mark.sprevak@ed.ac.uk

# Predictive coding I: Introduction

Mark Sprevak
*University of Edinburgh*

24 May 2021

## 1 Introduction

Predictive coding is a computational model of cognition. Like other computational models, it attempts to explain human thought and behaviour in terms of computations performed by our brains. It differs from more traditional approaches in at least three respects. First, it aspires to be *comprehensive*: it aims to explain, not just one domain of human cognition, but all of it – perception, motor control, decision making, planning, reasoning, attention, and so on. Second, it aims to *unify:* rather than explain cognition in terms of many different kinds of computation, it explains cognition by appeal to a single computation – one computational task and one computational algorithm are claimed to underlie all aspects of cognition. Third, it aims to be *complete*: it offers not just part of the story about the computation, but a model that stretches all the way from the details of neuromodulator release to abstract principles of rational action governing whole agents.

This is exciting stuff, however understanding precisely what predictive coding says, and whether it can achieve these ambitions, is not straightforward. For one thing, the term 'predictive coding' means different things to different people. For another, important features of the view, whatever its name, are liable to change or are underspecified in important respects. In this article and the three that follow it, my aim is to outline what predictive coding is, or aspires to be, and how it might fulfil these ambitions.

I claim that predictive coding should be understood as a loose alliance of three conceptually distinct claims. These claims, each of which may be precisified or

qualified in variety of ways, are made at Marr's *computational*, *algorithmic*, and *implementation* levels of description.[1] At Marr's computational level, predictive coding suggests that the *task* facing the brain in cognition is to minimise sensory prediction error. At the algorithmic level, predictive coding suggests that the *algorithm* by which our brains attempt to solve this task involves operation of a hierarchical artificial neural network of prediction and error units. This network may, in a further interpretative step, be understood as running a 'message passing' algorithm for approximate Bayesian inference. At Marr's implementation level, predictive coding suggests that some of the *physical resources* that implement this algorithm are located in the human neocortex: anatomically distinct areas in the neocortex implement functionally distinct layers of the abstract hierarchical artificial neural network and anatomically distinct cell populations inside each neocortical area implement prediction and error units.

Each of these claims is likely to be qualified in certain respects or supplemented by further details. Each needs to be stated more precisely and ideally associated with a quantitative mathematical formalisation. A path needs to be forged from the claims to supporting empirical evidence. Finally, one needs to show that the resultant model delivers the kinds of benefits originally promised – a comprehensive, unifying, and complete account of cognition. Different researchers within the predictive coding community have different opinions about how to do all this, and many of the details are simply currently left open. This means that the exact commitments of predictive coding are, to put it mildly, contentious. For all these reasons, it is more accurate to think of predictive coding as a *research programme* rather than as a mature theory that can be fully stated now. The aim of the research programme is to articulate and defend some sophisticated (likely heavily modified, precisified) descendent of the three claims above. As with any research programme, the merits of predictive coding should be judged in the round and, to some degree, prospectively: not just in terms of the raw predictive power and confirmation of what it says now, but also in terms of its potential future benefits, and its ability to inspire and guide fruitful future research.

Before saying what predictive coding is, it is helpful to first say what it is not. In this article, I outline five ideas that are often presented alongside predictive coding, but which should be carefully distinguished from predictive coding. In the three articles that follow, I focus on the positive content of the view. These explore predictive coding's claims at Marr's computational, algorithmic, and implementation levels respectively (Sprevak, forthcoming[a]; Sprevak, forthcoming[b]; Sprevak, forthcoming[c]). My strategy is to present what, in my opinion, are the 'bare bones' of the approach. As we will see, there are many ways in which these basic ideas may

[1] See Marr (1982), Ch. 1 for a description of these three levels.

be subsequently elaborated and refined. For readers new to this topic, I hope that this will provide you with a scaffold on which to drape your future, more nuanced understanding of the literature on predictive coding. Sometimes starting with a basic sketch is the best way to convey what is going on in a complex scene.

In the remainder of this article, I focus on five ideas that feature prominently in many expositions of predictive coding, but which are *not* predictive coding. The ideas are: (i) the brain employs an efficient coding scheme; (ii) cognition contains many top-down, expectation-driven effects; (iii) cognition involves minimising prediction error; (iv) cognition is a form of probabilistic inference; (v) cognition employs generative models. I argue that, while these ideas are used by predictive coding, they do not reflect what is unique about that research programme. They are shared by a wide variety of alternative computational approaches to cognition that have little otherwise in common with predictive coding. If one wishes to know what is special about predictive coding, then these ideas, whatever their value, can function as a potential distractors. An important corollary is that empirical evidence for predictive coding does not necessarily flow from the empirical evidence that supports these more general ideas. Empirical evidence for predictive coding should aim to *selectively* support predictive coding with respect to plausible contemporary rivals, not to confirm ideas that are shared by a wide variety of alternative approaches.

In both the present article and those that follow, I consider predictive coding only as a proposed model of *subpersonal* cognitive processing. I do not examine how predictive coding's computational model might be adapted, extended, or otherwise applied to describe personal-level thinking or conscious experience. Modelling conscious experience with predictive coding is a relatively new development that is gaining traction among philosophers. But it is a project that assumes we have a prior understanding of what predictive coding's computational model is. That prior question is the focus of this review.[2]

As already mentioned, some authors use the term 'predictive coding' to refer to only one aspect of the view: for example, the 'efficient coding' strategy of Section 2, or the artificial neural network described in Sprevak (forthcoming[b]). Likewise, some authors call the larger overarching research programme not 'predictive coding', but 'predictive processing', 'prediction error minimisation', or 'free energy minimisation'. In what follows, I choose to use the term 'predictive coding' to refer to the entire research programme, and I disambiguate alternative usages along the way. Readers should feel free to substitute alternative terms as they please.

[2]For examples of work that applies predictive coding's computational model to personal-level and conscious experience, see Clark (2019); Hohwy (2012); Kirchhoff and Kiverstein (2019); Seth (2017).

