Predictive coding I

Received: 10 March 2022 | Revised: 20 June 2023 | Accepted: 25 September 2023

DOI: 10.1111/phc3.12950

ARTICLE

WILEY

Predictive coding I: Introduction

Mark Sprevak

School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, UK

Correspondence
Mark Sprevak.
Email: mark.sprevak@ed.ac.uk

Abstract

Predictive coding – sometimes also known as 'predictive processing', 'free energy minimisation', or 'prediction error minimisation' – claims to offer a complete, unified theory of cognition that stretches all the way from cellular biology to phenomenology. However, the exact content of the view, and how it might achieve its ambitions, is not clear. This series of articles examines predictive coding and attempts to identify its key commitments and justification. The present article begins by focusing on possible confounds with predictive coding: claims that are often identified with predictive coding, but which are not predictive coding. These include the idea that the brain employs an efficient scheme for encoding its incoming sensory signals; that perceptual experience is shaped by prior beliefs; that cognition involves minimisation of prediction error; that the brain is a probabilistic inference engine; and that the brain learns and employs a generative model of the world. These ideas have garnered widespread support in modern cognitive neuroscience, but it is important not to conflate them with predictive coding.

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 the brain. 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 by appeal to a single, unified 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 cognition, but one that stretches all the way from the details of neuromodulator release to abstract principles of rational action governing whole agents.[1]

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.[2] For another, important features of the view, whatever its name, are liable to change or are underspecified in important respects. In this article and those that follow it, my aim is to sketch what predictive coding is, and how it might fulfil these ambitions.

I argue that predictive coding should be understood as a loose alliance of three 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.[3] At Marr's computational level, the claim is that the computational *task* facing the brain is to minimise sensory prediction error. At the algorithmic level, the claim is that the *algorithm* by which our brain attempts to solve this task involves the action of a hierarchical network of abstract prediction and error units. This network may be viewed, in a further step, as running a variational algorithm for approximate Bayesian inference. At Marr's implementation level, the claim is that the *physical resources* that implement the algorithm are primarily located in the neocortex: anatomically distinct cell populations inside neocortical areas implement distinct prediction and error units.

Each of these claims needs to be qualified in certain respects and 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 this, and many details are currently left open. This means that the exact commitments of predictive coding are, to put it mildly, contentious. For these reasons, it is more accurate to think of predictive coding as an ongoing research programme rather than 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 and precisified – descendent of the three claims above. As with any such 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 future potential, and its ability to inspire and guide fruitful research.[4]

Before saying what predictive coding is, it is first helpful to say what it is not. In this article, I outline five ideas that are often presented alongside predictive coding, but which should be distinguished from it. In the three articles that follow, I focus primarily 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, forthcoming-b, forthcoming-c). As we will see, there are many ways in which its basic ideas may be elaborated and refined. My strategy is to present what, in my opinion, are the 'bare bones' of the approach. For readers new to this topic, I hope that this will provide you with a scaffold on which to drape a more nuanced future understanding of the view.[5]

For the remainder of this article, I focus on five ideas that feature prominently in expositions of predictive coding, but which should be distinguished from predictive coding. These ideas are: (i) that the brain employs an efficient coding scheme; (ii) that perception has top-down, expectation-driven effects; (iii) that cognition involves minimisation of prediction error; (iv) that cognition is a form of probabilistic inference; (v) that cognition makes use of generative models. All these ideas are used by predictive coding but, I argue, they are also shared by a variety of other computational approaches. They do not reflect – taken either singly or jointly – what is distinctive about predictive coding's research programme. If one wishes to know what is special about predictive coding, these ideas, whatever their intrinsic value, can function as potential distractors. A corollary of this is that evidence for predictive coding does not necessarily flow from evidence that supports these more general ideas. Evidence for predictive coding should aim to selectively support predictive coding with respect to plausible contemporary rivals, not merely to confirm ideas that are shared by a wide variety of other approaches.

The literature on predictive coding is vast. In what follows, I ignore many interesting developments, proposals, and applications. My description is also inevitably partisan: there is too much disagreement within the primary literature to be able to characterise the view in a wholly uncontroversial way. If you disagree with my description, I hope that what I say at least provides a foil by which to triangulate your own views.

In both the present article and those that follow, I only consider predictive coding as a theory of subpersonal cognitive processing. I do not consider how its computational model might be adapted or extended to account for personal-level thought or conscious experience. Explaining conscious experience with predictive coding is a relatively recent development. However, it is a project that assumes we have a prior understanding of what predictive coding's computational model is. That question is the focus of this review.[6]

2 | EFFICIENT NEURAL CODING

A key idea that predictive coding employs is that the brain's coding scheme for storing and transmitting sensory information is, in a certain sense, efficient. The relevant form of efficiency is quantified by the degree to which the brain compresses incoming sensory information (measured in terms of Shannon information theory). To compress information, the sensory system should aim to transmit only what is 'new' or 'unexpected' or 'unpredicted' relative to its expectations. If the sensory system were to encode certain assumptions about its incoming sensory data, these would enable it to predict bits of that incoming sensory stream. This means that fewer bits would need to be stored or transmitted inwards from the sensory boundary, yielding a potential reduction in the costs of the brain physically storing and transmitting that data. The more accurately the brain's internal assumptions reflect its incoming sensory stream, the less information would need to be stored or transmitted inwards from the sensory periphery. All that would need to be sent inwards would be an error signal – what is new or unexpected – with respect to those predictions. A similar idea underlies coding schemes that allow electronic computers to store and transmit images and videos across the Internet (e.g. JPEG or MPEG).

The notion that our brains use a sensory 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 encoding sensory information based partly on the grounds that neurons in the early visual system have a limited physical dynamic range: the action potentials they send inwards to cortical centres are precious and should not be squandered to send information that those cortical centres already have.[7] Predictive coding adopts the same basic perspective, but elevates it to a universal design principle: not only the early stages of vision, but every aspect of cognition, should be viewed as an attempt by the brain to compress its incoming sensory data. To this, predictive coding adds a range of further assumptions about (i) the algorithm by which the incoming sensory data are compressed; (ii) how assumptions used for sensory compression are changed during learning; (iii) where physically in the brain all this takes place.

