Brain and Cognition 112 (2017) 84–91

Contents lists available at ScienceDirect

Brain and Cognition

journal homepage: www.elsevier.com/locate/b&c

To be precise, the details don't matter: On predictive processing, precision, and level of detail of predictions ☆

Johan Kwisthout *, Harold Bekkering, Iris van Rooij

Radboud University, Donders Institute for Brain, Cognition and Behavior, Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 20 September 2015
Revised 15 February 2016
Accepted 21 February 2016
Available online 22 April 2016

Keywords:
Predictive processing
Precision
Level of detail
Structured representations
Formal modeling
Causal Bayesian networks

ABSTRACT

Many theoretical and empirical contributions to the Predictive Processing account emphasize the important role of precision modulation of prediction errors. Recently it has been proposed that the causal models used in human predictive processing are best formally modeled by categorical probability distributions. Crucially, such distributions assume a well-defined, discrete state space. In this paper we explore the consequences of this formalization. In particular we argue that the level of detail of generative models and predictions modulates prediction error. We show that both increasing the level of detail of the generative models and decreasing the level of detail of the predictions can be suitable mechanisms for lowering prediction errors. Both increase precision, yet come at the price of lowering the amount of information that can be gained by correct predictions. Our theoretical result establishes a key open empirical question to address: How does the brain optimize the trade-off between high precision and information gain when making its predictions?

1. Introduction

The predictive processing account has received widespread attention. Building on pioneering hierarchical predictive coding models of perception (Friston, 2005; Lee & Mumford, 2003; Rao & Ballard, 1999), recent literature proposes that the whole of perception, cognition, and action (Clark, 2013a) or even the entire operation of the brain (Hohwy, 2013) can be summarized by a simple, unifying principle. Rather than processing inputs in a mere bottom-up fashion, the brain is assumed to predict its inputs in a hierarchical manner by generative (causal) models and to process only that part of the input that is yet unexplained – the so-called prediction error. Sometimes prediction errors stem from the inherent stochastic nature of the world. To illustrate, take for instance, the observation of the outcome of a coin toss. We will have high confidence in our prediction that the coin will either land on heads or that it will land on tails, each event having a probability of 0.5; the observation of the actual outcome – while generating one bit of information – will normally not surprise us, as both events are fully consistent with our experience and knowledge of tossing a (fair) coin. One's generative models will therefore presumably not be changed as a consequence of this prediction error.

However, sometimes prediction errors are the result of an incomplete, immature, or just plain wrong generative model; think of trying an unknown dish in a restaurant or standing on skates for the first time. The uncertainty here is due to a lack of knowledge, and the prediction error *will* have impact: It allows the brain to update and improve its generative models (Payzan-LeNestour & Bossaerts, 2011; Yu & Dayan, 2005). These different roles of prediction errors, depending on the source of the uncertainty (*irreducible*, i.e., due to the inherent (known) stochastic nature of the world; or *reducible*, i.e., due to our lack of knowledge) are captured by the *precision* of the prediction error: A context-specific weighting of the prediction error that drives less or more attention to prediction errors. The net effect of the observation is thus a function of the precision of the prediction (capturing the uncertainty of the outcome) and the precision of the prediction error (capturing the model confidence).

Traditionally, computational operationalizations of the predictive processing account formulate the generative models (i.e., the stochastic relation between hypothesized causes and the predicted effects thereof) as Gaussian densities. Recently, however, Friston et al. (2015) propose to use *categorical* (discrete) probability distributions to describe the stochastic generative models that give rise to the predictions. An important distinction between Gaussian

☆ This research was funded by a NWO-TOP grant (407-11-040) awarded to Harold Bekkering and Iris van Rooij. We thank the participants of the Lorentz Workshop on Human Probabilistic Inference and two anonymous reviewers for helpful discussions and constructive comments on an earlier version of this paper.

* Corresponding author at: Donders Centre for Cognition, Montessorilaan 3, 6525 HR Nijmegen, The Netherlands.
E-mail address: j.kwisthout@donders.ru.nl (J. Kwisthout).

http://dx.doi.org/10.1016/j.bandc.2016.02.008
0278-2626/

densities and categorical probability distributions is that in the latter the *state space granularity* (how detailed are the generative models and the predictions that follow from them) is crucial. Whereas the amount of uncertainty (or *precision*) in a Gaussian density can be adequately described by its variance, the precision in a categorical distribution must be described by its *entropy* (Shannon, 1948), which is a function of both the state space granularity and the nonuniformity of the distribution (Kwisthout & Van Rooij, 2015).

Note that this state space granularity is context-dependent. Crucial in the coin-tossing example is that we describe the outcome of 'tossing a coin' in terms of the side of the coin to land on top, disregarding all other information in the outcome (such as the amount of rotation of the coin in the plane) as irrelevant. Compare this with throwing a regular die. As all sides of the die are equally likely to land on top, one can expect an odd number to fall just as often as an even number. When the outcome of a die is predicted in terms of whether the number will be odd or even (and the result interpreted likewise), the precision of that prediction is equal to the precision of tossing a coin. However, if the outcome of a die is predicted in terms of the number of pips, and the result interpreted likewise, the prediction is more uncertain – simply, because there are more possible events ('1', …, '6'; rather than 'odd' or 'even') and each event is equally likely – therefore, the prediction will have lower precision because a prediction was made (and the outcome interpreted) at a higher level of detail (Fig. 1). The precision of a prediction, hence, is indeed a function of both state space and nonuniformity.

Disentangling precision into level of detail and nonuniformity becomes necessary when cognitive (neuro) scientists aim to describe predictions and observable outcomes in terms of discrete, categorical events (Kwisthout & Van Rooij, 2015). Such outcomes may be the result of a coin flip (heads, tails) or of a die throw (odd, even; or 1…6, depending on how detailed our prediction is); they may describe the next action of a car in front of us (turn left, turn right, park, brake, or just keep driving), or of our spouse's emotions (sad, frustrated, happy, angry, bored; or whatever distinctions one makes); it may be a description of what one expects to see in a forest ('trees and other life forms'; or, when looking more closely, a chestnut tree, a squirrel, moss, bugs, etcetera). The appropriate level of detail that describes such outcomes is typically highly context-specific and depends on the epistemic and practical goals of the observing agent.

