately, an arrangement of this kind only solves the problem in a degenerate way. If the error reported by every layer of the hierarchy is a residual in this way, all layers are then functionally linked together, and the attribution of hierarchical structure is called into question. Viewing the structure as a single, non-hierarchical model would seem equally justified.

Implementations of predictive processing would seem to face an inherent dilemma, then. Without coordination of layers, there is no reason why error reduction at higher layers should improve prediction of sensory input. With coordination of layers, the model's hierarchical structure is cast into doubt. The hierarchical structure identified seems to be essentially a projection. How this predicament can be resolved is not obvious. Working computational incarnations of predictive processing do exist, however. The predictive coding model of Rao and Ballard (1999) is often cited as a demonstration. If the scheme faces an irresolvable dilemma, how do we explain systems such as this, which seem to demonstrate its feasibility?

It is worth looking at the system described by Rao and Ballard in more detail. It is key to the design of this that predictions apply to sensory input directly. Layers are then representationally coordinated in the desired way. An error passed upward is always a residual error, on which basis reducing error at any layer of the hierarchy improves prediction of sensory data. In principle,

[9] In Williams's (2018) description, 'the data for every level of the hierarchy—with the exception of the first—consists of representations at the level below' (Williams, 2018, p. 151).

this arrangement calls into question the hierarchical structure of the model, as noted. But in the Rao and Ballard system, the model's hierarchical structure has a validity that is seen to exist independently.

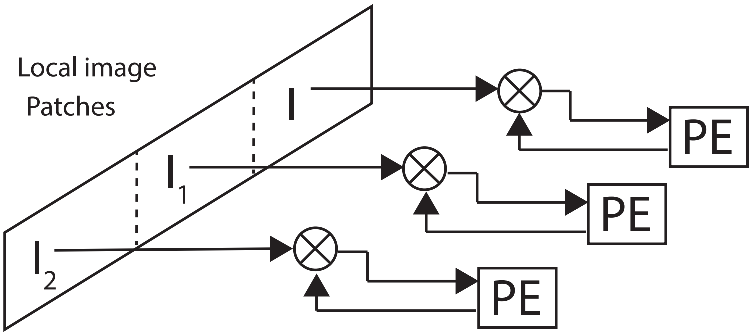

Figure 5: Detail from part C of Fig. 1, Rao and Ballard (1999) showing an initial stage in the hierarchical organization of sensory data. The 'PE' units are predictive estimators. Higher level predictive estimators reference receptive fields that are hierarchically composed from fields referenced by lower level estimators.

It is part of the system's design that the presentation of sensory input is itself hierarchically organized. Sensory data derives from the responses of visual receptive fields, as seen in Figure 5. Higher layers make reference to receptive fields that are hierarchically composed out of those referenced by lower layers. An error passed upward is then a residual error in a spatial sense. It refers to effects that extend beyond the limits of receptive fields for the current layer, but which remain *within* the composite fields referenced by the higher layer. Reducing error at any layer of the hierarchy then has the effect of tuning the model to the hierarchical structure of the input stream. The system is then seen to 'learn a hierarchical internal model of its natural image inputs' (Rao and Ballard, 1999, p. 80).

The Rao and Ballard system is of great interest, but there is a risk of overinterpreting it. The temptation is to assume that its derivation of a specifically hierarchical model is a consequence of the predictive processing it carries out. This potentially leads to predictive processing being viewed as a general, data-driven method for learning hierarchically structured models.[10] A computational assessment of the situation suggests the method's capabilities fall short of this, and that it can be applied only to models whose hierarchical structure is pregiven and in some way externally validated.

[10] Consider, for example, Lupyan and Clark (2015) who suggest that a '...remarkable consequence of this [predictive processing] arrangement is that seeking to reduce the overall prediction error produces representations at multiple levels of abstraction, flexibly incorporating whatever sources of knowledge help to reduce the overall prediction error' (Lupyan and Clark, 2015, p. 279).