Predictive coding has particular views about how compression of sensory signals works – see (i)–(iii) above. It also adopts the rather extreme position that sensory compression is the brain's *only* goal. As Barlow made clear in his later work, even if one thinks that compressing incoming sensory data is one thing that the brain does, it is not obvious that it 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 would count as 'efficient' coding for image data on a PC. If all one wishes to do is to transmit an image across the Internet, then compressing it using a redundancy reducing code

(e.g. JPEG) might be a good solution, since it would reduce the number of physical signals one would need to send. Similarly, if one only wishes to store the image on a hard disk drive, then compressing it would mean that fewer physical resources would be required for its storage.[8] However, if one wishes to *transform* the image or *perform an inference* over it, then a redundancy reducing code like JPEG may not be the best or most efficient solution. Compressed data are often harder to work with. If you ask a PC to rotate an image 23° clockwise, the machine will generally not attempt to execute this operation on a compressed encoding of the image data. Instead, it will switch to an uncompressed version of the image (e.g. 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 defined over their more compressed counterparts.[9] 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.[10]

If all that matters to the brain in cognition are the costs of transmitting and storing incoming sensory data, then it may make sense for the brain to aim to maximally compress that incoming sensory data. However, if speed, simplicity, and ease of inference matter, then it may make sense to add or preserve redundant structure within incoming sensory data.[11] Reducing redundancy is not the only possible objective for a cognitive system that aims at efficient sensory coding.

It is common for contemporary work on efficient coding to acknowledge this point.[12] Predictive coding, in its strongest and purest form, adopts a rather extreme view: it equates efficiency with sensory redundancy reduction, and it claims that the entire brain (not just certain areas in the sensory cortex) is devoted to this task; it also claims that the sensory compression is accomplished by a specific algorithm and representational scheme. Although predictive coding employs the idea of efficient coding, the general idea is not unique to predictive coding. Similarly, although evidence for efficient sensory coding in, e.g. early stages in the visual cortex, may be compatible with predictive coding, it may also be compatible with a range of other, more modest proposals about efficient coding in cognition.

3 | TOP-DOWN, EXPECTATION-DRIVEN EFFECTS IN PERCEPTION

Top-down, expectation-driven effects in perception are instances in which an agent's prior beliefs systematically affect that agent's perceptual experience. Top-down, expectation-driven effects are sometimes presented as a hallmark feature of predictive coding. Predictive coding's computational model is thought to imply that perception is top-down or expectation-laden: 'What we perceive (or think we perceive) is heavily determined by what we know' (Clark, 2011). Evidence for top-down effects in perception is also thought to somehow confirm predictive coding's computational model: we should give higher credence to predictive coding's computational proposal based on observation of top-down effects in perception.[13]

However, the relationship between predictive coding and top-down, expectation-driven effects in perception is more complex and less direct than this.

First, top-down effects in perception are standardly defined in terms of a relationship between an agent's *personal-level* states: what an agent *believes* affects their *perceptual experience*.[14] Predictive coding, at least in the first instance, makes a claim about the agent's subpersonal computational states and processes. The 'top' and 'bottom' in predictive coding's computational model refer, as we will see, to subpersonal computational states of the agent. 'High-level' neural representations (implemented deep in the cortical hierarchy) are assumed to have a 'top-down' influence on 'low-level' representations (implemented in the early sensory system). How this kind of subpersonal 'top-down effect' relates to personal-level top-down effects observed in psychology is presently unclear.

One might argue that, at a minimum, personal-level top-down effects require *some* subpersonal information to flow from high-level cognitive centres to low-level sensory systems. However, it is difficult to know what can be inferred from this assumption regarding personal-level experience. Not every piece of subpersonal information posited by predictive coding's computational model features in the contents of either personal-level belief or perceptual experience. Only a tiny fraction of subpersonal information appears to be present at the personal level. For

predictive coding to say something specific about the existence or character of top-down effects at the personal level, it would need to say *which* aspects of that subpersonal information give rise to *which* personal-level states (beliefs and perceptual contents). These assumptions – which connect the subpersonal level to the personal level – are currently not to be found anywhere within predictive coding's computational model. Ideas about these connections have been proposed, but exactly how subpersonal states of the computational model map onto personal-level beliefs and perceptual experiences remains a highly speculative matter.[15] 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 psychology.[16]

Second, positing top-down subpersonal information flow inside a computational model is not a characteristic that is unique to predictive coding. Almost any plausible computational model of cognition is likely to claim that information flows both 'upwards' (from lower-level sensory systems to high-level cognitive centres) and 'downwards' (from high-level cognitive centres to lower-level sensory systems). As Ira Hyman observed 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'.[17] This could even be said of so-called 'bottom-up' computational models, such as the account of vision proposed by Marr (1982). Those models might appear to ignore top-down processes, but this is not because they hold that top-down influences do not exist in the brain or are unimportant, but rather because they are not necessary to explain a particular phenomenon of interest.[18] Indeed, it has been for a long time standard practice in cognitive science to invoke top-down information flow to account for endogenous attention, semantic priming, and to explain how the brain handles ambiguity, noise, and uncertainty in its sensory input.[19] The mammalian brain contains a huge number of 'backward' cortical connections which suggest that signals carried from cortical centres to peripheral sensory areas have a significant computational role in cognitive processing. Even if one were to ignore these connections, Firestone and Scholl (2016) observe that there are many other causal routes by which high-level cognitive centres should be expected to systematically affect processing in low-level sensory systems – the decision to 'shut one's eyes' causes one's eyelids to close, which changes low-level sensory inputs, systematically affecting the contents of states in subpersonal low-level sensory systems, for example.[20] When advocates of predictive coding suggest that their model has a special relationship with top-down, expectation-driven effects observed at the personal level, a challenge they face is to explain why predictive coding's specific set of top-down computational pathways is *uniquely* or *best* suited to explain these effects.