In this paper, we explore the computational and theoretical consequences of formalizing predictive processing in categorical probability distributions. After describing the predictive processing account more specifically, we introduce *level of detail* as a concept that intuitively captures the state space granularity, and together with the nonuniformity of the distribution describes its precision or entropy. We define the key computational processes in predictive processing in terms of (hierarchical, dynamical, multi-dimensional) causal Bayesian networks (Pearl, 2000). We show that manipulating the level of detail of generative models and/or predictions allows for the modulation of precision: For example, we can increase the precision of a prediction by decreasing the level of detail of the prediction. This, however, comes at the loss of information that can be gained by correct predictions. How this trade-off between predictions with high precision and predictions with high information gain is resolved in the brain is a key open theoretical (and empirical) question to address.

2. Predictive processing

The Predictive Processing (hereafter PP) account is becoming more and more popular as a unifying theory of what drives our cortical processes.[1] It encompasses key concepts such as the Bayesian brain (the brain encodes probability measures and balances prior expectations to sensory evidence according to the laws of probability theory, in particular Bayes' theorem; Knill & Pouget, 2004), the brain as prediction machine (the brain continuously makes predictions about future sensory evidence based on its current best model of the causes of such evidence; Dayan, Hinton, & Neal, 1995; Hohwy, 2007), the free energy principle (the brain minimizes overall expected prediction error as a proxy to minimize free energy; Friston, 2010) and the hierarchical organization of the brain (Friston, 2005, 2008). In particular it is claimed that the PP account applies to the entire cortex (Clark, 2013a) and that the same generic apparatus and mechanisms are used for both lower and higher cognition, e.g., both low-level vision and high-level intention attribution (Clark, 2013b; Kilner, Friston, & Frith, 2007; Koster-Hale & Saxe, 2013). However, to account for "higher cognitive phenomena such as thought, imagery, language, social cognition, and decision-making" there is still "plenty of work to do" (Hohwy, 2013, p. 3). In particular it is as yet unknown "What […] the local approximations to Bayesian reasoning look like as we depart further and further from the safe shores of basic perception and motor control? What new forms of representation are then required, and how do they behave in the context of the hierarchical predictive coding regime?" (Clark, 2013a, p. 201).

PP can be understood as a cascading hierarchy of increasingly abstract (e.g., in time scale or space) hypotheses about the world, where the predictions on one level of the hierarchy are identified with the hypotheses at the subordinate level. The inference process, i.e., inferring assumed causes from stimuli, is presumed to be facilitated by having predictions (stemming from the generative, top-down process) at each level of the hierarchy, comparing these predictions with the observed (or inferred) observations, and using the prediction error to update both the current hypothesis and to learn for future predictions.

For example, in the action understanding domain, the hierarchy can include the actual visual, auditory, tactile, or olfactory *inputs*, like a series of visual inputs, at the lowest level; one level above we may situate the *kinematics*, like a grasping movement of the hand, followed by the more abstract object-oriented *actions* (picking up a cup). Eventually, the hierarchy may include complex social cognitive constructs such as future *intentions*, social conventions, world knowledge, context etcetera (Kilner et al., 2007). However, the PP account is currently computationally fleshed out predominantly at the basic perception and motor control level (Clark, 2013a; see also Hohwy, 2013). In particular, computational implementations of PP (typically grouped under the denominator *hierarchical predictive coding* or HPC), such as those suggested by Rao and Ballard (1999), Lee and Mumford (2003), and Friston (2005, 2010), reside at that level.

In a probabilistic interpretation, making a prediction based on the current hypothesis in any of the assumed levels corresponds to computing a posterior probability distribution $P(Pred|Hyp)$ over a space of candidate predictions, given the current estimated distribution over a space of hypotheses.[2] Computing (the magnitude of) a prediction error corresponds to computing the relative entropy

[1] An illustration of this might be the observation that a separate outlet (Cleeremans & Edelman, 2013) was created to allow for the large number of commentaries to Clark's (2013a) target article in *Behavioral and Brain Sciences*. Also indicative is that a search on "predictive coding" and "predictive processing" on Google Scholar together found about 2500 papers published in 2014.

[2] There appears to be some ambiguity in the literature about whether a prediction (hypothesis) refers to a distribution over candidate predictions (hypotheses), or the *mode* of that distribution; see, e.g., Kilner et al. (2007, p. 161), Hohwy, Roepstorff, and Friston (2008, p. 691), and Hohwy (2013, p. 61) for examples that suggest the latter. In this paper we adhere to the view (e.g., Knill & Pouget, 2004; Lee & Mumford, 2003; Friston, 2009) that suggests that whole distributions (or approximations thereof) are maintained, without claiming that this debate has fully settled yet. In the remainder, unless explicitly noted, *hypothesis* refers to a probability distribution over a space of candidate hypotheses, and similar for predictions.