At a more fine-grained level of detail, it is questions of calculation that become the main concern. What kinds of calculation should be used to progress the upward and downward flows that PP requires? Given the scheme's links to the Bayesian brain hypothesis, the expectation is that inferential calculations will suffice. As seen above, this is not what is found. A purely Bayesian way of implementing the downward flow can be identified—this is simply the process of 'pulling down' priors which theorists such as Hohwy (2013) and Clark (2013) have long viewed as integral to the scheme. Implementation of the upward flow is more problematic, however.

The PP scheme envisages an upward flow in the form of prediction errors that are passed from layer to layer. The problem is that a hierarchical Bayesian model naturally gives rise to an upward flow in a different form. Derivations of posterior probability (i.e., applications of Bayes' rule) are also constitutive of an upward flow. The computational relationship between the two flows is then difficult to reconcile. The assumption that they proceed in parallel faces difficulties, as does the assumption that they are integrated. One possibility is to assume the inferential flow mediates the transmission of prediction error. The difficulty then faced is that there are meaningful predictive scenarios in which posterior probabilities cannot be derived.

Attempting to implement predictive processing in a Bayesian/inferential way faces serious obstacles, then. Arguably, this is due to the incapacity of the Bayesian framework to deal satisfactorily with the phenomenon of prediction. Cases of prediction that cannot be conceptualized in Bayesian terms are easily identified. Any outcome may be predicted to have a probability that exceeds its observed probability. In some contexts, this is the norm. Rain that is observed to have a probability of .4 may be predicted—on evidential or theoretical grounds—to occur with probability .7. A car observed to break down with probability .2 may be predicted to break down with probability .8. A student observed to attend seminars with probability .6 may be predicted to attend one with probability .9, and so on. These predictions are in no way abnormal; all may turn out to yield errors in the usual way. From the Bayesian point of view, however, they are all intractable.

It is not that the Bayesian apparatus cannot be applied. A predicted probability can be expressed as the product of a likelihood and a hypothesis prior in the usual way. An observed probability can be expressed as the corresponding evidential prior. The difficulty is that the posterior probability of the hypothesis is then compromised. If the numerator in Bayes' rule exceeds the denominator, as it does in the cases above, what is obtained is a value greater than one. This cannot be a probability. An inferential approach to situations such as those above always fails in this way. It is arguable that the Bayesian framework is fundamentally ill-suited to deal with the phenomenon of prediction for this reason. The difficulty of obtaining a Bayesian implementation of predictive processing may well be a reflection of this.

The main finding of the present inquiry is that information theory is of more use for clarifying the computational details of predictive processing. In this respect, the analysis follows in the footsteps of Friston and colleagues (Friston, 2005, 2010; Friston et al., 2012; Friston, 2013; Friston et al., 2017) and the informational conception of prediction that they advocate. This emphasizes the degree to which prediction and prediction error are intimately related to uncertainty and information. Successful prediction must reduce uncertainty, and reduction of uncertainty accomplishes information gain (Mackay, 2003). Optimizing prediction of sensory input is equivalent to maximizing information gain in this sense. The objective of predictive processing can then be viewed as the task of minimizing the average surprisal (informational uncertainty) of sensory data. The conceptual perspective that this leads to is well summarized by Friston's dictum that 'Predictive coding is a consequence of surprise minimization' (Friston, 2013, p. 32).[11]

With predictive processing reconceptualized in this way, the proposal's scientific implications change to some degree. The idea that emerges is less that of a Bayesian brain and more that of an *infotropic* brain. Bayes' rule is seen to be replaced by information gain, as the underlying principle that guides processing (cf. Thornton, 2014, 2017).[12] For present purposes, however, it is the practical benefits of the informational approach that are mainly of interest. The framework of information theory is found to provide all that is required for a computationally precise interpretation of predictive processing.

A key attraction of this is its capacity to address the problem of layer coordination. Improving prediction of a state that itself predicts badly will not generally have the effect of improving the state's own predictive performance. A critical element of the predictive processing scheme is thus the requirement for representational coordination of layers. The problem, as noted above, is that this cannot be met without effectively eliminating the all-important hierarchical structure of the referenced model.