To be clear, predictive coding's computational model is *compatible* with personal-level top-down effects in perception occurring; it is also broadly *suggestive* that such effects would occur. What is not clear is that it is *better* suited to account for these effects than any number of other models that also incorporate subpersonal top-down information flow (e.g. other kinds of recurrent neural networks or classical computational models with loops). For these reasons, it is not clear how personal-level top-down effects is distinctively associated with, or selectively confirms, predictive coding.

4 | MINIMISING PREDICTION ERROR

It is common in contemporary artificial intelligence (AI) to characterise learning and inference in terms of minimising prediction error. During learning, an AI system might attempt to change its parameters to better predict its training data. During inference, an AI system might search for values of its variables that would result in it generating predictions that minimise prediction errors – that are as close to 'ground truth' as possible.[21] Different AI systems might differ in the types of data they try to predict, the mathematical model they use for prediction, or the way they revise parameters of that model during learning.[22] Prediction error might also be measured in a number of ways. A common formalisation is mean-squared error – the average of the squares of the differences between the predicted values and the true values of the data.[23]

The logical space of possible computational systems that aim to minimise their prediction error is vast. One can get some idea of the size and diversity of that space by opening up any current textbook on machine learning

or statistics.[24] A maximally simple example of a system that aims to minimise its prediction error would be one that performs linear regression on its training data. Here, minimising prediction error reduces to just fitting a straight-line mathematical model to the training data and using that straight-line model to make predictions about unseen cases. Learning consists in finding the value of two parameters (slope and *y*-intercept) that would define a straight line that minimises mean-squared error over the training data. Classical statistics contains many algorithms for finding those values (e.g., the ordinary least squares algorithm). Deep neural networks provide more complicated examples of computational systems that aim to minimise their prediction error. Here, learning consists in finding the values of not just two, but millions or billions of parameters. Algorithms like backpropagation are commonly used to find these values. During inference, a deep neural network might execute a long sequence of mathematical operations over many variables in an effort to yield an output that is as close to the ground truth as possible.

Predictive coding suggests that the brain, like many other computational systems, aims to minimise a measure of prediction error. What distinguishes predictive coding from other proposals is that it makes specific claims about the *data*, *model*, and *algorithm* used in this task; a distinctive claim is also made about the *role* of this instance of prediction error minimisation within the brain's wider cognitive economy.

Regarding the *data*, predictive coding claims that the brain aims to minimise prediction error concerning incoming *sensory signals*. This should be distinguished from other approaches that claim that the brain aims to minimise prediction error concerning other forms of data, such as *reward signals*.[25] The mathematical *model* the brain uses to generate its predictions is encoded in an abstract hierarchical network containing prediction and error units linked by weighted connections. This network is similar to the connectionist networks found in deep learning, although the behaviour of individual units and the overall topology of the network differs from those commonly used in deep learning. The *algorithm* that adjusts the parameters of the network during learning is also different. Deep learning tends to use some version of backpropagation; predictive coding suggests that the brain uses a Hebbian learning algorithm.[26] Finally, a special *role* is accorded to prediction error minimisation in cognition. Predictive coding holds that minimising prediction error over sensory signals is not just one among many objectives undertaken by the brain, but its only or fundamental objective.

It is common to find prediction error minimisation occuring inside a computational model of cognition. What marks out predictive coding as special is the claim that cognition exclusively involves prediction error minimisation over a specific set of data, with a specific mathematical model, and using a specific algorithm for learning and inference. Evidence for prediction error minimisation occuring in the brain, although it may be compatible with predictive coding, may also be compatible with any number of other computational models that also employ prediction error minimisation.

5 | COGNITION AS A FORM OF PROBABILISTIC INFERENCE

Brains receive noisy, incomplete, and sometimes contradictory information via their sensory organs. They need to weigh this information rapidly and integrate it with (sometimes conflicting) background knowledge in order to reach a decision and generate behaviour. Probabilistic models of cognition provide a broad framework by which to understand how brains do this. According to these models, brains do not represent the world in purely categorical way (e.g. 'the person facing me is my father'), but instead represent multiple possibilities (e.g. 'the person facing me is my father, my uncle, his cousin, …') along with some measure of uncertainty regarding those outcomes.[27] Computational models typically formalise this by ascribing mathematical *subjective probability distributions* to brains. These probability distributions measure the brain's degree of confidence in a range of different possibilities.[28] Cognitive processing is then modelled as a series of operations in which one subjective probability distribution conditions, or updates, another. The exact manner in which this happens may vary between different computational models. In principle, cognitive processing may maintain this probabilistic character until the brain 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).

A particularly influential example of this approach is the *Bayesian brain hypothesis*.[29] On this view, Bayes' rule, or some approximation to it, is assumed to describe how the brain combines and updates its subjective probability distributions.[30] Because exact Bayesian inference is computationally intractable, advocates of the Bayesian brain hypothesis generally assume that the brain implements some version of approximate Bayesian inference. Approximate Bayesian inference can be achieved in a variety of ways, the most popular of which being *sampling algorithms* (which use multiple categorical samples to create an empirical distribution that approximates the true Bayesian posterior) and *variational algorithms* (which change the parameters of some simpler, more computationally tractable distribution in order to try to find a posterior distribution that is close to the true Bayesian posterior).[31] Both forms of approximate Bayesian inference are common in AI and machine learning. Proponents of the Bayesian brain hypothesis do not agree about whether the brain uses a sampling method, a variational method, or something else entirely.[32]

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 as that of minimising sensory prediction error. If combined with appropriate simplifying assumptions, this task can be shown to entail approximate Bayesian inference.[33] 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.[34] Predictive coding proposes that these numerical parameters, and 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 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.[35]

Someone might endorse the idea that the brain engages in probabilistic inference, or even the Bayesian brain hypothesis, but reject some or all of these further assumptions. For example, someone might not accept that a single probabilistic model underlies every aspect of cognition, or that the subjective probability distributions in the brain are always Gaussian, or that the brain uses the specific version of variational Bayesian inference proposed by predictive coding, or that the brain's subjective probability distributions are encoded in the neocortex.[36] Predictive coding is an example of a probabilistic model of cognition, but there are many possible alternative probabilistic models. Endorsement of, or evidence for, a probabilistic approach to cognition cannot straightforwardly be read as endorsement of, or evidence for, 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 predict which sights, sounds, smells you would experience. At the highest level of abstraction, you might conceive of a generative model as a black box that takes, as input, a hidden state of the world and that yields, as output, the sensory signals that would be likely to be observed. It is widely thought that generative models – and in particular, probabilistic generative models – play an important role in cognition. This is for at least three reasons.