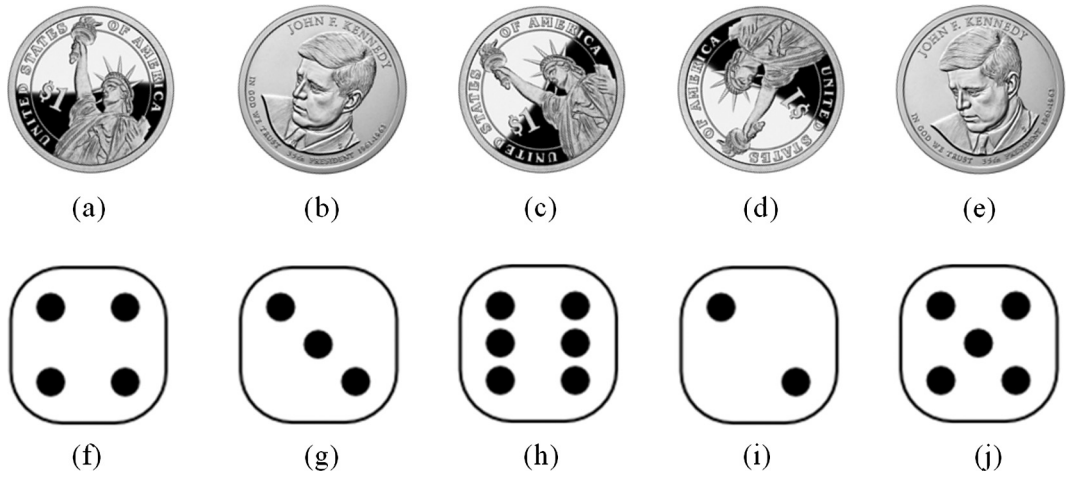

Fig. 1. We are typically inclined to see outcomes (a), (c), and (d) as belonging to the same category "Tails" and (b) and (e) to the category "Heads"; in contrast, outcomes (f) to (j) would normally each be categorized in a different category, even though we could group them by 'odd' and 'even' and come out with the same number of distinct 'events' as for the coin. The choice of the appropriate state granularity or level of detail decides how much relevant information is conveyed in each outcome, and correspondingly, what the entropy (or uncertainty) of the predicted outcome will be.

or Kullback–Leibler divergence $D_{KL}(Obs||Pred)$ between the actual and the predicted observation[3,4] (Friston, 2010). This prediction error is weighted with its *precision*; a parameter that reflects the amount of reducible uncertainty (or *estimation uncertainty*; Payzan-LeNestour & Bossaerts, 2011) in the current context. Which precision is warranted in a particular context depends on so-called *hyperpriors* defined as "prior beliefs about the precision of beliefs about the state of the world" (Friston, Lawson, & Frith, 2012, p. 1; see also Friston, 2005). The precision-weighted prediction error is used to bring prediction and observation closer to each other; either by revising the original hypothesis with an updated hypothesis (typically implemented by some form of gradient descent (Friston, 2002) or variational Bayes (Friston & Kiebel, 2009)), by revising the parameters of the generative model that generated the predictions (Friston, 2003), by obtaining additional information, i.e., by sampling the world (Friston, Adams, Perrinet, & Breakspear, 2012), or by manipulating the state of the world, i.e., by active inference (Brown, Friston, & Bestmann, 2011). Which strategy is employed depends on various aspects, for example the amount of reducible versus irreducible uncertainty in the environment (Yu & Dayan, 2005). On a longer time scale, prediction errors (or the relative absence thereof) also shape the generative models, including hyperpriors, to improve future predictions (Dayan, 2012).

The PP account has traditionally focused on visual perception (Rao & Ballard, 1999); for example, it has been used to explain binocular rivalry (Hohwy et al., 2008) and perceptual categorization (Kiebel, Von Kriegstein, Daunizeau, & Friston, 2009). One step further is the entanglement of perception and motor control, known as *active inference* (Brown et al., 2011; Friston, Daunizeau, Kilner, & Kiebel, 2010). Here, motor acts are seen as a consequence of proprioceptive prediction errors, i.e., motor acts *actively* change the sensory inputs in order to overcome errors between what is perceived and what is expected.

However, PP has been proposed for domains that do extend from Clark's "safe shores of basic perception and motor control". For example, Kilner et al. (2007) proposed predictive processing as a mechanism for action understanding, suggesting that the mirror neuron system plays a role in making hierarchical predictions of other's actions. Brown and Brüne (2012) extended this idea to incorporate social interaction in a broader context, pointing at emerging evidence of a shared neural representation and internal models of other's actions that allows us to understand the other's goals and intentions. Koster-Hale and Saxe (2013) reviewed evidence for key signatures of the PP framework in brain areas that are associated with theory of mind, such as the superior temporal sulcus (STS), temporo-parietal junction (TPJ) and medial prefrontal cortex (mPFC).

Proposed applications of PP at such higher levels of cognitive processing, however, do not yet postulate the "new forms of representation" Clark called for. They tend to focus on more conceptual or verbal models (e.g., the *Dr. Jekyll and Mr. Hyde*-example in Kilner et al. (2007, p. 164)) or seek to find evidence for key characteristics of PP (such as the inhibition of predicted signals in particular brain areas like STS and TPJ (Koster-Hale & Saxe, 2013) and the anterior insular cortex (AIC; Seth, Suzuki, & Critchley, 2012), or a beta/gamma dissociation in the Granger-causal connectivity in the Action-Observation Network (AON; Van Pelt et al., in press). While no computational models at these higher levels have been proposed yet, Koster-Hale and Saxe (2013) indeed argued that such models "must be quite abstract, and include expectations that actions will be rational and efficient, and consistent with, for example, the individual's beliefs, personality traits, and social norms" (Koster-Hale & Saxe, 2013, p. 839).

In cognitive science it is well known that higher-order cognitive processing requires some form of *structured* representations (Dietrich & Markman, 2003; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Markman, 1999), i.e., representations that can encode not only lists of features or variables but also (higher-order) relations between them. Here, we propose a formal version of PP for cognitive processing using structure representations suited for the PP framework, specifically causal Bayesian networks (Pearl, 2000). Such networks allow for complex, structured, and categorical stochastic dependences between hypotheses and predictions at various levels of detail, as a required for cognitive processing (Griffiths et al., 2010). In the next section we will discuss how the basic concepts of the PP account can be defined in this computational framework.

[3] We make the assumption that the observation is a joint probability distribution over the prediction variables that may be deterministic. Our formalism allows for both uncertain and certain observations; the latter is represented with a deterministic probability distribution that assigns a probability of 1 to the observed joint value assignment of the joint distributions.

[4] The difference between the predicted and observed distribution is also referred to as the *complexity* in the model comparison literature, and is also known as *Bayesian surprise* (Itti & Baldi, 2009).