## 2 Efficient neural coding

An important idea that predictive coding employs is that the brain's coding scheme for storing and transmitting sensory information is efficient. Neural activity, or rather certain aspects of it, should be seen through the lens of *compressing information* (and information should be understood in terms of Shannon information theory). To compress Shannon information, a system should aim to transmit only what is 'new' or 'unexpected' or 'unpredicted' relative to its expectations. If the brain embodies certain assumptions about its incoming sensory data, these could allow it to predict certain 'bits' of that incoming sensory stream. This means that fewer bits would need to be stored or transmitted inwards from the sensory periphery, yielding a gain in efficiency in storing and transmitting sensory information from the sensory organs to the rest of the brain. The more accurately the brain's assumptions reflect its incoming sensory stream, the less information would need to be stored or transmitted inwards from the periphery. All that needs to be sent is the error with respect to its predictions. The same idea underlies efficient coding schemes used on electronic computers for storing and transmitting images and movies across the Internet (e.g. JPEG or MPEG).

The notion that our brains use a coding scheme that is efficient in this respect dates back at least to the work of Attneave (1954) and Barlow (1961). They argued that the brain uses a compressing, 'redundancy reducing' code for sensory data partly based on the grounds that neurons in the early visual system have very limited physical dynamic range: the bits they store or transmit are precious and should not be squandered.[3] Predictive coding employs the same basic idea, but elevates it to a universal design principle that governs all aspects of cognition and neural functioning. According to predictive coding, the brain is ruled by a single imperative – to maximally compress its incoming sensory stream. To this, predictive coding adds a range of assumptions about (i) the particular algorithm and representational scheme by which the incoming sensory signals are predicted and compressed; (ii) how the assumptions used for sensory compression are modified during learning; (iii) where physically in the brain all this takes place.

The idea that brains code sensory information 'efficiently' is not unique to predictive coding. For one thing, predictive coding has rather specific views about how its sensory compression scheme operates – see (i)–(iii) above. For another, predictive coding holds the rather extreme position that redundancy reduction for sensory data is the brain's *only* goal. As Barlow made clear in his later work, even if one thinks that sensory compression is one thing the brain does, it is not obvious that it

[3] See Simoncelli and Olshausen (2001); Sterling and Laughlin (2015); Stone (2018) for reviews of the contemporary literature on efficient coding in the sensory system.

is the only thing. In some circumstances, it may pay the brain *not* to compress:

> The point Attneave and I failed to appreciate is that the best way to code information depends enormously on the use that is to be made of it … if you simply want to transmit information to another location, then redundancy-reducing codes economizing channel capacity are what you need … But the brain is not just a communication system, and we now need to survey cases where compression is not the best way to exploit statistical structure. (Barlow, 2001, p. 246).

One can appreciate Barlow's point by considering what is the most efficient coding scheme for image data on an electronic computer. What counts as the most efficient coding scheme depends, not just on how many bits one would save during storage or transmission of the image, but also on what one wishes to do with the image data. If all one wishes to do is transmit the image across a low-bandwidth channel (e.g. a slow Internet connection), then compressing it using a scheme like JPEG makes sense, since it reduces the amount of data you need to transport. Similarly, if all one wishes is to store the image on a small hard drive, then JPEG may be a good scheme, since it minimises how much storage would be used.[4] However, if you want to *transform* the image or *perform an inference* over it, then a scheme like JPEG may not be the best or most efficient coding system. Compressed data are often harder to work with. If you ask your computer to rotate an image 23 degrees clockwise, the computer will generally not do this with the compressed data. Instead, it will switch to an uncompressed version of the image (a two-dimensional array of RGB values at X, Y pixel locations). Image processing algorithms defined over uncompressed data tend to be shorter, simpler, and faster than those over their compressed counterparts.[5] Uncompressed images have extra structure, and that structure can make the job of an algorithm that operates on them easier, even if it adds extra overhead to store or transmit.[6]

If the only costs that matter to the brain are the storage and transmission costs of incoming sensory data, then it may make sense for the brain to exclusively aim

[4] Other coding schemes are better than JPEG at redundancy reduction. Wavelet-based codes (Usevitch, 2001) and deep neural networks (Toderici et al., 2016) can both outperform it. Notably however, these schemes tend to impose higher processing burdens during decoding for inference or transformation of an image.

[5] This is an instance of a more general trade-off in computer science between saving time and space. Compressing data saves space, but generally has an adverse effect on the time (computing cycles) required to manipulate that data for many other tasks. You would have experienced this trade-off any time you waited for a '.zip' archive to uncompress before being able to work on its contents.

[6] A related point is that uncompressed data are more resistant to noise during storage and transmission.

to compress that incoming sensory data. However, if other considerations matter – e.g. speed, simplicity, and ease in inference – then it may make sense to add or preserve redundant structure in incoming sensory data.[7] Optimising for redundancy reduction of sensory data is not the only possible objective for an cognitive system that seeks efficiency.

It is not uncommon for contemporary work on the 'efficient coding' hypothesis in computational cognitive science to acknowledge this point.[8] Predictive coding has rather strong and unusual views about this: it equates efficient coding with sensory redundancy reduction, it claims that the entire brain (not just certain areas in the sensory cortex) is devoted to this sensory compression, and it claims that this is accomplished by a specific algorithm and representational scheme. The basic idea that *one* of the things that the brain does is sensory compression is, however, not unique to predictive coding.

## 3 Top-down, expectation-driven effects in perception

Top-down, expectation-driven effects in perception are cases in which what an agent 'thinks' systematically affects what they 'perceive'. Top-down, expectation-driven effects are often presented as part and parcel of predictive coding. Predictive coding's model is thought to imply that human perception is fundamentally top-down or expectation-laden: 'What we perceive (or think we perceive) is heavily determined by what we know' (Clark, 2011). Experimental evidence for top-down, expectation-driven effects is sometimes presented as evidence that supports predictive coding. The reasoning here appears to be that if predictive coding is true, then top-down, expectation-driven effects are to be expected; they can be predicted and explained in terms of the two-way flow of prediction and error signals inside predictive coding's computational model.[9]

However, the relationship between predictive coding and top-down, expectation-driven effects in perception is complex.