First, a generative model could help the brain to distinguish between changes to its sensory data that are *self-generated* and *externally generated*. When our eyes move, our sensory input changes. How does the brain know which changes are due to movement of our sensory organs and which are due to movement of external objects in the environment? Helmholtz (1867) proposed that our brain makes a copy of its upcoming motor plans and uses this copy (the 'efference copy') to predict how its plans are likely to affect incoming sensory data. A generative model (the 'forward motor model') predicts the likely sensory consequences of a planned movement (e.g. how sensory data would be likely to change if the eyeballs rotate). These predictions are then fed back to the sensory system and 'subtracted away' from incoming sensory data. This would allow the brain to compensate for changes its own movement introduces into its sensory data stream.[37]

Second, a generative model would help the brain to overcome some of the inherent latency, noise, and gaps in its sensory data. When you execute a complex, rapid motion – e.g. a tennis serve – your brain needs to have accurate, low-latency sensory feedback. It needs to know where your limbs are, how its motor plan is unfolding, whether any unexpected resistance is being met, and how external objects (like the tennis ball) are moving. Due to the limits of the brain's physical hardware, this sensory feedback is likely to arrive late, with gaps, and with noise. A generative model would help the brain to alleviate these problems by regulating its motor control based, not on actual sensory feedback, but on expected sensory feedback. When the incoming sensory data do arrive, the brain could then integrate them into its predictions in a way that takes into account any background information that it has about bias, noise, and uncertainty in that sensory signal. Franklin and Wolpert (2011) argue that this would allow the brain to make 'optimal' use of its sensory input during motor control – optimal in the sense that the brain would make use of all its available information.[38]

Third, if a generative model takes a probabilistic form, it could, in principle be, inverted to produce a *discriminative* model.[39] Discriminative models are of obvious utility in many areas of cognition. A discriminative model tells the cognitive system, given some sensory signal, which state(s) of the world are most likely to be responsible for its observations.[40] Discriminative models are needed in visual perception, object categorisation, speech recognition, detection of causal relations, and social cognition. A discriminative model and a generative model facilitate inference in opposite directions: whereas a discriminative model tells the cognitive system how to make the inferential leap from sensory data to the value of latent unobserved variables, a generative model tells the cognitive system how to make the inferential leap from the value of latent variables to sensory observations. The latter form of inference might not initially appear to be useful, but if the system applies Bayes' theorem, a generative model can be used to infer a discriminative model. Moreover, this may be a computationally attractive strategy because generative models are often easier to learn, more compact to represent, and less liable to break when background conditions change.[41] In AI, a common strategy for tackling a discriminative problem is to first learn a generative model of the domain and then invert it using Bayes' theorem. This strategy is frequently suggested as the way in which the brain tackles discriminative problems in certain domains of cognition.[42]

A generative model is a common feature in a modern computational model of cognition. Its content and structure, the methods by which it is updated, and how it might be physically implemented in the brain, might be filled out in many ways, including ways that depart substantially from those suggested by predictive coding. In the context of predictive coding, a single probabilistic generative model is claimed to be employed across all domains of cognition. This generative model is claimed to have a specific hierarchical structure, content, and to be implemented in a specific way in the brain.

Someone might accept that generative models play a role in cognition, but reject these further assumptions. For example, they might hold that multiple distinct generative models exist in the brain in relative functional isolation from each other – e.g., there might be a domain-specific generative model dedicated to motor control.[43] They might hold that the brain does not use a generative model to solve every inference problem – the brain might sometimes attempt to learn and use a discriminative model of a domain directly, or employ some other, non-model-based strategy to reach a decision.[44] They might disagree about the content of the generative model or how the generative model is physically implemented in the brain.[45]

Generative models appear in many computational accounts of cognition. Predictive coding employs the idea, but that idea is not unique to predictive coding. The proposal that the brain uses a generative model should not simply be equated with predictive coding and 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 cognition from predictive coding. Many philosophers first encounter these ideas in the context of predictive coding. However, it is important to recognise that those

ideas exist in a broader intellectual landscape and they are employed by approaches that have little or nothing to do with predictive coding. Accepting one or more of these ideas does not constitute an endorsement of predictive coding. Similarly, evidence that supports one or more of the ideas should not be taken as evidence that unambiguously supports predictive coding. If one wants to understand the distinictive 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 someone defining the words 'predictive coding' to refer to some broad, non-specific synthesis of these five ideas. On such a deflationary reading, one could say, without fear of contradiction, that predictive coding is already widely accepted and empirically 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 and that it faces genuine jeopardy with respect to future evidence. If these claims are to be taken seriously, one would need to show (i) that the view departs from plausible rivals; and (ii) that it is not so anodyne as to be consistent with any likely empirical evidence. To this end, Clark warns against interpreting predictive coding as an 'extremely broad vision'; it should be interpreted as a 'specific proposal' (Clark, 2016, p. 10). Hohwy observes that there is often an ambiguity which renders presentations of predictive coding 'both mainstream and utterly controversial' (Hohwy, 2013, p. 7). He argues that in order for it to meaningfully make contact with empirical evidence, it should be understood as a specific, detailed proposal (Hohwy, 2013, pp. 7–8).[46]

What is that specific, detailed version of predictive coding? In what follows, I argue that what distinguishes predictive coding from contemporary rivals is a combination of three claims, each of which may be precisified 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 a cautionary note. As already mentioned, the specific, detailed content of predictive coding is in no way a settled matter. Researchers disagree about which features of the view are essential, whether the model should be applied to all domains of cognition, whether the computational, algorithmic, and implementation level claims should be combined, and the exact form each of these claims should take. Cutting across this disagreement and uncertainty, however, is a set of ideas that has inspired many researchers: a simple, bold, and unifying picture of the mind, its abstract computational structure, and its physical implementation. This somewhat idealised version of predictive coding will be the focus of the next three papers.