3. Predictive processing in causal Bayesian networks

Causal Bayesian networks are graphical models that represent stochastic causal relationships between discrete random variables. They extend 'ordinary' Bayesian networks in that the arcs in the graph reflect not merely bidirectional stochastic dependences in the probability distribution, but causal interactions between random variables. The formal theory is built on the notion of *interventions* in probability distributions (Pearl, 1988). An intervention $do(X = x)$ intervenes 'from outside the model' by forcing the random variable X to take on the value x; the effect on a variable Y, i.e., $P_{do(X=x)}(Y)$ is to be distinguished from the conditional probability $P(Y|X = x)$. The former represents the probability distribution of Y, given that X is (externally) forced to take on the value x; the latter represents the conditional probability distribution of Y if X were observed to be x. The distinction between these notions may be illustrated with the following example. Assume that a light switch operates a light bulb: Putting the light switch to the *On* position causes the light bulb to go on. If we *observe* the light bulb to be on, we can infer that the light switch is most probably in the *On* position (rather than there being a short circuit, etc.), and vice versa, if we observe the light switch to be in the *On* position, we can infer that the light bulb is most probably on. Intervention works a bit different. If we *intervene* in the system by putting the light switch on, we can again infer that the light bulb is most probably on; however, if we intervene by throwing a stone at the light bulb and scattering it, nothing can reasonably be inferred about the state of the light switch: interventions on effects do not influence causes.

In our implementation of PP, every level of the hierarchical model is depicted by a causal Bayesian network B, where the variables of this network are partitioned into hypothesis variables *Hyp* (jointly representing the set of working hypotheses in this level), prediction variables *Pred* (jointly representing the predictions that are based on these hypotheses), and intermediate variables *Int* that are neither hypothesis nor prediction variables but do influence the outcome of the predictive process. These intermediate variables may, e.g., represent contextual effects, lateral connections (e.g., between different modalities), or latent variables. The arcs in the network represent causal relations.[5] The network is *dynamic*, in the sense that the current probability distribution over the variables is not only statically dependent on the causal relationships, but also on dynamic interactions in time. For example, the current prediction depends not only on the current hypothesis, but also on the prediction one time slice ago. This dynamical process is represented by inter-time slice connections between (some of) the variables in each level of the hierarchy. We do not impose structural constraints on the network structure, other than that we require that all prediction variables within a (static) time-slice of the dynamic network are sinks (have no outgoing arcs) and all hypothesis variables within a time-slice are sources (have no incoming arcs). The prediction variables at each level are identified with the hypothesis variables at the subordinate level. The hierarchical structure is depicted in Fig. 2a, while an example level is illustrated in Fig. 2b.

Hypothesis, prediction, observation and prediction error are defined in our formalization as follows. A hypothesis is simply the prior joint probability distribution P_{Hyp} over the hypothesis variables *Hyp*. We define a prediction as the posterior probability distribution P_{Pred} over the prediction variables, an observation as a (possibly deterministic) probability distribution P_{Obs} over the prediction variables that corresponds to observed or inferred information about the state of these variables, the prediction error $\delta(P_{Pred}, P_{Obs})$ as the *net residual* of subtracting P_{Pred} *from* P_{Obs}, and the *size* of the prediction error as the Kullback–Leibler divergence $D_{KL}(P_{Obs}||P_{Pred})$ between the two distributions.[6]

So far, the mathematical formalizations of *prediction* and *prediction error* in our formalization are defined quite similarly as in the (non-structural) Gaussian models that characterize conventional predictive coding formalizations. A crucial distinction, however, is that in our formalization the precision of hypotheses and predictions is not defined by variance but by entropy, viz., the combination of level of detail and non-uniformity. In the next section we will describe both aspects of precision in our formalization.

4. Uncertainty and level of detail in Predictive Processing

A prediction in Predictive Processing always refers to a probability distribution over a set of candidate predictions. The precision of a prediction refers to the entropy of the distribution, that is, the amount of uncertainty, given a particular state space. The more precise a prediction is at a given state space, the lower the entropy. The prediction error depends on the precision of the prediction *and* on the actual observation: If a coin that is biased toward tails falls on heads nevertheless the prediction error is much larger than when it indeed falls on tails; the prediction error of a fair coin falling on either heads or tails would be in between the two extremes (Fig. 3).

Orthogonal to the precision of the prediction is the precision of the prediction error. Whereas the precision of the prediction is a measure of the amount of uncertainty with respect to the prediction, the precision of the prediction error is a measure of the *nature* of the uncertainty; that is, whether it is reducible (can be decreased by learning) or irreducible (is due to the inherent stochastic nature of the world). When we have confidence in the generative model that generated the prediction, all uncertainty in the prediction can be fully explained by the inherent stochastic nature of the many-to-many mapping between causes and effects. In contrast, if we do not have confidence in the model – for example, because we are still learning its statistical regularities – then part of the uncertainty in the prediction cannot be explained.

In contrast to the common use of the term 'precise', a *precise* prediction is not necessarily a *detailed* prediction, despite what the term 'precision' may suggest in everyday usage. For example, "the visual stimulus is a robin" is a prediction on a fairly high level of detail, whereas "the visual stimulus is a bird" is a prediction on a lower level of detail; both predictions can have a high or a low precision, depending on the entropy of these predictions. For example, typically we will recognize a bird when it flies in front of us, but we may be uncertain when we perceive fluttering in the visual periphery (Friston et al., 2012). Precision thus can be specified for predictions at various levels of detail, such as illustrated in Table 1.

Formally, we define the uncertainty of a prediction as the *entropy* (Shannon, 1948) of the distribution corresponding with that prediction, we define the nonuniformity of a prediction as the *relative Shannon redundancy* (Shannon, 1948) of the distribution,[7] and we define the level of detail of a prediction as the *state space granularity* (Kwisthout, 2013) of the distribution. The entropy $H(Pred)$ of a prediction *Pred* (described in *bits*) is $-\sum_{(x \in Pred)} P(x)\log_2 P(x)$, where x denotes a (concrete) candidate prediction from the probability distribution. The relative Shannon redundancy $R(Pred)$ of a prediction *Pred* is defined as $H(Pred)/\log_2 |\Omega(Pred)|$, where $\Omega(\cdot)$ describes a state space and $|\Omega(\cdot)|$ describes the *size* of that state

[5] While our theory is worked out for discrete causal Bayesian networks, it can be in principle generalized to other formalizations. For example, one can model non-causal relationships, e.g., stochastic co-occurrences between visual and auditory inputs, using so-called chain graphs (Lauritzen & Richardson, 2002).

[6] But see Thornton (2017) for a different characterization of prediction errors and their sizes.

[7] We are indebted to one of the reviewers for suggesting this measure to us.