For one thing, top-down effects are normally identified with a relationship between an agent's *personal-level* states: what an agent *believes* affects their perceptual *experience*.[10] Predictive coding is, at least in the first instance, a claim about the agent's

[7] Gardner-Medwin and Barlow (2001) list examples in which adding redundancy to sensory signals increases the chances of fast and reliable inference over sensory data.

[8] For example, Simoncelli and Olshausen (2001) suggest that the details of the task a cognitive system currently faces, and not the mere imperative for redundancy reduction, should be considered when calculating the efficiency of a coding scheme (p. 1210).

[9] For examples, see Clark (2013), p. 190; Lupyan (2015).

[10] See Macpherson (2012); Firestone and Scholl (2016).

subpersonal computational processes. The 'top' and 'bottom' in predictive coding, as we will see, refer to subpersonal computational states. Predictive coding proposes that 'high-level' neural representations (implemented deep in the cortical hierarchy) have a 'top-down' influence on 'low-level' representations (implemented in the early sensory system).[11] How this kind of subpersonal top-down effect relates to personal-level top-down effects observed in psychology is presently unclear.

Plausibly, the existence of any personal-level top-down effects requires *some* information flow at the subpersonal level from high-level cognitive centres to low-level sensory systems. However, only a tiny fraction of the top-down subpersonal information processing posited by predictive coding is reflected in the contents of personal-level belief or perceptual experience. For predictive coding to say something specific about the existence or character of personal-level top-down effects, it would need to say *which* aspects of that subpersonal information flow give rise to *which* personal-level states (beliefs and perceptual contents). These assumptions are not currently to be found in predictive coding's core computational model. Ideas about these have been proposed, but exactly how subpersonal states in the computational model map onto personal-level beliefs and perceptual experiences remains a highly speculative matter. Absent confidence in such assumptions however, it is simply unclear how predictive coding's computational architecture bears, or if it bears at all, on personal-level top-down effects observed in perception.[12]

A separate issue complicating the relationship between predictive coding and top-down effects is that positing top-down subpersonal information flow in a computational model is not a feature that is *unique* to predictive coding. Indeed, almost any computational model of cognition will involve information flowing downwards from high-level cognitive centres to low-level sensory systems. There is an obvious need for top-down subpersonal influence to account for the action of endogenous attention and to explain how and why certain sensory processes get suppressed or enhanced based on the agent's background knowledge and assumptions.[13] A further commonly cited consideration is the huge number of descending neural

[11] Sprevak (forthcoming[b]), Section 3; Sprevak (forthcoming[c]), Section 2.

[12] See Macpherson (2017); Drayson (2017) for in-depth discussion of this line of argument. They suggest – for reasons similar to those indicated here – that predictive coding's computational model is compatible with *no* personal-level top-down effects occurring at all.

[13] As Ira Hyman observes in his introduction to the reprinting of Neisser's classic 1967 textbook: 'Cognitive psychology has been and always will be an interaction of bottom-up and top-down influences' (Neisser, 2014, p. xvi). See Firestone and Scholl (2016), p. 14, where despite alternative explanations being sought, attention is introduced as an unavoidable source of subpersonal top-down influence. See Gregory (1997); Poeppel and Bever (2010); Yuille and Kersten (2006) for appeal to subpersonal top-down influences to explain how the brain resolves ambiguities in its incoming sensory data, how it handles noise, the persistence of knowledge-based perceptual illusions, and semantic priming effects.

pathways in the mammalian brain that carry information in the cerebral cortex backwards from higher cognitive areas to lower sensory areas. Hypotheses abound about the computational function of these neural backwards connections. Even if one were to disregard them, there are plenty of other routes by which information in high-level cognition is likely to systemically influence low-level sensory systems. As Firestone and Scholl (2016) point out, the decision to "shut one's eyes", which causes one's eyelids to close, has an evident effect one's sensory input, thereby opening up an information-bearing channel (looping out into the world) via which high-level representations can influence low-level sensations.[14] Finally, even so-called feedforward computational models – e.g. the account of the early visual system proposed by Marr (1982), Ch. 3 – are normally qualified with a rider that, *of course*, additional top-down, expectation-driven subpersonal influences exist, even if they have been omitted from the model for the sake of simplicity.[15]

When suggesting that it can account for personal-level top-down, expectation-driven effects in perception, the case predictive coding has to make is why its specific set of top-down computational pathways is *uniquely* or *best* suited to explain personal-level effects. There are a vast number of alternative computational architectures that allow for some degree and manner of subpersonal top-down influence. These include endless varieties of artificial neural network and classical, symbolic architectures that contain loops. It is presently unclear why predictive coding's proposal is the best one to explain personal-level top-down effects in psychology. To be clear, predictive coding *allows* for top-down effects in perception to occur; it is also broadly *suggestive* that such effects will occur. What is not clear is that it is better suited to account for these effects than any number of alternative computational models. For these reasons, it is not clear how empirical evidence of personal-level top-down selectively confirms predictive coding.

## 4 Minimising prediction error

Minimising prediction error is a common objective in modern artificial intelligence and machine learning. Researchers often define learning tasks or inference tasks as problems of minimising prediction error (about reward or other kinds of data). A computational system that learns tries to tweak the parameters of its mathematical

[14]Dennett (1991) argues that these kind of top-down 'virtual wires' can produce extremely sophisticated forms of information processing, including those that are characteristic of high-level human thought and reasoning (pp. 193–199).

[15]See Marr (1982), pp. 100–101: '… top-down information is *sometimes used and necessary* … The interpretation of some images involves more complex factors as well as more straightforward visual skills. This image [a black-and-white picture of a Dalmatian] devised by R. C. James may be one example. Such images are not considered here.' (emphasis mine)

model to fit or predict its data. A computational system that performs inference often tries to make predictions that will minimise error or be as close to reality ('ground truth') as possible.[16] Different types of computational system differ in the kinds of data they try to predict, the mathematical model they use, and the methods they use to fit their model or perform inference.[17] Prediction error can also be measured in many different ways. A commonly used measure is the mean value of the squared difference between predictions and the actual data – 'mean-squared error'. Many computational systems change their mathematical models or their variables to minimise the magnitude of their mean-squared errors.[18]

The space of possible computational models that attempt to minimise prediction error is vast. You can get some idea of its size and diversity by opening up any current textbook on machine learning or statistics.[19] A relatively simple example of such a model is one that performs regression. Regression is an attempt to fit a polynomial function of a certain degree – a smooth curve – to observed data and use that curve to make predictions about unobserved cases. Classical regression techniques in statistics tell the computational agent how to find such polynomial functions. The simplest version of this method is linear regression, which uses a straight line as its model of the data (a polynomial of degree 1). Minimising prediction error reduces to the task of finding the value of the two numerical parameters (slope and y-intercept) that define a straight line that minimises the mean-squared error in predicting known data.

