 mark.sprevak@ed.ac.uk

Predictive coding II: The computational level

Mark Sprevak
University of Edinburgh

20 May 2022

1 Introduction

When we encounter a new computing device, we often try to describe its computational characteristics in terms of the task it faces: this shop's cash register has the task of *adding numbers*, this computer programme has the task of *sorting names into alphabetical order*, this Excel spreadsheet has the task of *calculating losses*. As well as asking a *how*-question about the device – How does it work? – we might ask a *what*-question: What is the problem the device is trying to solve? A theory at Marr's computational level aims to provide an answer to this question. It aims to identify the *computational task* that a device faces.[1]

What is the computational problem faced by the brain? Conventional approaches in computational cognitive science tend to start from the assumption that the brain faces many computational problems. Different aspects of cognition – e.g. perception, motor control, decision making, language learning – require the brain to respond to different information-processing challenges. Each challenge has its own computational nature and is likely to deserve its own Marrian computational-level description. On such a picture, it makes sense for computational cognitive science to adopt a *divide et impera* strategy to modelling cognition: it should break up

[1]Marr's use of the term 'computational' here is not meant to imply that his other levels of description are not computational. His usage of the term derives from mathematical logic, where a 'computational' theory denotes relationships between tasks that are blind to differences in algorithms or physical implementation (as in the identification of relations of computational equivalence).

human cognition into multiple constituent computational problems, each of which should be described in turn.

Predictive coding suggests that this *divide et impera* strategy, and the 'many problems' assumption on which it is based, is wrong. During cognition, the brain faces a *single* computational problem. At Marr's computational level, one unified story should be told. Apparent differences between different challenges encountered in perception, motor control, decision making, language learning, and so on mask an underlying unity that all these problems share. They are all instances of a single overarching problem: to *minimise sensory prediction error*.

Sections 2–4 attempt to unpack what is meant by this. Sections 5–8 turn to the claim's justification. I outline three main strategies an advocate of predictive coding might draw on to defend it: the *case-based* defence (Section 7), the *free-energy* defence (Section 8), and the *instrumental-value* defence (Section 9).

2 Minimising sensory prediction error

What does it mean to say that the brain is trying to minimise its sensory prediction error? As we will see, there are a variety of ways of formalising this task in mathematical language. However, an advocate of predictive coding often starts with an *informal* description of the task. Subsequent mathematical descriptions aim to codify this informal description more precisely and open it up to proposals that it is tackled by various numerical algorithms. In predictive coding, there is currently some degree of uncertainty about exactly the right way to formalise the task of minimising sensory prediction error in mathematical terms. However, there is broad agreement about the *informal* nature of the problem. We will begin with this informal description.

The task of *minimising sensory prediction error* may be informally characterised as follows. Brains have sensory organs and their sensory organs supply them with a continuous stream of input from the outside world. Brains also have complicated endogenous physical structures and activities that determines how they react to that stream of input. According to predictive coding, the computational task that a brain faces in cognition is to ensure that these endogenously generated responses (the brain's 'inference' over its 'generative model') cancel out or suppress the incoming flux of physical signals conveyed by the sensory organs from the outside world (that it 'predicts' the incoming 'sensory evidence'). The degree to which this happens, or fails to happen, is measured by the *sensory prediction error*. This quantity measures the discrepancy between the contribution of the brain's endogenously generated activities and the incoming physical signals from the world. According to predictive coding, the problem that the brain faces, in all aspects of cognition, is to minimise

this discrepancy. If the brain were to succeed at doing this then, at the sensory boundary, two opposing forces – the world's sensory input (excitatory/stimulating) and the brain's endogenously generated predictions (inhibitory/suppressing) – would exactly cancel out. The brain's anticipatory signal would 'quench' incoming excitation from the world. In more colourful and metaphorical language:

> … this is the state that the cortex is trying to achieve: perfect prediction of the world, like the oriental Nirvana, as Tai-Sing Lee suggested to me, when nothing surprises you and new stimuli cause the merest ripple in your consciousness. (Mumford, 1992, p. 247, n. 5)