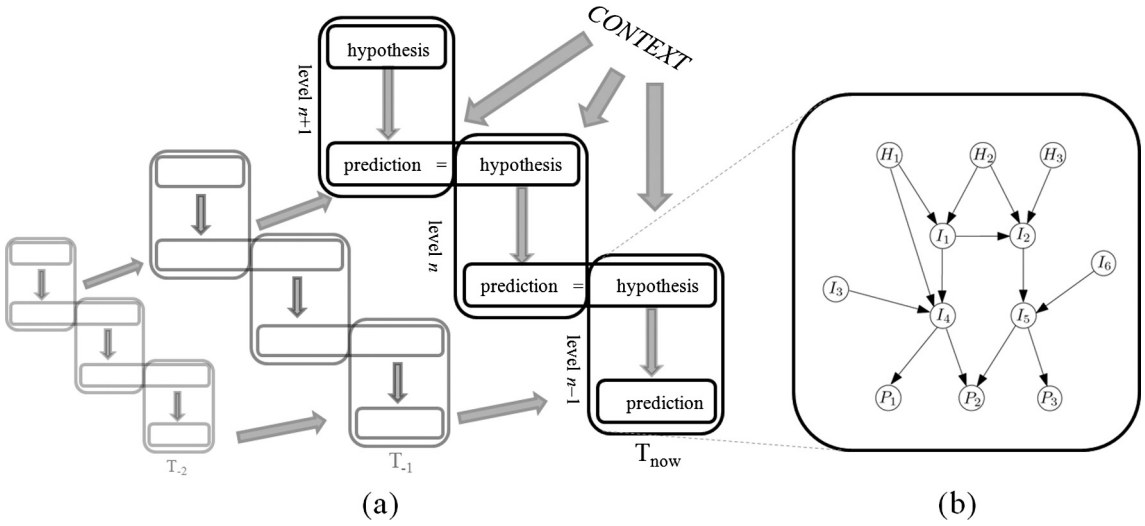

Fig. 2. (a) Illustrates the hierarchical and dynamical nature of Predictive Processing, where a prediction on level n is identified with a hypothesis at level $n-1$. When zooming in on a level (b), we see that the stochastic dependency of the prediction variables ($Pred = \{P_1, P_2, P_3\}$) on the hypothesis variables ($Hyp = \{H_1, H_2, H_3\}$) is mediated by intermediate variables ($Int = \{I_1, \ldots, I_6\}$).

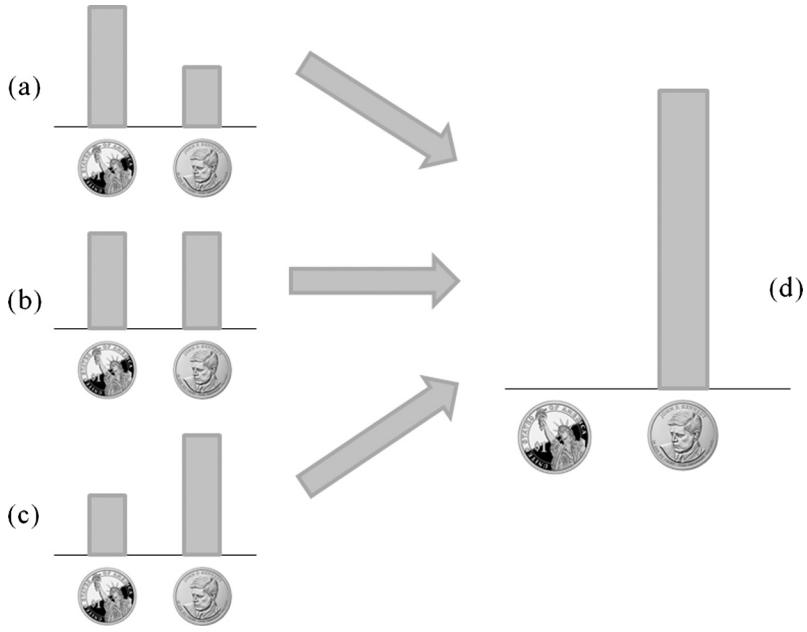

Fig. 3. Prediction errors will be higher when we expected the coin to fall on tails (a) compared to when we expected the coin to fall on heads (c). For uniform expectations (b), prediction error will be somewhat in between. Note that there is a non-zero prediction error in either of the three cases (a), (b), and (c) as there is always unpredicted information in the outcome.

Table 1
Illustration of probability distributions at a high and low level of detail, with high and low uncertainty.

	High level of detail	Low level of detail
High uncertainty	Pr(robin) = 0.20 Pr(nightingale) = 0.15 Pr(lark) = 0.19 Pr(brimstone) = 0.22 Pr(Peacock butterfly) = 0.24	Pr(bird) = 0.54 Pr(butterfly) = 0.46
Low uncertainty	Pr(robin) = 0.91 Pr(nightingale) = 0.01 Pr(lark) = 0.03 Pr(brimstone) = 0.02 Pr(Peacock butterfly) = 0.03	Pr(bird) = 0.95 Pr(butterfly) = 0.05

space. For example, the entropy of the predictions with various level of detail, such as depicted in Table 1 would be as follows: H(high detail, high uncertainty) = 2.30, H(high detail, low uncertainty) = 0.61, H(low detail, high uncertainty) = 1.00, H(low detail, low uncertainty) = 0.29. The corresponding relative Shannon redundancy is R(high detail, high uncertainty) = 0.99, R(high detail, low uncertainty) = 0.26, R(low detail, high uncertainty) = 1.00, R(low detail, low uncertainty) = 0.29; capturing that the nonuniformity of the distributions is similar in the high detail and in the low detail cases.