Deep neural networks offer examples of much richer and more complicated mathematical models that also aim to minimise prediction error. The predictions generated by a deep neural network may involve stringing together a huge sequence of mathematical operation with many variables. Learning for these models consists in finding the values, not of just two, but of millions of parameters that minimise the model's prediction error. Learning techniques for deep neural networks, e.g. various versions of backpropagation, attempt to iteratively modify the model's many parameters to produce a model that does better at minimising prediction error.

[16] Bishop (2006), pp. 1–12 and Hohwy (2013), pp. 42–46.

[17] Note that a prediction is not necessarily about the future. It is an *estimate* of some data the computational system has not already observed. A prediction in this sense might well concern past or present unobserved data. For a helpful review of the relevant sense of 'prediction', see de Lange, Heilbron and Kok (2018), p. 766, Box 2.

[18] More accurately, they aim to minimise a *cost function*, which describes the overall cost of a prediction and of which prediction error is one component. A common cost function is the prediction error plus the sum of the squares of all the model's parameters. The latter serves as a 'regularisation' that penalises (increases the cost of) learning more complex models. For an introduction to prediction errors, regularisation, and cost functions, see Russell and Norvig (2010), pp. 709–713.

[19] For example, Bishop (2006); MacKay (2003); Barber (2012); Matloff (2017).

Predictive coding says that the brain aims to minimise prediction error. What distinguishes predictive coding from other contemporary approaches is that it makes distinctive claims about the *data*, *model*, and *algorithm* used to minimise prediction error; a further point of difference is that predictive coding makes a special claim about the *role* of minimising prediction error in the brain's overall cognitive economy.

The *data* that the brain attempts to predict are, according to predictive coding, the brain's sensory signals. Predictive coding claims that the brain aims to minimise its prediction error over (a weighted average of) its incoming sensory signals.[20] This should be distinguished from alternative hypotheses about data over which predictions are made, such as the claim that the brain attempts to minimise prediction error over its *reward* signals.[21] The *model* the brain uses to generate its predictions consists in a deep artificial neural network with a specific topology and pattern of connections between prediction and error units. This artificial neural network is organised in a way that is quite unlike those commonly found in modern commercial applications of deep learning and artificial intelligence. The *algorithm* predictive coding ascribes to the brain for finding the right parameters of this artificial neural network is also unusual. Unlike in most commercial applications of deep learning – which rely on some version of backpropagation – predictive coding proposes that learning occurs via Hebbian learning.[22] A special *role* is also accorded to prediction error minimisation. Predictive coding claims that minimising sensory prediction error is not just one objective among many that the brain faces, but its *only* goal. Minimising prediction error should be understood as the objective of all aspects of cognition (and not just, say, something that it does in perceptual learning or classification).

Many contemporary computational models of cognition advert to the notion of minimising prediction error. What marks out predictive coding as special is that it proposes that the brain uses a specific dataset, a specific mathematical model, a specific algorithm, and it accords this task a special role in cognition. Evidence that the brain contains prediction errors or that it is sometimes engaged in the task of minimising prediction error, even if it is compatible with what predictive coding says, is liable to also fit many other approaches. These might also posit prediction errors, but minimise them in different ways, or not grant them a universal role in

[20] Sprevak (forthcoming[a]), Sect XX.

[21] For example, see Schultz, Dayan and Montague (1997); Niv and Schoenbaum (2008). The relationship between minimising *reward* prediction error and minimising *sensory* prediction error is an active area of research and not yet fully understood. See Friston et al. (2013); Schwartenbeck et al. (2015) for an attempt to redescribe the task of minimising reward prediction error as that of minimising sensory prediction error.

[22] Sprevak (forthcoming[b]), Sect XX.

governing all aspects of cognition. In order to selectively confirm predictive coding, one needs further details about the nature, role, and function of these prediction error signals in the brain – Are they exclusively sensory predictions? How were they created? How would they revised? What is their role across different cognitive tasks?

## 5 Cognition as a form of probabilistic inference

Cognitive systems receive noisy, incomplete, and sometimes contradictory information from the world. They need to weigh this information rapidly and efficiently, and integrate it with (perhaps conflicting) background knowledge in order to reach a decision or to generate behaviour. Probabilistic computational models have been widely adopted in computational cognitive science to help shed light on this.

According to these models, brains represent multiple incompatible possibilities (e.g. 'the person facing me is my father, my uncle, his cousin, …') along with some measure of uncertainty about those various outcomes. Subsequent steps in the cognitive processing will take each of these different possibilities into account, weighted somehow by the cognitive system's degree of uncertainty. The essential idea is that the brain does not 'put all its money' on one outcome at any given moment, but rather keeps track of many possibilities, along with its degree of uncertainty about them.

Computational models tend to develop this idea by ascribing mathematical *subjective probability distributions* to cognitive systems. These subjective probability distributions describe the cognitive system's degree of uncertainty or confidence across many possible outcomes. Cognitive processes are modelled as a series of elementary steps in which one subjective probability distribution conditions, or updates, another. Cognition might maintain this probabilistic character right until the moment the cognitive system is forced to plump for a specific outcome in action (e.g. the agent is required to respond 'yes'/'no' in a forced-choice task). The probabilistic rules that govern this processing – the steps by which subjective probabilities are combined or updated in the brain – may vary between different approaches. It is important to note however, that the subjective probabilities that are ascribed may not be personal-level states of the whole agent (e.g. as in the classical degrees of belief discussed by de Finetti, 1990; Ramsey, 1990). Subjective probability distributions are often ascribed to subpersonal parts of the agent (e.g. cognitive subsystems, neural regions, cell populations, or even individual neurons) (for example, see Deneve,