ACKNOWLEDGMENTS

I would like to thank Jonathan Birch, Matteo Colombo, Matt Crosby, Krzysztof Dolega, Jonny Lee, Edouard Machery, Christian Michel, Nina Poth, Wolfgang Schwarz, Dan Williams, Wanja Wiese, and Sam Wilkinson for helpful comments and discussion on earlier drafts of this paper.

ORCID

Mark Sprevak https://orcid.org/0000-0002-1413-5534

ENDNOTES

[1] For examples of these broad claims, see Clark (2013); Clark (2016); Hohwy (2013); Friston (2009, 2010).

[2] Some authors use 'predictive coding' to refer to only one aspect of the view: for example, to the efficient coding strategy described in section 2, or to the algorithm described in section 2 of Sprevak, (forthcoming-b). Some authors call the overall research programme 'predictive processing', 'prediction error minimisation', or 'free energy minimisation'. In what follows, I use the term 'predictive coding' to refer to the overall research programme.

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

[4] The term 'research programme' is used here to indicate that the precise details, goals, and conditions of correct application of a scientific model are often not to be decided in advance and are liable to change over time. It is not meant to indicate commitment to a specific philosophical understanding of a scientific research programme (e.g. that of Lakatos, 1978 or Laudan, 1977). In what follows, I use the terms 'framework', 'approach', 'view', 'account', 'theory', and 'model' interchangeably with 'research programme', with alternative uses flagged along the way.

[5] To help build that understanding, helpful reviews include Aitchison and Lengyel (2017); Friston (2003, 2005, 2009, 2010); Kanai et al. (2015); Keller and Mrsci-Flogel (2018). For reviews that focus on the describing the mathematical and computational framework, see Bogacz (2017); Gershman (2019); Jiang and Rao (2022); Spratling (2017); Sprevak and Smith (2023). For reviews that focus on the possible neural implementation, see Bastos et al. (2012); Jiang and Rao (2022); Lange et al. (2018); Kok & Lange (2015). For reviews that focus on philosophical issues and possible applications to existing problems in philosophy, see Clark (2013, 2016); Friston et al. (2018); Hohwy (2013, 2020); Metzinger and Wiese (2017); Roskies and Wood (2017).

[6] For examples of work that applies predictive coding's computational model to explain conscious experience, see Clark (2019, 2023); Dolega and Dewhurst (2021); Hohwy (2012); Kirchhoff and Kiverstein (2019); Seth (2017, 2021).

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

[8] Other coding schemes such as wavelet-based codes (Usevitch, 2001) or deep neural networks (Bühlmann, 2022; Toderici et al., 2017) would outperform JPEG in these respects. However, these schemes tend to impose even higher computing burdens than JPEG if one wishes to decode or transform an image.

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

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

[11] Gardner-Medwin and Barlow (2001) list examples in which adding redundancy to sensory signals produces faster and more reliable inference over sensory data.

[12] For example, Simoncelli and Olshausen (2001) suggest that the nature of the downstream task a cognitive system faces in a specific context should be considered when measuring the overall efficiency of a coding scheme, not merely the degree of compression of the incoming sensory signal (p. 1210).

[13] For examples of this kind of reasoning, see Clark (2013, p. 190); Lupyan (2015).

[14] See characterisations in Macpherson (2012); Firestone and Scholl (2016). One could also define a 'top-down effect' in terms of how various high-level states in predictive coding's subpersonal computational model change the subject's physically (non-intentionally) characterised behaviour (e.g. physical button presses by a subject during a psychophysics experiment). Such a claim *would* plausibly fall within the scope of predictive coding's model, but its relationship to top-down effects as standardly defined is not obvious. Thanks to Matteo Colombo for this point.

[15] For critical discussion of this point with respect to Seth (2021)'s proposals about personal-level experience, see Sprevak (2022).

[16] See Macpherson (2017); Drayson (2017) for further development of this line of argument. They suggest that predictive coding's computational model is compatible with *no* top-down effects occurring at the personal level at all.

[17] Neisser (2014, p. xvi).

[18] For example, Marr (1982): '… 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.' (pp. 100–101).

[19] See Gregory (1997); Poeppel and Bever (2010); Yuille and Kersten (2006). Firestone and Scholl (2016) suggest that endogenous attention requires subpersonal top-down information flow inside a computational model (p. 14).

[20] Dennett (1991) argues that these kinds of external 'virtual wires', which loop into the environment, can enable sophisticated forms of top-down information processing, including those characteristic of rational thought (pp. 193–199).

[21] For example, see Bishop (2006, pp. 1–12) and Hohwy (2013, pp. 42–46).

[22] Note that a 'prediction' need not be about the future. A prediction is an estimate concerning something that the system does not already know. In principle, a prediction might concern what happened in the past, what is happening in the present, or what will happen in the future. For a helpful review of the relevant notion of prediction, see Lange et al. (2018, p. 766, Box 2) and Forster (2008).

[23] Strictly speaking, AI systems normally aim to minimise a *cost function*, which combines prediction error with other factors. A commonly used cost function is the prediction error plus the sum of the squares of the model's parameters. The latter serves as regularisation term that penalises more complex models. For discussion, see Russell and Norvig (2010, pp. 709–713).

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

[25] There are a wide range of computational models of learning and decision-making that attribute the goal of minimising prediction error over reward signals to the brain (Niv & Schoenbaum, 2008; Schultz et al., 1997). Although these models bear a family resemblance to predictive coding, advocates of predictive coding are generally clear that the two approaches are distinct (Friston, 2009). However, see Friston et al. (2013); Schwartenbeck et al. (2015) for an attempt to show that minimising reward prediction error can be reconceptualised as minimising a measure of expected free-energy that is also associated with sensory prediction error.