Level of detail, in contrast, is a measure on how *fine-grained* a probability distribution is. From a given distribution at the highest level of detail we can 'zoom out' by aggregating or clustering the values that the distribution can take. This can be done both for

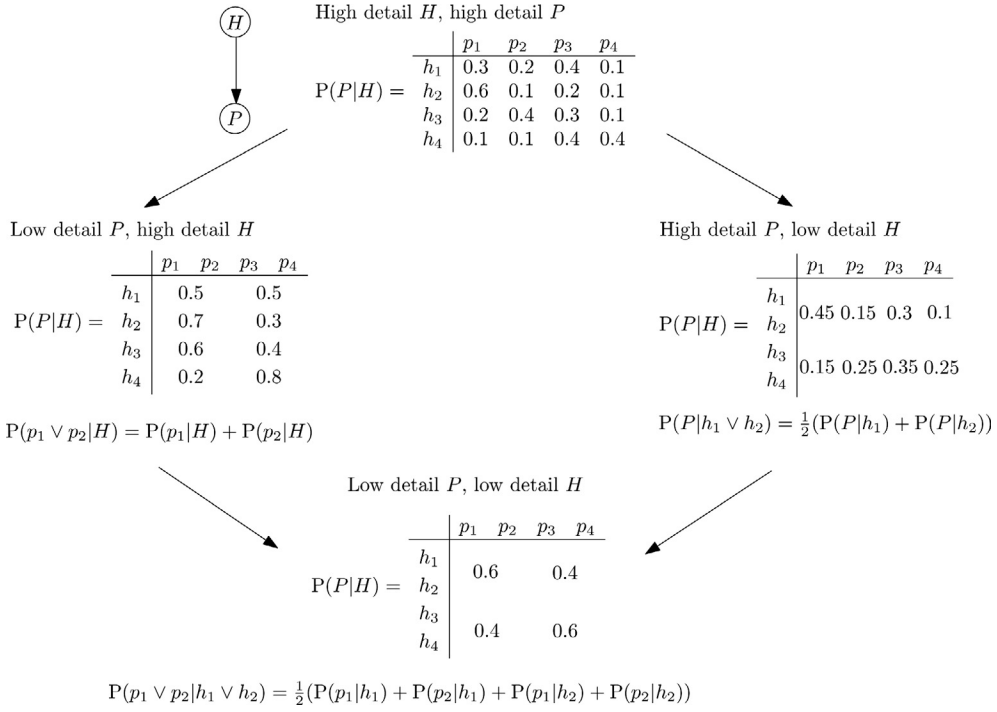

Fig. 4. When the level of detail of the hypothesis space decreases (i.e., the model of the causes becomes less fine-grained), the conditional probability of the prediction is averaged over the hypotheses that are aggregated. Likewise, if the level of detail of the prediction space decreases (i.e., the model of the predicted effects becomes less fine-grained), the conditional probability of the predictions that are aggregated is added.

hypotheses and for predictions (or observations) at any level of the hierarchy. Zooming in or out does not influence a particular hypothesis–prediction relation; it just describes this relation (including probabilistic dependencies), at a different level of abstraction. This is illustrated in Fig. 4 using a conditional probability table describing *P*(*Pred*|*Hyp*), i.e., the probability distribution over the prediction variables, given a particular distribution over the hypothesis variables for both high and low detailed hypothesis and prediction state spaces.

The precision of a prediction at a particular level of detail can be seen as a (statistical) property of the generative model at that level of detail. *Which* levels of detail are identified, and *what* level of detail is appropriate in a particular context, is described by a family of causal Bayesian networks, and a context-dependent hyperprior, respectively. Describing level of detail both for the hypothesis and prediction and partially ordering the networks in a lattice, such as in Fig. 4, gives us a mathematically elegant way of formulating the modulation of level of detail both on the hypothesis level and on the prediction level.

5. Dealing with prediction errors

In the face of a prediction error, the brain can achieve a decrease of this prediction error in a number of different ways (Friston, 2010; Friston et al., 2012; see also Kwisthout, 2014). In particular, the brain could revise the hypothesized causes; revise the causal model that generated the predictions from the hypothesized causes; it may lower prediction error by observation of latent variables, or by intervention in the world (i.e., acting). In addition, the brain may bring prediction and observation closer to each other by lowering the level of detail of both, or by increasing the level of the hypothesis space. In this section, we explore and formalize these different ways of dealing with prediction error in the context of the lattices of causal Bayesian networks we described in the previous section. These different ways are summarized in Fig. 5.

5.1. Revising the hypotheses

Prediction error can be lowered by *changing the distribution over the hypothesis variables* while keeping the model intact. It is typically appropriate in a situation where we encounter a situation that may be unexpected, but which is perfectly possible within our world model and no reason to update this model. This can be seen as a situation in which all uncertainty is irreducible (Yu & Dayan, 2005) – the environment's statistics are fully known. An example might be the situation where we throw three dice that all land with sixes on top. It may be rather unexpected, but it is nevertheless consistent with our generative model of the outcome of throwing dice. This hypothesis revision is formally defined as the computation of the posterior probability *P*(*Hyp*|*Pred*), that is, updating the prior probability over *Hyp* in the light of the (observed or inferred) distribution $P_{Obs}(Pred)$. Note that this distribution $P_{Obs}(Pred)$ is fully determined by the prediction $P_{Pred}(Pred)$ and the prediction error $\delta(P_{Pred}, P_{Obs})$ as the latter is defined as the result of subtracting P_{Pred} *from* P_{Obs}.

5.2. Revising the causal model

Prediction error can also be lowered by *changing the model or some of its parameter probabilities*. This would be appropriate in a situation where prediction error is caused by a change in the environmental properties, an apparent misrepresentation of the environment, or simply ignorance with respect to the environment. This situation can be seen as *unexpected uncertainty* respectively estimation uncertainty (Nassar, Wilson, Heasly, & Gold, 2010; Payzan-LeNestour & Bossaerts, 2011; Yu & Dayan, 2005). For example, we might expect two persons to shake hands when they greet each other, but we may learn a different cultural experience like a fist bump. This would require us to update the causal model by altering the stochastic dependences and/or introducing new (values of) variables. Formally, this corresponds to adding