2008; Pouget, Dayan and Zemel, 2003).[23]

A particularly influential example of this approach is the so-called 'Bayesian brain' hypothesis. This suggests that Bayes' rule, or some approximation to it, describes the rules by which the human brain combines and updates its subjective probability distributions. Exact Bayesian inference is computationally costly, so most advocates of the Bayesian brain hypothesis believe that the brain performs some version of *approximate Bayesian inference*.[24] There are a vast range of algorithms to choose from here. Bayesian algorithms tend to fall into two camps: *sampling methods* (which aim to follow the trajectory of multiple categorical samples through inference to create a posterior empirical distribution that approximates the true Bayesian posterior) and *variational methods* (which use simplified, less computationally demanding subjective probability distributions, varying their parameters to find an analytical result that is close to the true Bayesian posterior). Both approaches for approximating Bayesian inference are common in modern artificial intelligence and machine learning.[25] Advocates of the Bayesian brain hypothesis do not agree about whether the brain uses a sampling method or a variational method.[26]

Predictive coding is one example of a probabilistic model of cognition and an instance of the Bayesian brain hypothesis. Predictive coding identifies the task the brain faces in cognition is that of minimising sensory prediction error. If combined with appropriate simplifying assumptions, this task can be shown to entail approximate Bayesian inference.[27] The numerical values that feature in predictive coding's artificial neural network can be interpreted as parameters of subjective probability distributions (namely, as the means and variances of Gaussian distributions). Predictive coding's algorithm can be interpreted as a particular version of variational Bayesian inference.[28] Predictive coding proposes that these numerical parameters, hence the subjective probability distributions manipulated in cognition, are encoded in the average firing rates of neural populations of layers in the neocortex, and the

[23] Against this, Rahnev (2017) suggests that brains do not store full subjective probability distributions, but instead only store a small number of samples or summary statistics (e.g. the mean and variance of some distribution). Colombo, Elkin and Hartmann (2018) review a number of other non-probabilistic ways in which the brain might encode uncertainty.

[24] Chater and Oaksford (2008); Knill and Pouget (2004).

[25] For an introduction to sampling methods (e.g. Markov chain Monte Carlo methods or particle filtering), see Bishop (2006), Ch. 11. For an introduction to variational methods, see Bishop (2006), Ch. 10.

[26] For proposals that the brain uses sampling, see Fiser et al. (2010); Griffiths, Vul and Sanborn (2012); Hoyer and Hyvärinen (2003); Moreno-Bote, Knill and Pouget (2011); Sanborn and Chater (2016); Sanborn and Chater (2017). Predictive coding is an example of a proposal that the brain uses a variational method to approximate Bayesian inference.

[27] Sprevak (forthcoming[a]), Section XX; Sprevak (forthcoming[d]), Section XX.

[28] Sprevak (forthcoming[b]), Section XX.

manner in which these subjective probability distributions condition one another in inference is encoded in the strength of the synaptic connections between distinct neocortical areas.[29]

Someone might endorse the idea that the brain engages in probabilistic inference, or even the Bayesian brain hypothesis, but reject some or all of predictive coding's specific assumptions about how all of this works. They might, for example, not accept that *all* aspects of human cognition involve Bayesian inference, or that *every* aspect of cognition involves inference over the *same* probabilistic model, or that the subjective probability distributions are always *Gaussian*, or that the brain's rules for manipulating these distributions are predictive coding's specific version of *variational* Bayes, or that average firing rates in neocortical layers encode the parameters of the brain's subjective probability distributions.[30] The space of possible computational models that treat cognition as involving some kind of probabilistic inference is vast. Evidence in favour of a probabilistic approach to cognition cannot straightforwardly be treated as evidence that confirms predictive coding as opposed to any number of other views.

## 6 Cognition uses a generative model

A generative model is a special kind of representation that describes how observations are produced by unobserved ('latent') variables in the world. If a generative model were supplied with the information that your best friend enters the room, it might tell you about which sights, sounds, smells you would experience. There is a growing acceptance in computational cognitive science (and AI) that generative models – and in particular, *probabilistic* generative models – are likely to play an important role in cognition. This is for at least three (interrelated) reasons.

First, a generative model would help a cognitive system solve the problem of distinguishing between changes in its sensory data that are *self-generated* and *externally generated*. When our eyes move, the pattern of light projected onto our retinas changes. How does our brain distinguish these kinds of self-generated change from the changes that would be produced by the movement of external objects in our environment? von Helmholtz (1867) suggested that the brain makes a copy of its motor plans and uses this copy (the 'efference copy') to predict how its planned movements are likely to affect future sensory data. When the cognitive system issues a motor command (e.g. to rotate its eyeballs), it sends a copy of the command to a generative model (the 'forward model' or 'motor emulator'), which predicts

[29] Sprevak (forthcoming[c]), Section XX.

[30] Aitchison and Lengyel (2017) consider what might happen if, at the algorithmic level, predictive coding's variational methods were replaced by a sampling method (pp. 223–224).

the sensory consequences that will flow from that motor command (how sensory data are likely to change if the eyeballs rotate). These consequences are then fed back to the sensory system and the brain uses them to 'subtract away' estimated self-generated changes from the incoming data. A generative model could thus help the brain to distinguish changes to the sensory data wrought by itself from those that are caused by external objects.[31]

Second, a generative model could help the brain overcome some of the latency, noise, variability, and gaps in sensory input that potentially cause problems for motor control. When you execute a complex, rapid motion – e.g. a tennis serve – your brain needs to have accurate, low-latency sensory feedback. During the motion, your brain needs to where your limbs are, how its intended motor plan is unfolding, if any resistance is being met, and how the positions of external objects (like the ball) are changing. Complex, rapid motor control needs to be *regulated* by sensory feedback. The problem the brain faces is that, due to limitations in its hardware, this sensory feedback tends to arrive late, with many gaps, and a great deal of noise and variability. A generative model would allow the brain to partly overcome these limitations by introducing regulation based, not on *actual* sensory feedback, but on *expected* feedback. This would mean that the brain would not need to wait for (slow, noisy, gappy) sensory data to arrive. It could control motion based on expected sensory data, updating its generative model as and when the actual sensory data do arrive. Potentially, that updating might take into account all sorts of background information that the brain has about systematic bias, noise, or uncertainty in the sensory data or sensory organs. Advocates of this approach suggest that a probabilistic generative model, updated using Bayesian rules, could allow the brain to make *optimal* use of its background knowledge and sensory data to regulate motor control and motor learning.[32]

Third, a generative model that takes a probabilistic form could, in principle, be inverted using Bayes' theorem to yield a *discriminative* model of a domain. Discriminative models are of obvious value in many areas of cognition. A discriminative model tells the cognitive system, given some sensory signal, which state(s) of the

[31]For a description, see Keller and Mrsci-Flogel (2018), pp. 424–425. Blakemore, Frith and Wolpert (1999) use a model of this kind to explain why it is difficult to tickle yourself.