Predictive coding is a theory is about the subpersonal computational machinery of cognition, not our conscious personal-level experience, but the basic idea is correct. The computational task the brain faces is to avoid being perturbed or surprised by incoming sensory inputs (in the Shannon sense of 'surprise', i.e. unpredicted). The brain's goal is to arrange itself and its physical responses to anticipate and cancel upcoming sensory input. This goal – 'Nirvana' in the above quotation – is unlikely to ever be achieved, or achieved in any sustained way, because the sensory inputs supplied by the world are too rich and complex for our brains to always predict them with perfect accuracy. Nevertheless, *trying* to predict them is the task the brain faces in cognition.

Predictive coders suggest that the various computational problems that the brain faces in perception, learning, motor control, decision making, and so on, are all instances of minimising sensory prediction error. Our various cognitive capacities (sensing, planning, and so on), which have traditionally been viewed as dissociable responses to distinct problems (faced in perception, motor control, and so on), should perhaps be reconceived as parts of a seamless, unified response by the brain to a single problem. This all suggests that we might need to rethink how we describe and individuate our cognitive capacities, and potentially blur the boundaries between them. It is in this sense that, at Marr's computational level, predictive coding aims to offer a grand, unified theory of cognition.

It worth stressing that is not novel or unusual to suggest that minimising sensory prediction error is *one* of the computational challenges faced by the brain. Contemporary models often suggest that early vision involves compression of sensory signals (Sprevak, forthcoming[a], Section 2) and certain inference and learning tasks are often described as minimising sensory prediction error (ibid., Section 4). What marks out predictive coding as special in this context is that it says that minimising sensory prediction error is the brain's *only* computational task. It is not one among many objectives pursued by the brain, but the only or the fundamental objective. The elevated status of this one task is the primary feature that differentiates predictive coding from other approaches.

3 Formal and informal descriptions

Theories at Marr's computational level are often precise and characterised in mathematical language. They are usually *formal* and *quantitative*. Typically, a theory at Marr's computational level will ascribe computation of a mathematical function to the brain as well as offering an explanation of why computing that function would help the brain solve a problem that is informally characterised. For example, in his account of vision, Marr ascribed computation of the mathematical function $\nabla^2 G * I$ to the brain. Marr related this problem to the informally characterised task of *edge detection*: finding the location of boundaries between objects in the visual field.[2] Marr argued that edge detection is an important problem that the brain faces in early vision and that solving it is a preliminary step to solving other problems such as object recognition, depth perception, or binocular fusion. Marr proposed that the informal task of edge detection could be more precisely described as the task of computing of this mathematical function.

In Marr's formal description, I is a two-dimensional matrix of numerical values. These values quantify the magnitude of light falling on a two-dimensional array of photoreceptors on the retina. G is a Gaussian filter which is convolved ($*$) with the two-dimensional image (I) and the Laplacian, second-derivative operator (∇^2) is applied to the result. Marr argued that if the brain were to compute the zero-crossings of this function for various sizes of Gaussian filter, it would identify areas in the retinal image that correspond to sharp changes in light intensity. These, Marr argued, tend to coincide with the edges of objects in the visual field. Hence, the task of computing the zero-crossings of this mathematical function provides a precise, mathematically codified formalisation of the (informally characterised) problem of edge detection.[3]

One way in which this relationship is described is that between a 'what' element and a 'why' element of a computational-level theory.[4] The 'what' element in a computation-level theory describes the mathematical function that the device needs to compute. In the above case, this would be $\nabla^2 G * I$. The 'why' element links the task of computing that mathematical function to some informally characterised information-processing problem. It draws a connection between the values that

[2] Marr (1982), pp. 68–74. The full story about the informal task is complex, and 'edges' should be understood to include not only the boundaries of objects, but also regions of the visual field where there are changes in reflectance, illumination, depth, or surface orientation.

[3] Marr thought that this computation was accomplished by the action of retinal ganglion cells: 'Take the retina. I have argued that from a computational point of view, it signals $\nabla^2 G * I$ (the X channels) and its time derivative $\partial/\partial t(\nabla^2 G * I)$ (the Y channels). From a computational point of view, this is a precise specification of what the retina does.' (Marr, 1982, p. 337).