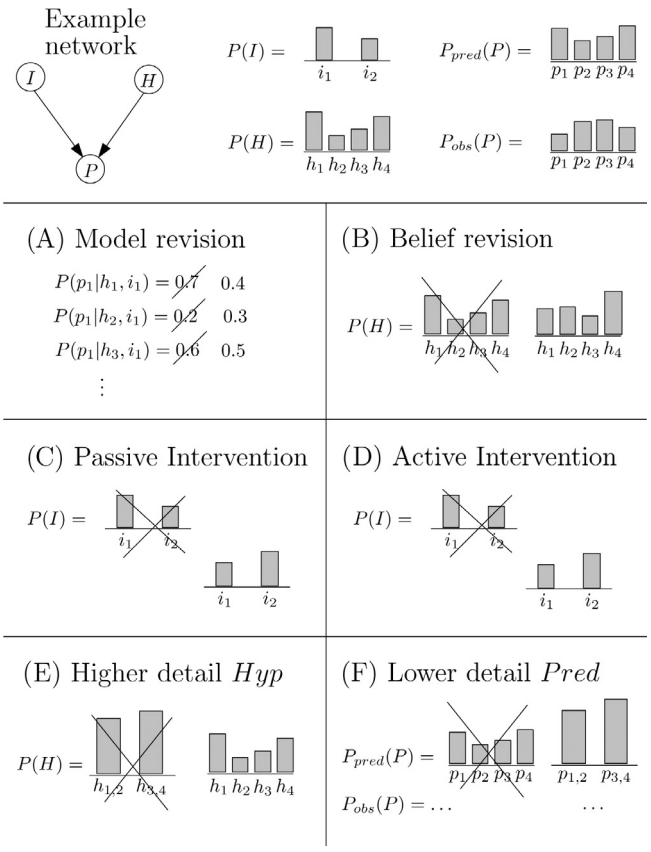

Fig. 5. Six ways of lowering prediction error: by revising the model (A), revision the priors on the hypothesis space (B), by observing (C) or setting (D) the value of intermediate variables, by increasing the detail of the hypothesis space (E) or by lowering the detail of the prediction space (F).

new variables to the Bayesian networks, increasing or decreasing the number of values that a variable can take, or updating the conditional probability distributions in the networks.

5.3. Evidence gathering

One may be able to reduce prediction error by gathering additional observations, or maybe reconsideration of observations we already have. A typical example could be when one is sitting in a train that is standing still at a railway station. When there is a train next to us at the opposite platform, and we observe movement, there is ambiguity whether *we* move or *the opposite train* moves. This can be seen as a prediction error as the (more or less) uniform hypothesis distribution leads to ambiguous or contradictory perceptual and vestibular predictions. A natural way of resolving this ambiguity is by seeking evidence that can discriminate between these hypotheses, in particular by looking at a stationary point, e.g., the railway station, in order to lower prediction error.

Technically, we can formalize this by seeking observations for previously unobserved intermediate variables in the network. For example, the singleton hypothesis node *Hyp* with uniformly distributed values h_1 and h_2 connects to both the prediction variable *Pred* with values p_1 and p_2 and the intermediate variable *Int* with values i_1 and i_2; $P(Pred = p_1|Hyp = h_1) = 1$, $(Pred = p_2|Hyp = h_2) = 1$, $P(Int = i_1|Hyp = h_1) = 1$, and $(Int = i_2|Hyp = h_2) = 1$. Now, $P(Pred)$ is uniformly distributed if $P(Hyp)$ is uniformly distributed, but the observation $Int = i_1$ will influence the distribution of *Pred* via the distribution of *Hyp*; $P(Pred = p_1|Int = i_1) = 1$, and likewise (but in the other direction) will the observation $Int = i_2$.

5.4. Intervention in the world

Prediction error between observation and prediction can be lowered by bringing prediction closer to match the observation, but also by intervening in the world, thus changing the actual inputs. The canonical example here is active inference (Brown et al., 2011): if there is a proprioceptive prediction error between the expected and actual position of one's limb, we engage in motor acts that intervene such as to bring the actual position closer to the expected position.

Technically this can be formalized by an *intervention* in the generative causal model: rather than (passively) observing the value a particular intermediate variable, we actively set it (by external intervention) to its desired value (Pearl, 2000).

5.5. Modulation of level of detail

Finally, we may be able to reduce prediction error by increasing the level of detail of the hypotheses or by lowering the level of detail of the predictions and actual observations. For example, when you observe me leaving the office with a coffee mug in my hand you may predict that I desire to get some coffee and that I will place my cup in the coffee machine. An observed movement toward the sink will lead to prediction error if you predicted a movement toward the coffee machine. One way of lowering the prediction error is by re-interpreting the prediction and observation in a less precise, more abstract manner, i.e., by expecting "activities related to coffee making (such as cleaning one's mug)" and interpreting the observations accordingly. Alternatively, you may increase the level of detail of the model, taking into account that I am holding a *filthy* coffee mug, yielding a different prediction.

Technically we can formalize this by changing the current causal Bayesian network in the family that corresponds to the current layer to a different network, either with a more fine-grained hypothesis space (to increase the level of detail of hypotheses) or with a less fine-grained prediction space (to decrease the level of detail of predictions).

5.6. Summary

The causal Bayesian network framework allows for the representation of complex relationships in generative models, including contextual influences, non-monotone relations, and related models with varying state space granularity. Prediction, precision, prediction error, and prediction error minimization can be elegantly described as computational processes that operate on these structures.

6. Conclusion and further work

In the previous sections we described how to formulate making predictions, computing prediction errors, and lowering prediction errors by various means in a causal Bayesian network formalization of PP. We thus proposed a concrete computational-level (Marr, 1982) characterization of the representations and processes crucial in the predictive processing account that allow this account to be explanatory when we focus on higher cognition, affording the PP framework to gain formal application beyond the scope of low-level perception and motor control. A crucial aspect of this formalization is *level of detail*, capturing the state space granularity of the generative models; the level of detail and the entropy of a distribution together describe its precision (Kwisthout & Van Rooij, 2015).

We identified six computational mechanisms for lowering prediction error. Four of them have been proposed in the literature before, but have not previously been formalized for categorical

probability distributions. Perceptual inference (Friston & Stephan, 2007) corresponds in our formalization with *hypothesis revision*, i.e., changing the current probability distribution over the hypothesis variables without altering the generative model. Active inference (Brown et al., 2011) corresponds to *active intervention*, i.e., intervening in the generative model by clamping some variables to their desired values. Sampling the world (Friston et al., 2012) corresponds to *passive intervention*, i.e., observing the values of some of the variables in the generative model. Finally, learning (Friston, 2003) is captured by *model revision*, i.e., by structurally 'rewiring' the generative model, adding new candidate hypotheses or updating the causal relationships. We thus have a tight coupling between psychological explanations in the predictive processing account and computational-level processing in our formalization.