[32]See Franklin and Wolpert (2011); Grush (2004); Körding and Wolpert (2004); Körding and Wolpert (2006).

world are most likely to be responsible for its observations.[33] Discriminative models are needed in visual perception, object categorisation, speech recognition, detection of causal relations, and social cognition. Whereas a discriminative model tells the cognitive system how make the inferential leap *from* sensory data *to* the value of latent unobserved variables, a generative model tells the cognitive system how to make an inference *from* the value of latent variables *to* sensory observations. That inverse information might not initially appear to be useful, but if the system applies Bayes' theorem, a generative model can be flipped to create a discriminative model. What is more, building a generative model of a domain might be a computationally attractive strategy because generative models are often easier to learn, easier to update, more compact to represent, and less liable to break as background conditions change than discriminative models.[34] Therefore, an effective method for answering a discriminative query (what is the value of a latent variable, given my sensory input?) is sometimes to learn and maintain a generative model of the domain in question and then invert it as and when needed using Bayes' theorem to answer the query. It is common to see this generative strategy used in contemporary machine learning and computational cognitive science.[35]

Nowadays, it is not usual for a computational model of cognition to include a generative model. However, the considerations above do not specifically support predictive coding's proposal about the nature, content, and function of a generative model. They do not, for example, commit to the idea that the brain only has a one generative model, or that computation over that generative model is its exclusive method of inference, or to predictive coding's ideas about the particular content or structure of the brain's probabilistic generative model, the algorithms by which it is updated or used in inference, or where in the brain it is physically implemented. As far as the points above are concerned, there may be multiple generative models in cognition. Distinct models might exist in relative informational isolation inside different cognitive modules – there might, for example, be a domain-specific gen-

[33] A *discriminative model* is typically defined as a model that tells one the conditional probability of some unobserved target variable $Y$, given an observation $x$, $P(Y \mid X = x)$. A *generative model* is defined either as a model that tells one the likelihood function, i.e. the conditional probability an observation, $X$, given some hidden state of the world, $y$, $P(X \mid Y = y)$; or, as a model that tells one the joint probability distribution, $P(X, Y)$. In practice, the difference between the two does not matter as the joint probability distribution is equal to the product of the likelihood function and the system's priors over those unobserved states: $P(X, Y) = P(X \mid Y)P(Y)$, and both likelihood and priors are needed to invert the model under Bayes' theorem.

[34] The reasons for this complex and depend on the contingent way our world is often structured. For a brief intuitive explanation, see Russell and Norvig (2010), pp. 497, 516–517.

[35] See Bishop (2006), Ch. 4 on creating classifiers using generative models. See Chater and Manning (2006); Kriegeskorte (2015); Poeppel and Bever (2010); Tenenbaum et al. (2011); Yuille and Kersten (2006) for various proposals for using generative models to answer discriminative queries in cognition.

erative model dedicated exclusively to motor control.[36] The considerations above are also consistent with brain using other methods alongside generative models to solve problems. When faced with a discriminative problem, for example, the brain might sometimes learn and use a discriminative model of that domain directly, or adopt any number of hybrid generative-discriminative approaches.[37] Finally, there are endless ways in which the content and structure of a generative model might be filled out, methods by which a generative model might be updated and used, and proposals for how it might be physically implemented in the brain.[38]

Generative models feature in many contemporary computational models of cognition. Predictive coding employs the idea, but the idea is not unique to predictive coding. The proposal that the brain uses a generative model should not be equated with predictive coding. One should not assume that empirical evidence that favours the hypothesis that the brain employs a generative model is also evidence that supports predictive coding's specific proposal about the character and role of a generative model in cognition.

## 7 Conclusion

The aim of this paper is to separate five influential ideas about computational modelling of cognition from predictive coding. Many philosophers first encounter these five ideas in the context of predictive coding. They should be aware that the ideas are not unique to that view: they exist in a broader intellectual landscape and they are employed by approaches that have little or nothing to do with predictive coding. Endorsement of one or more of the ideas should not be interpreted as an implicit endorsement of predictive coding. Empirical evidence that supports one or more of the ideas should not be interpreted as straightforwardly evidence for predictive coding (rather than evidence for any number of other views). If one wants to understand the *distinctive* content of predictive coding, or to evaluate the empirical evidence for it, one needs to disentangle it from these other ideas.

Of course, there is nothing to stop one from adopting the deflationary view that 'predictive coding' refers to some broad, unspecified synthesis of the five ideas. On such a view, one could say, without fear of contradiction, that predictive coding is

[36] Wolpert, Ghahramani and Flanagan (2001); Grush (2004) propose this. They also suggest that the generative model used by motor control is not implemented in the neocortex, but in the cerebellum.

[37] See Ng and Jordan (2002) for conditions under which learning and using a discriminative model of a domain is more efficient than learning a generative model and inverting it. For examples of hybrid discriminative-generative approaches, see Raina et al. (2003); Lasserre, Bishop and Minka (2006).