[4] See Marr (1982), p. 22.

feature in the function and physical quantities and the concrete adaptive problems faced by the embodied device. In the case above, it involves explaining why computing $\nabla^2 G * I$ would help an embodied system solve the problem of detecting edges in the visual field. Marr's 'what' element provides a formal, mathematical characterisation of the task; the 'why' element explains the appropriateness and adequacy of that mathematical description to the task as informally conceived.[5]

There are many possible ways one might attempt to formalise the task of minimising sensory prediction error. Predictive coding has not yet settled on a single canonical formalisation. A simple example of a formalisation is given in Sprevak (forthcoming[b]), Section 2.1.[6] However, even in more complex mathematical treatments, it is common to assume a highly simplified or stripped-down version of the task. For example, it is common to consider systems with only one or two sensory input channels, to only attempt to minimise their current prediction errors, or to only consider predictions from a linear generative model. Such simplifications not only help to keep the formalisation manageable, they may also serve to highlight specific features of interest in the intended model.

Nevertheless, some broad generalisations can be made about predictive coding's formal task description. All mathematical formalisations tend to treat the task as a numerical *optimisation problem*. That problem is regarded as having two free variables – the *generative model* and the *prediction values*. Those variables are changed, on different timescales, in order to find the global minimum of an objective function – the *sensory prediction error*. In the simplest case, the generative model is formalised as a two-dimensional matrix of values. Prediction values are formalised as a vector that, when combined with a generative model by multiplication, produce another vector, the *sensory prediction*. The *sensory input* is another vector with the same dimensionality, each of whose components encode the actual incoming activity of each physical sensory channel. The *sensory prediction error* measures how close the sensory prediction is to the sensory input. It is often treated as the (weighted) sum or mean of the squares of the difference between the sensory input vector and the sensory prediction vector. The task the brain faces is to select prediction values and generative model such that its prediction errors over sensory inputs are minimised. Describing the problem in this way allows many existing numerical optimisation algorithms – including the vast range of algorithms that employ some form of gradient descent – to be brought to bear as proposals about how the brain attempts to solve its problem.

[5] See Shagrir and Bechtel (2013); Shagrir (2010) for a helpful explanation of the 'what' and 'why' at Marr's computational level.

[6] A range of other formalisations can be found in Bogacz (2017); Friston (2003); 1330–1339; Friston (2005), pp. 819–821; Friston (2009), p. 296; Spratling (2017), pp. 92–93.

4 Precision weighting of prediction errors

An important element that has not yet been mentioned is that not all sensory prediction errors are weighted equally in the task of minimising sensory prediction error. Predictive coding has a third type of variable, *precision weighting*, which describes the relative weight of each sensory prediction error. The brain's task is thus to minimise its *precision-weighted* sensory prediction error. Errors that have a high degree of precision weighting should be prioritised during this; errors that have a low precision weighting are given a lower priority or ignored. Precision weighting describes a scaling factor or 'gain' that is applied to each component of the sensory prediction error.

Precision weighting is a critically important part of predictive coding's task description. It can make certain sensory prediction errors dominate the optimisation process and others small enough to be irrelevant. It can exercise this control in very fine-grained, nuanced ways. Precision weighting can potentially modify the gain on prediction errors associated with each individual sensory channel independently. Precision weighting is usually treated as a distribution that determines which sensory prediction errors are boosted and which are dampened down at any given moment. The shape of that distribution may be complex and it may change radically and rapidly over time (e.g. within milliseconds). Formally, and in the simplest case, precision weighting is represented as a two-dimensional matrix that is multiplied by the raw sensory prediction error vector to scale its elements.[7]