Two other mechanisms have, to the best of our knowledge, as of yet not found their way into the neuroscience literature. However, the relevance of introducing level of detail of predictions in categorical distributions, and the modulation thereof as a means of lowering prediction error, has recently been recognized (Friston, 2015, in response to Kwisthout & Van Rooij, 2015). As we argued elsewhere, there is a trade-off between making predictions that are very likely to be correct (due to their generality) but carry little relevant information, and predictions that allow for much information gain (due to their specificity) but are likely to be incorrect (Kwisthout & Van Rooij, 2015).

Which mechanism actually will be applied when faced with a prediction error? How is this trade-off between information gain and expected prediction error resolved? These are questions that can only be answered by subsequent empirical investigations and theory forming. At this point, we can only speculate that, for example, model revision will be more dominant than hypothesis revision when the precision of the prediction error is high, that in early development predictions are made with low detail (due to immature generative models), and that the information gain vs. prediction error trade-off may fall under the regime of free energy minimization (cfm. Friston, 2015). It is beyond the scope of the current paper to address these questions. Instead we hope that a precise articulation of them, as afforded by our computational-level characterization of Predictive Processing, will form an impetus for new empirical research addressing these questions.

References

Brown, E. C., & Brüne, M. (2012). The role of prediction in social neuroscience. *Frontiers in Human Neuroscience, 6*, e147.

Brown, H., Friston, K., & Bestmann, S. (2011). Active inference, attention, and motor preparation. *Frontiers in Psychology, 2*, e218.

Cleeremans, A., & Edelman, S. (Eds.). (2013). Forethought as an evolutionary doorway to emotions and consciousness [Special issue]. *Frontiers in Theoretical and Philosophical Psychology, 4*.

Clark, A. (2013a). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*, 191–253.

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 Theoretical and Philosophical Psychology, 4*, e270.

Dayan, P. (2012). Twenty-Five lessons from computational neuromodulation. *Neuron, 76*(1), 240–256.

Dayan, P., Hinton, G. E., & Neal, R. M. (1995). The Helmholtz machine. *Neural Computations, 7*, 889–904.

Dietrich, E., & Markman, A. B. (2003). Discrete Thoughts: Why cognition must use discrete representations. *Mind and Language, 18*(1), 95–119.

Friston, K. J. (2002). Functional integration and inference in the brain. *Progress in Neurobiology, 590*, 1–31.

Friston, K. J. (2003). Learning and inference in the brain. *Neural Networks, 16*(9), 1325–1352.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences, 360*(1456), 815–836.

Friston, K. J. (2008). Hierarchical models in the brain. *PLoS Computational Biology, 4*(11), e1000211.

Friston, K. J. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Science, 13*(7), 293–301.

Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*, 127–138.

Friston, K. J. (2015). Response to commentaries: From complexity to epistemic emotions. *Cognitive Neuroscience, 6*(4), 225–227.

Friston, K. J., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology, 3*, 151.

Friston, K. J., Daunizeau, J., Kilner, J., & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics, 102*(3), 227–260.

Friston, K. J., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*, 1211–1221.

Friston, K. J., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience, 13*, 1–28.

Friston, K. J., & Stephan, K. (2007). Free energy and the brain. *Synthese, 159*(3), 417–458.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences, 14*(8), 357–364.

Hohwy, J. (2007). Functional integration and the mind. *Synthese, 159*(3), 315–328.

Hohwy, J. (2013). *The predictive mind*.Oxford, UK: Oxford University Press.

Hohwy, J., Roepstorff, A., & Friston, K. L. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition, 108*, 671–701.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research, 49*(10), 1295–1306.

Kiebel, J., Von Kriegstein, K., Daunizeau, J., & Friston, K. J. (2009). Recognizing sequences of sequences. *PLoS Computational Biology, 5*(11), e1000464.

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: An account of the mirror neuron system. *Cognitive Processing, 8*(3), 159–166.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *TRENDS in Neurosciences, 27*(12), 712–719.

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron, 79*, 836–848.

Kwisthout, J. (2013). Most inforbable explanations: Finding explanations in Bayesian networks that are both probable and informative. In L. C. van der Gaag (Ed.), *Proceedings of the 12th European conference on symbolic and quantitative approaches to reasoning with uncertainty. Lecture Notes in AI 7958* (pp. 328–339). Springer.

Kwisthout, J. (2014). Minimizing relative entropy in Hierarchical Predictive Coding. In L. C. van der Gaag & A. J. Feelders (Eds.), *Proceedings of the seventh European workshop on probabilistic graphical models. Lecture Notes in AI 8754* (pp. 254–270). Springer.

Kwisthout, J., & Van Rooij, I. (2015). Free energy minimization and information gain: The devil is in the details. Commentary on Friston, K., Rigoli, F., Ognibene, D., Mathys, C., FitzGerald, T., and Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience, 6*(4), 216–218.

Lauritzen, S. L., & Richardson, T. S. (2002). Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*, 321–348.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A, 20*(7), 1434–1448.

Markman, A. B. (1999). *Knowledge representation*.Mahwah, NJ: Lawrence Erlbaum Associates.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*.New York: Freeman.

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian Delta-rule model explains the dynamics of belief updating in a changing environment. *The Journal of Neuroscience, 30*(37), 12366–12378.

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Computational Biology, 7*(1), e1001048.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*.San Francisco, CA: Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, reasoning and inference*.Cambridge, MA: MIT Press.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79–87.

Seth, A. K., Suzuki, K., & Critchley, H. D. (2012). An interoceptive predictive coding model of conscious presence. *Frontiers in Consciousness Research, 2*, e395.

Shannon, C. R. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*, 379–423.

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

Van Pelt, S., Heil, L., Kwisthout, J., Ondobaka, S., van Rooij, I., & Bekkering, H. (2016). Beta- and gamma activity reflect predictive coding in the processing of causal events. *Social Cognitive and Affective Neuroscience* (in press). http://dx.doi.org/10.1093/scan/nsw017.

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron, 46*(4), 681–692.