[26] See Sprevak, (forthcoming-b), section 2.3.

[27] For examples, see Chater et al. (2006); Danks (2019).

[28] The subjective probabilities in question are formally handled in a similar manner to subjective probabilities inside classical formulations of Bayesianism – i.e. as degrees of belief or credences of some reasoning agent (de Finetti, 1990; Ramsey, 1990). However, unlike in traditional treatments, these subjective probabilities need not be ascribed to the entire agent; they may be ascribed to subpersonal parts of the agent (e.g. to individual brain regions, neural populations, or single neurons) (for example, see Deneve, 2008; Pouget et al., 2013). For discussion of how the concept of subjective probability should be applied to subpersonal parts of agents, see Icard (2016); Rescorla (2020).

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

[30] Bayesian updating is not the only option for handling inference under uncertainty. Plenty of rules and heuristics do not fit the Bayesian norms but still generate adaptive behaviour (Bowers & Davis, 2012; Colombo et al., 2021; Eberhardt & Danks, 2011; Rahnev & Denison, 2018). Rahnev (2017) considers the possibility that brains do not store full probability distributions, but only a few categorical samples or summary statistics (e.g. variance, skewness, kurtosis) and use these partial measures to generate adaptive behaviour.

[31] 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).

[32] For exploration of the idea that the brain uses a sampling method, see Fiser et al. (2010); Griffiths et al. (2012); Hoyer and Hyvärinen (2003); Moreno-Bote et al. (2011); Sanborn and Chater (2016, 2017). Predictive coding is an example of a view that holds that the brain uses a variational method for approximate Bayesian inference.

[33] Sprevak, (forthcoming-a), section 8; Sprevak and Smith (2023).

[34] Sprevak, (forthcoming-b), section 5.

[35] Sprevak, (forthcoming-c), section 3.

[36] Aitchison and Lengyel (2017) consider how predictive coding's proposals might be changed if its algorithm for variational Bayesian inference were replaced with a sampling algorithm (pp. 223–224).

[37] Keller and Mrsci-Flogel (2018, pp. 424–425). Blakemore et al. (1999) use a model of this kind to explain why it is difficult to tickle yourself.

[38] Grush (2004); Körding and Wolpert (2004, 2006); Rescorla (2018).

[39] Bayes' theorem is $P(Y|X) = P(Y|X)P(Y)/P(X)$, and follows from standard axioms and definitions of probability theory. Bayes' rule (referenced in Section 5) says that an agent's subjective probabilities should be updated using Bayesian conditionalisation, $P_{t+1}(Y) = P_t(Y|X)$; its justification does not follow from the axioms of probability (Strevens, 2017).

[40] A discriminative model estimates the probability of a latent variable, Y, given an observation, x, i.e. $P(Y|X = x)$. A generative model is defined either as the likelihood function, i.e. the probability of an observation, X, given some hidden state of the world, y, $P(X|Y = y)$; or, as the full joint probability distribution, $P(X, Y)$. The difference between these rarely matters in practice as the joint probability distribution equals the product of the likelihood and the system's priors over those unobserved states, $P(X, Y) = P(X|Y)P(Y)$, and both likelihood and priors need to be known to invert the model under Bayes' theorem.

[41] The reasons why generative models provide these advantages are complex and depend partly on the contingent way our world is structured. For a brief intuitive explanation, see Russell and Norvig (2010, pp. 497, 516–517).

[42] See Bishop (2006, ch. 4) on creating discriminative 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 about how the brain uses generative models to answer discriminative queries in cognition.

[43] Wolpert et al. (2001); Grush (2004) suggest this. They also suggest that this motor model is not implemented in the neocortex but in the cerebellum.

[44] Ng and Jordan (2002) consider conditions under which it is more efficient to learn a discriminative model of a domain directly than learn a generative model first and then invert it. Raina et al. (2003); Lasserre et al. (2006) examine a range of hybrid discriminative-generative approaches to inference.

[45] See Sprevak, (forthcoming-b), section 2.5; Sprevak, (forthcoming-c), section 6.

[46] Colombo (2017) argues that Clark sometimes interprets predictive coding as a broad vision.

REFERENCES

Aitchison, L., & Lengyel, M. (2017). With or without you: Predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology*, *46*, 219–227. https://doi.org/10.1016/j.conb.2017.08.010

Attneave, F. (1954). Informational aspects of visual perception. *Psychological Review*, *61*(3), 183–193. https://doi.org/10.1037/h0054663

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

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

Barlow, H. B. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems*, *12*(3), 241–253. https://doi.org/10.1080/net.12.3.241.253

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711. https://doi.org/10.1016/j.neuron.2012.10.038

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blakemore, S.-J., Frith, C. D., & Wolpert, D. M. (1999). Spatio-temporal prediction modulates the perception of self-produced stimuli. *Journal of Cognitive Neuroscience*, *11*(5), 551–559. https://doi.org/10.1162/089892999563607

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, *76*, 198–211. https://doi.org/10.1016/j.jmp.2015.11.003

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *128*(3), 389–414. https://doi.org/10.1037/a0026450

Bühlmann, M. (2022). Stable diffusion based image compression. https://pub.towardsai.net/stable-diffusion-based-image-compresssion-6f1f0a399202

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*(7), 335–344. https://doi.org/10.1016/j.tics.2006.05.006

Chater, N. & Oaksford, M. (Eds.). (2008). *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford University Press.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287–291. https://doi.org/10.1016/j.tics.2006.05.007

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

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–253. https://doi.org/10.1017/s0140525x12000477

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Clark, A. (2019). Consciousness as generative entanglement. *The Journal of Philosophy*, *116*(12), 645–662. https://doi.org/10.5840/jphil20191161241

Clark, A. (2023). *The experience machine: How our minds predict and shape reality*. Allen Lane.