[38] See Sprevak (forthcoming[b]), Sect XX; Sprevak (forthcoming[c]), Sect XX.

already widely accepted and experimentally confirmed. However, there are good reasons to resist such a move. Advocates of predictive coding are keen to stress that their view is both novel with respect to contemporary rivals and that it has testable empirical content. To the extent that these claims are justified, an advocate of predictive coding should be able to show that predictive coding departs from other views and that it does not make a claim that is so anodyne as to be consistent with any future evidence. Clark, for example, warns against interpreting predictive coding as 'extremely broad vision of the brain as an engine of multilevel probabilistic prediction' (Clark, 2016, p. 10). Predictive coding should be a 'specific proposal', not a 'broad vision' (ibid.). Hohwy observes that there can be an ambiguity in how the view is presented which means that it is 'both mainstream and utterly controversial' (Hohwy, 2013, p. 7). Hohwy says that in order to make meaningful contact with empirical evidence, a specific version of the theory is needed (Hohwy, 2013, pp. 7–8).

What is that specific, constrained version of predictive coding? In what follows, I propose that what distinguishes predictive coding consists in a combination of three, potentially dissociable, claims, each of which may be further developed or qualified in various ways. These claims concern how cognition works at Marr's *computational*, *algorithmic*, and *implementation* levels.

It is worth tempering what follows with the cautionary note that the content of predictive coding is in no way a settled matter. Researchers differ about which features of the view matter, how they should be articulated, whether the resulting model will have a truly universal applicability to every aspect of human cognition, and whether the computational, algorithmic, and implementation level claims should all be asserted together, or packaged into a single framework in the way proposed. Cutting through this disagreement and uncertainty however, is an idea that has inspired many researchers: a simple, bold, unifying picture of the mind, its computational architecture, and its physical implementation. This (perhaps deliberately idealised and simplified) version of the view will be the primary target of the next three papers.

## Bibliography

Aitchison, L. and M. Lengyel (2017). "With or without you: Predictive coding and Bayesian inference in the brain". In: *Current Opinion in Neurobiology* 46, pp. 219–227.

Attneave, F. (1954). "Informational aspects of visual perception". In: *Psychological Review* 61, pp. 183–193.

Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge: Cambridge University Press.

Barlow, H. B. (1961). "Possible principles underlying the transformations of sensory messages". In: *Sensory Communication*. Ed. by W. A. Rosenblith. Cambridge, MA: MIT Press, pp. 217–234.

— (2001). "Redundancy reduction revisited". In: *Network: Computation in Neural Systems* 12, pp. 241–253.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York, NY: Springer.

Blakemore, S. -J., C. D. Frith and D. M. Wolpert (1999). "Spatio-temporal prediction modulates the perception of self-produced stimuli". In: *Journal of Cognitive Neuroscience* 11, pp. 551–559.

Chater, N. and C. D. Manning (2006). "Probabilistic models of language processing and acquisition". In: *Trends in Cognitive Sciences* 10, pp. 335–344.

Chater, N. and M. Oaksford, eds. (2008). *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*. Oxford: Oxford University Press.

Clark, A. (2011). "What scientific concept would improve everybody's cognitive toolkit?" In: *Edge*. URL: https://www.edge.org/response-detail/10404.

— (2013). "Whatever next? Predictive brains, situated agents, and the future of cognitive science". In: *Behavioral and Brain Sciences* 36, pp. 181–253.

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

— (2019). "Consciousness as generative entanglement". In: *The Journal of Philosophy* 116, pp. 645–662.

Colombo, M., L. Elkin and S. Hartmann (2018). "Being realist about Bayes, and the predictive processing theory of mind". In: *The British Journal for the Philosophy of Science*. DOI: 10.1093/bjps/axy059.

de Finetti, B. (1990). *Theory of Probability*. Vol. 1. New York, NY: Wiley & Sons.

De Lange, F. P., M. Heilbron and P. Kok (2018). "How do expectations shape perception?" In: *Trends in Cognitive Sciences* 22, pp. 764–779.

Deneve, S. (2008). "Bayesian spiking neurons I: Inference". In: *Neural Computation* 20, pp. 91–117.

Dennett, D. C. (1991). *Consciousness Explained*. Boston, MA: Little, Brown & Company.

Drayson, Z. (2017). "Modularity and the predictive mind". In: *Philosophy and Predictive Processing*. Ed. by T. Metzinger and W. Wiese. 10.15502/9783958573031: MIND Group. DOI: 10.15502/9783958573024.

Firestone, C. and B. J. Scholl (2016). "Cognition does not affect perception: Evaluating the evidence for "top-down" effects". In: *Behavioral and Brain Sciences* 39, E229.

Fiser, J., P. Berkes, G. Orbán and M. Lengyel (2010). "Statistically optimal perception and learning: From behavior to neural representations". In: *Trends in Cognitive Sciences* 14, pp. 119–130.

Franklin, D. W. and D. M. Wolpert (2011). "Computational mechanisms of sensorimotor control". In: *Neuron* 72, pp. 425–442.

Friston, K., P. Schwartenbeck, T. FitzGerald, M. Moutoussis, T. Behrens and R. J. Dolan (2013). "The anatomy of choice: active inference and agency". In: *Frontiers in Human Neuroscience* 7, p. 598.

Gardner-Medwin, A. R. and H. B. Barlow (2001). "The limits of counting accuracy in distributed neural representations". In: *Neural Computation* 13, pp. 477–504.

Gregory, R. L. (1997). "Knowledge in perception and illusion". In: *Philosophical Transactions of the Royal Society of London, Series B* 352, pp. 1121–1128.

Griffiths, T. L., E. Vul and A. N. Sanborn (2012). "Bridging levels of analysis for probabilistic models of cognition". In: *Current Directions in Psychological Science* 21, pp. 263–268.

Grush, R. (2004). "The emulator theory of representation: Motor control, imagery, and perception". In: *Behavioral and Brain Sciences* 27, pp. 377–442.

Hohwy, J. (2012). "Attention and conscious perception in the hypothesis testing brain". In: *Frontiers in Psychology* 3, pp. 1–14.

— (2013). *The Predictive Mind*. Oxford: Oxford University Press.

Hoyer, P. O and A. Hyvärinen (2003). "Interpreting neural response variability as Monte Carlo sampling of the posterior". In: *Advances in Neural Information Processing Systems 15*. Ed. by S. Becker, S. Thrun and K. Obermayer. Cambridge, MA: MIT Press, pp. 277–284.

Keller, G. B. and T. D. Mrsci-Flogel (2018). "Predictive processing: A canonical cortical computation". In: *Neuron* 100, pp. 424–435.