Under the informational implementation, this catch-22 is dealt with to some degree. What is conveyed upward in this arrangement is not prediction error as such. It is informational quantities that are derived from those which define the sensory data. The upward flow consists of a series of abstractions of (derivations from) the sensory data in this sense. Predictions at higher layers then do refer to the sensory data, but at differing removes. There is then no reason to doubt the efficacy of the processing scheme. Reducing error higher in the hierarchy can have the effect of reducing it at the lowest layer as well. Equally, there is no reason to doubt the hierarchical structure of the model.

[11] Friston's position is that surprisal must be minimized indirectly, via a bound termed free energy; the ground proposition is termed the free energy principle accordingly. The identified connection with surprisal remains unaffected, however. As Wiese and Metzinger note, '... free energy constitutes a tight bound on the surprisal of sensory signals. Hence, minimizing free energy by changing sensory signals will, implicitly, minimize surprisal' (Wiese and Metzinger, 2017, p. 12). Furthermore, free energy '... on most PP accounts would amount to the long-term average of prediction error' (Wiese and Metzinger, 2017, p. 18).

[12] It also suggests neural signals convey information encoded in a positive or negative (i.e., bidirectional) form. Potentially relevant to this is the observation that neural signals in the dopaminergic system can be modulated in a bidirectional way (Keller and Mrsic-Flogel, 2018, p. 425). Spratling (2017), however, takes the view that positive and negative firing rates are 'biologically implausible' (Spratling, 2017, p. 94).

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis (2nd ed.)*, Springer-Verlag.

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

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

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

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

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

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

Friston, K., Rigoli, F., Schwartenbeck, P. and Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, *29*, No. 1 (pp. 1-49).

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).

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.

Howson, C. and Urbach, P. (1989). *Scientific Reasoning: The Bayesian Approach*, Chicago, IL, US: Open Court Publishing Co.

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.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*, Cambridge, UK: Cambridge University Press.

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.

Keller, G. B. and Mrsic-Flogel, T. D. (2018). Predictive Processing: A Canonical Cortical Computation. *Neuron*, *100*, No. 2 (pp. 424-435).

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).

Kobayashi, H. (1974). Image Data Compression by Predictive Coding I: Prediction Algorithms. *Journal of Research and Development*, *18*, No. 2 (pp. 164-171).

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).

Lupyan, G. and Clark, A. (2015). Words and the World: Predictive Coding and the Language-Perception-Cognition Interface. *Current Directions in Psychological Science*, *24*, No. 4 (pp. 279-284).

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.

Pensiri, F. and Auwatanamongkol, S. (2012). A lossless image compression algorithm using predictive coding based on quantized colors. *WSEAS Transactions on Signal Processing*, *8*, No. 2.

Pouget, A., Beck, J., Ma, W. J. and Latham, P. (2013). Probabilistic brains: Knowns and unknowns. *Nature Neuroscience*, *16*, No. 9 (pp. 1170-1178).

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.

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).

Spratling, M. V. (2017). A review of predictive coding algorithms. *Brain and Cognition*, *112* (pp. 92-97).

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

Thornton, C. (2017). Predictive processing simplified: The infotropic machine. *Brain & Cognition*, *112* (pp. 13-24).

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.

Wiese, W. and Metzinger, T. (2017). Vanilla PP for Philosophers: A Primer on Predictive Processing. In Metzinger and Wiese (Eds.), *Philosophy and Predictive Processing: 1*, Frankfurt am Main: MIND Group.

Williams, D. (2018). Predictive Processing and the Representation Wars. *Minds & Machines*, *28* (pp. 141-172).

Yon, D., de Lange, F. P. and Press, C. (2019). The Predictive Brain as a Stubborn Scientist. *Trends in Cognitive Sciences*, *23*, No. 1 (pp. 6-8).

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