Colombo, M. (2017). Review of Andy Clark, Surfing Uncertainty: Prediction, Action, and the Embodied Mind. *Minds and Machines*, *27*(2), 381–385. https://doi.org/10.1007/s11023-017-9420-y

Colombo, M., Elkin, L., & Hartmann, S. (2021). Being realist about Bayes, and the predictive processing theory of mind. *The British Journal for the Philosophy of Science*, *72*(1), 185–220. https://doi.org/10.1093/bjps/axy059

Danks, D. (2019). Probabilistic models. In M. Sprevak & M. Colombo (Eds.), *The Routledge handbook of the computational mind* (pp. 149–158). Routledge.

de Finetti, B. (1990). *Theory of probability* (Vol. 1). Wiley & Sons.

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in Cognitive Sciences*, *22*(9), 764–779. https://doi.org/10.1016/j.tics.2018.06.002

Deneve, S. (2008). Bayesian spiking neurons I: Inference. *Neural Computation*, *20*(1), 91–117. https://doi.org/10.1162/neco.2008.20.1.91

Dennett, D. C. (1991). *Consciousness explained*. Little, Brown & Company.

Dolega, K., & Dewhurst, J. E. (2021). Fame in the predictive brain: A deflationary approach to explaining consciousness in the prediction error minimization framework. *Synthese*, *198*(8), 7781–7806. https://doi.org/10.1007/s11229-020-02548-9

Drayson, Z. (2017). Modularity and the predictive mind. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing*. MIND Group. https://doi.org/10.15502/9783958573024

Eberhardt, F., & Danks, D. (2011). Confirmation in the cognitive sciences: The problematic case of Bayesian models. *Minds and Machines*, *21*(3), 389–410. https://doi.org/10.1007/s11023-011-9241-3

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for 'top-down' effects. *Behavioral and Brain Sciences*, *39*, E229. https://doi.org/10.1017/s0140525x15000965

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences*, *14*(3), 119–130. https://doi.org/10.1016/j.tics.2010.01.003

Forster, M. (2008). Prediction. In S. Psillos & M. Curd (Eds.), *The Routledge companion to philosophy of science* (pp. 405–413). Routledge.

Franklin, D. W., & Wolpert, D. M. (2011). Computational mechanisms of sensorimotor control. *Neuron*, *72*(3), 425–442. https://doi.org/10.1016/j.neuron.2011.10.006

Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, *16*(9), 1325–1352. https://doi.org/10.1016/j.neunet.2003.06.005

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London, Series B*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. https://doi.org/10.1016/j.tics.2009.04.005

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K., Fortier, M., & Friedman, D. A. (2018). Of woodlice and men: A Bayesian account of cognition, life and consciousness: An interview with Karl Friston. *ALIUS Bulletin*, *2*, 17–43.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2013). The anatomy of choice: Active inference and agency. *Frontiers in Human Neuroscience*, *7*, 598. https://doi.org/10.3389/fnhum.2013.00598

Gardner-Medwin, A. R., & Barlow, H. B. (2001). The limits of counting accuracy in distributed neural representations. *Neural Computation*, *13*(3), 477–504. https://doi.org/10.1162/089976601300014420

Gershman, S. J. (2019). What does the free energy principle tell us about the brain? *Neurons, Behavior, Data Analysis, and Theory*, *2*(3). https://doi.org/10.51628/001c.10839

Gregory, R. L. (1997). Knowledge in perception and illusion. *Philosophical Transactions of the Royal Society of London, Series B*, *352*(1358), 1121–1128. https://doi.org/10.1098/rstb.1997.0095

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263–268. https://doi.org/10.1177/0963721412447619

Grush, R. (2004). The emulator theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, *27*(3), 377–442. https://doi.org/10.1017/s0140525x04000093

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, *3*, 1–14. https://doi.org/10.3389/fpsyg.2012.00096

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

Hohwy, J. (2020). New directions in predictive processing. *Mind and Language*, *35*(2), 209–223. https://doi.org/10.1111/mila.12281

Hoyer, P. O., & Hyvärinen, A. (2003). Interpreting neural response variability as Monte Carlo sampling of the posterior. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15, pp. 277–284). MIT Press.

Icard, T. (2016). Subjective probability as sampling propensity. *Review of Philosophy and Psychology*, *7*(4), 863–903. https://doi.org/10.1007/s13164-015-0283-y

Jiang, L. P., & Rao, R. P. N. (2022). Predictive coding theories of cortical function. In S. M. Sherman (Ed.), *Oxford research encyclopedia of neuroscience*. https://doi.org/10.1093/acrefore/9780190264086.013.328

Kanai, R., Komura, Y., Shipp, S., & Friston, K. (2015). Cerebral hierarchies: Predictive processing, precision and the pulvinar. *Philosophical Transactions of the Royal Society of London, Series B*, *370*(1668), 20140169. https://doi.org/10.1098/rstb.2014.0169

Keller, G. B., & Mrsci-Flogel, T. D. (2018). Predictive processing: A canonical cortical computation. *Neuron*, *100*(2), 424–435. https://doi.org/10.1016/j.neuron.2018.10.003

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

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. https://doi.org/10.1016/j.tins.2004.10.007

Kok, P., & de Lange, F. P. (2015). Predictive coding in the sensory cortex. In B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience* (pp. 221–244). Springer.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247. https://doi.org/10.1038/nature02169

Körding, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, *10*(7), 319–326. https://doi.org/10.1016/j.tics.2006.05.003

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Lakatos, I. (1978). *The methodology of scientific research programmes: Philosophical papers, Volume 1*. (J. Worrall & G. Currie Eds.). Cambridge University Press.

Lasserre, J. A., Bishop, C. M., & Minka, T. P. (2006). Principled hybrids of generative and discriminative models. In *Proceedings of the 2006 IEEE computer society conference on computer vision and pattern recognition* (pp. 87–94). IEEE Computer Society.

Laudan, L. (1977). *Progress and its problems*. University of California Press.

Lupyan, G. (2015). Cognitive penetrability of perception in the age of prediction: Predictive systems are penetrable systems. *Review of Philosophy and Psychology*, *6*(4), 547–569. https://doi.org/10.1007/s13164-015-0253-4

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press.