Kirchhoff, M. D. and J. Kiverstein (2019). *Extended consciousness and predictive processing*. Abingdon: Routledge.

Knill, D. C. and A. Pouget (2004). "The Bayesian brain: the role of uncertainty in neural coding and computation". In: *Trends in Neurosciences* 27, pp. 712–719.

Körding, K. P. and D. M. Wolpert (2004). "Bayesian integration in sensorimotor learning". In: *Nature* 427, pp. 244–247.

— (2006). "Bayesian decision theory in sensorimotor control". In: *Trends in Cognitive Sciences* 10, pp. 319–326.

Kriegeskorte, N. (2015). "Deep neural networks: A new framework for modeling biological vision and brain information processing". In: *Annual Review of Vision Science* 1, pp. 417–446.

Lasserre, J. A., C. M. Bishop and T. P. Minka (2006). "Principled hybrids of generative and discriminative models". In: *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. New York, NY: IEEE, pp. 87–94.

Lupyan, G. (2015). "Cognitive penetrability of perception in the age of prediction: Predictive systems are penetrable systems". In: *Review of Philosophy and Psychology* 6, pp. 547–569.

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

Macpherson, F. (2012). "Cognivite penetration of colour experience: Rethinking the issue in light of an indirect mechanism". In: *Philosophy and Phenomenological Research* 84, pp. 24–62.

— (2017). "The relationship between cognitive penetration and predictive coding". In: *Consciousness and Cognition* 47, pp. 6–16.

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

Matloff, N. (2017). *Statistical Regression and Classification*. Boca Raton, FL: CRC Press.

Moreno-Bote, R., D. C. Knill and A. Pouget (2011). "Bayesian sampling in visual perception". In: *Proceedings of the National Academy of Sciences* 108, pp. 12491–12496.

Neisser, U. (2014). *Cognitive Psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Ng, A. Y. and M. I. Jordan (2002). "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes". In: *Advances in Neural Information Processing Systems 14*. Ed. by T. G. Dietterich, S. Becker and Z. Ghahramani. Cambridge, MA: MIT Press, pp. 841–848.

Niv, Y. and G. Schoenbaum (2008). "Dialogues on prediction errors". In: *Trends in Cognitive Sciences* 12, pp. 265–272.

Poeppel, D. and T. G. Bever (2010). "Analysis by synthesis: A (re-)emerging program of research for language and vision". In: *Biolinguistics* 4, pp. 174–200.

Pouget, A., P. Dayan and R. S. Zemel (2003). "Inference and computation with population codes". In: *Annual Review of Neuroscience* 26, pp. 381–410.

Rahnev, D. (2017). "The case against full probability distributions in perceptual decision making". In: *bioRxiv*. DOI: 10.1101/108944.

Raina, R., Y. Shen, A. McCallum and A. Y. Ng (2003). "Classification with hybrid generative/discriminative models". In: *Advances in Neural Information Processing Systems 16*. Ed. by S. Thrun, L. K. Saul and B. Schölkopf. Cambridge, MA: MIT Press, pp. 545–552.

Ramsey, F. P. (1990). *Philosophical Papers*. Ed. by D. H. Mellor. Cambridge: Cambridge University Press.

Russell, S. and P. Norvig (2010). *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River, NJ: Pearson.

Sanborn, A. N. and N. Chater (2016). "Bayesian brains without probabilities". In: *Trends in Cognitive Sciences* 20, pp. 883–893.

— (2017). "The sampling brain". In: *Trends in Cognitive Sciences* 21, pp. 492–493.

Schultz, W., P. Dayan and P. R. Montague (1997). "A neural substrate of prediction and reward". In: *Science* 275, pp. 1593–1599.

Schwartenbeck, P., T. FitzGerald, C. Mathys, R. J. Dolan and K. Friston (2015). "The dopaminergic midbrain encodes the expected certainty about desired outcomes". In: *Cerebral Cortex* 25, pp. 3434–34445.

Seth, A. K. (2017). "The cybernetic brain: From interoceptive inference to sensorimotor contingencies". In: *Philosophy and Predictive Processing*. Ed. by T. Metzinger and W. Wiese. Frankfurt am Main: MIND Group. DOI: 10.15502/9783958570108.

Simoncelli, E. P. and B. A. Olshausen (2001). "Natural image statistics and neural representation". In: *Annual Review of Neuroscience* 24, pp. 1193–1216.

Sprevak, M. (forthcoming[a]). "Predictive coding II: The computation". In: *TBC*.

— (forthcoming[b]). "Predictive coding III: The algorithm". In: *TBC*.

— (forthcoming[c]). "Predictive coding IV: The implementation". In: *TBC*.

— (forthcoming[d]). "Predictive coding: Appendix". In: *TBC*.

Sterling, P. and S. Laughlin (2015). *Principles of Neural Design*. Cambridge, MA: MIT Press.

Stone, J. V. (2018). *Principles of Neural Information Theory: Computational Neuroscience and Metabolic Efficiency*. Sebtel Press.

Tenenbaum, J. B., C.. Kemp, T. L. Griffiths and N. D. Goodman (2011). "How to grow a mind: Statistics, structure, and abstraction". In: *Science* 331, pp. 1279–1285.

Toderici, G., D. Vincent, N. Johnston, S. J. Hwang, D. Minnen, J. Shor and M. Covell (2016). "Full resolution image compression with recurrent neural networks". URL: http://arxiv.org/abs/1608.05148.

Usevitch, B. E. (2001). "A tutorial on modern lossy wavelet image compression: foundations of JPEG 2000". In: *IEEE Signal Processing Magazine* 18, pp. 22–35.

Von Helmholtz, H. (1867). *Handbuch der physiologischen Optik*. Hamburg und Leipzig: Leopold Voss.

Wolpert, D. M., Z. Ghahramani and J. R. Flanagan (2001). "Perspectives and problems in motor learning". In: *Trends in Cognitive Sciences* 5, pp. 487–494.

Yuille, A. and D. Kersten (2006). "Vision as Bayesian inference: analysis by synthesis?" In: *Trends in Cognitive Sciences* 10, pp. 301–308.