Macpherson, F. (2012). Cognivite penetration of colour experience: Rethinking the issue in light of an indirect mechanism. *Philosophy and Phenomenological Research*, *84*(1), 24–62. https://doi.org/10.1111/j.1933-1592.2010.00481.x

Macpherson, F. (2017). The relationship between cognitive penetration and predictive coding. *Consciousness and Cognition*, *47*, 6–16. https://doi.org/10.1016/j.concog.2016.04.001

Marr, D. (1982). *Vision*. W. H. Freeman.

Matloff, N. (2017). *Statistical regression and classification*. CRC Press.

Metzinger, T., & Wiese, W. (2017). Vanilla PP for philosophers: A primer on predictive processing. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing*. MIND Group. https://doi.org/10.15502/9783958573024

Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, *108*(30), 12491–12496. https://doi.org/10.1073/pnas.1101430108

Neisser, U. (2014). *Cognitive psychology*. Prentice-Hall.

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems 14* (pp. 841–848). MIT Press.

Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences*, *12*(7), 265–272. https://doi.org/10.1016/j.tics.2008.03.006

Poeppel, D., & Bever, T. G. (2010). Analysis by synthesis: A (re-)emerging program of research for language and vision. *Biolinguistics*, *4*(2–3), 174–200. https://doi.org/10.5964/bioling.8783

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: Knows and unknowns. *Nature Neuroscience*, *16*(9), 1170–1178. https://doi.org/10.1038/nn.3495

Rahnev, D. (2017). The case against full probability distributions in perceptual decision making. *bioRxiv*, 108944. https://doi.org/10.1101/108944

Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences*, *41*, 1–66. https://doi.org/10.1017/s0140525x18000936

Raina, R., Shen, Y., McCallum, A., & Ng, A. Y. (2003). Classification with hybrid generative/discriminative models. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems* (Vol. 16, pp. 545–552). MIT Press.

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

Rescorla, M. (2018). Motor computation. In M. Sprevak & M. Colombo (Eds.), *The Routledge handbook of the computational mind* (pp. 424–435). Routledge.

Rescorla, M. (2020). A realist perspective on Bayesian cognitive science. In A. Nes & T. Chan (Eds.), *Inference and consciousness* (pp. 40–73). Routledge.

Roskies, A. L., & Wood, C. C. (2017). Catching the prediction wave in brain science. *Analysis*, *77*(4), 848–857. https://doi.org/10.1093/analys/anx083

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Pearson.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893. https://doi.org/10.1016/j.tics.2016.10.003

Sanborn, A. N., & Chater, N. (2017). The sampling brain. *Trends in Cognitive Sciences*, *21*(7), 492–493. https://doi.org/10.1016/j.tics.2017.04.009

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Schwartenbeck, P., FitzGerald, T., Mathys, C., Dolan, R. J., & Friston, K. (2015). The dopaminergic midbrain encodes the expected certainty about desired outcomes. *Cerebral Cortex*, *25*(10), 3434–34445. https://doi.org/10.1093/cercor/bhu159

Seth, A. K. (2017). The cybernetic brain: From interoceptive inference to sensorimotor contingencies. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing*. MIND Group. https://doi.org/10.15502/9783958570108

Seth, A. K. (2021). *Being you: A new science of consciousness*. Faber & Faber.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*(1), 1193–1216. https://doi.org/10.1146/annurev.neuro.24.1.1193

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition*, *112*, 92–97. https://doi.org/10.1016/j.bandc.2015.11.003

Sprevak, M. (forthcoming-a). *Predictive coding II: The computational level*. Philosophy Compass.

Sprevak, M. (forthcoming-b). *Predictive coding III: The algorithmic level*. Philosophy Compass.
Sprevak, M. (forthcoming-c). *Predictive coding IV: The implementation level*. Philosophy Compass.
Sprevak, M. (2022). Understanding phenomenal consciousness while keeping it real. *Philosophical Psychology*, *36*(2), 438–441. https://doi.org/10.1080/09515089.2022.2092465
Sprevak, M., & Smith, R. (2023). An introduction to predictive processing models of perception and decision-making. *Topics in Cognitive Science*. https://doi.org/10.1111/tops.12704
Sterling, P., & Laughlin, S. (2015). *Principles of neural design*. MIT Press.
Stone, J. V. (2018). *Principles of neural information theory: Computational neuroscience and metabolic efficiency*. Sebtel Press.
Strevens, M. (2017). Notes on Bayesian confirmation theory. http://www.strevens.org/bct/BCT.pdf
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788
Toderici, G., Vincent, D., Johnston, N., Hwang, S. J., Minnen, D., Shor, J., & Covell, M. (2017). Full resolution image compression with recurrent neural networks. *arXiv*, 1608.05148. https://doi.org/10.48550/arXiv.1608.05148
Usevitch, B. E. (2001). A tutorial on modern lossy wavelet image compression: Foundations of JPEG 2000. *IEEE Signal Processing Magazine*, *18*(5), 22–35. https://doi.org/10.1109/79.952803
von Helmholtz, H. (1867). *Handbuch der physiologischen Optik*. Leopold Voss.
Wolpert, D. M., Ghahramani, Z., & Flanagan, J. R. (2001). Perspectives and problems in motor learning. *Trends in Cognitive Sciences*, *5*(11), 487–494. https://doi.org/10.1016/s1364-6613(00)01773-3
Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308. https://doi.org/10.1016/j.tics.2006.05.002

AUTHOR BIOGRAPHY

Mark Sprevak is Senior Lecturer in Philosophy at the University of Edinburgh. His primary research interests are philosophy of mind, philosophy of science, and metaphysics, with particular focus on the cognitive sciences. He has published articles in, among other places, *The Journal of Philosophy*, *The British Journal for the Philosophy of Science*, *Synthese*, *Philosophy, Psychiatry & Psychology*, and *Studies in History and Philosophy of Science*.

How to cite this article: Sprevak, M. (2023). Predictive coding I: Introduction. *Philosophy Compass*, e12950. https://doi.org/10.1111/phc3.12950