Precision weighting plays a number of conceptually distinct functional roles within predictive coding. First, under a probabilistic interpretation of predictive coding's algorithm, it is assumed to be connected to the brain's estimation of *uncertainty* associated with its sensory predictions. Predictions about which the brain is more certain have a smaller variance, which is equivalent to them having a higher precision weighting associated with their corresponding prediction errors (Friston, 2003).[8] Second, precision weighting is claimed to be connected to the *direction of fit* of sensory predictions. Sensory prediction errors with a high degree of precision weighting are the ones that will dominate the optimisation process and on which the brain is more likely to act; they will function as quasi motor commands (see Section 7). In contrast, prediction errors with a low degree of precision weighting are less likely to feed into action and might be used to simulate or imagine actions of the agent or of other agents without danger of producing an associated motor response (Clark, 2016; Friston, Mattout and Kilner, 2011, Ch. 5; Pickering and Clark, 2014).[9] Third, precision weighting is claimed to be connected to the allocation of *attention*.

[7] See Sprevak (forthcoming[b]), Section 2.4.
[8] See Sprevak (forthcoming[b]), Section 5.
[9] See Sprevak (forthcoming[b]), Section 6.1.

When the cognitive system attends to certain features, the components of the sensory signals associated with those features are the ones for which the corresponding prediction errors have been assigned a higher precision weight. When the cognitive system shifts the focus of its attention, this entails a rebalancing of the distribution of precision weightings away from those features (Feldman and Friston, 2010).[10] Finally, and most controversially, precision weighting is sometimes used as a kind of 'fudge factor' to accommodate observations that do not straightforwardly fit into the prediction-error-minimisation framework. If the brain fails to minimise a sensory prediction error, then an advocate of predictive coding might interpret that failure, not as evidence against predictive coding, but as evidence that the brain has assigned a low precision weighting to that particular sensory error. If a scientist is allowed to assume any distribution of precision weightings at any moment in time, almost any observation can be accommodated under predictive coding's task description.[11] Obviously, constraints are needed on how precision weightings are assigned to a brain. Finding a sufficient number of empirically motivated constraints on this remains an open problem for predictive coding.[12]

The distribution of precision weighting intuitively captures 'what matters' to the brain when it is attempting to minimise its sensory prediction error. No version of predictive coding can afford to omit precision weighting: it would simply be implausible to think that every sensory prediction error matters equally to the brain. However, introducing precision weighting into predictive coding's task description raises a number of puzzles. It plays many roles within predictive coding's model and it is not obvious how all those various roles cohere. It is also not clear which independent, empirically motivated constraints lie on the assignment of precision weightings considering its tremendous power to reshape the computational problem facing the brain.

5 Long-term prediction error and the dark-room objection

A second important element of the task description not yet mentioned is that the objective should be understood as that of minimising *long-term* sensory prediction error. That goal might be glossed in various ways, with expressions such as 'global' prediction error (Lupyan, 2015), 'upcoming' prediction error (Muckli, 2010, p. 137),

[10] See Sprevak (forthcoming[c]), Section 5.

[11] See Clark (2013a) for examples of how precision weighting can explain a range of otherwise puzzling cases for the view (e.g. habit-based action and behaviour during model-free learning). See Miller and Clark (2018), p. 2568 for their response to the objection that precision weighting functions as a 'magic modulator' that allows predictive coding to accommodate every possible behaviour.

[12] For further discussion of this problem, see Sprevak (forthcoming[c]), Section 8.

'long-term average' of prediction error (Hohwy, 2013, p. 90, 175, 176), or 'long-term average surprise' (Schwartenbeck et al., 2013).

The mathematical nature of this long-term objective is, however, not entirely clear. It is to minimise some form of average of individual (precision-weighted) sensory prediction errors over time. However, what type of average, and how far in time that period should extend, is not clear. It is unknown whether, and to what degree, future prediction errors should be discounted. It is unknown whether the objective should be to reduce prediction errors relative to the system's own expectations (its subjective probability) of making future sensory prediction errors, or relative to the objective chances (objective probability) of it making such errors. It is unknown whether the relevant time period to minimise errors is of the order of hours, days, years, the entire future lifespan of the organism, or further to include the lifespans of all its possible descendants and evolutionary successors. It is unknown how this average (which weights prediction errors over time) should interact with precision weighting (which weights the current error signals) – i.e. whether precision weighting should be understood as having a prospective component to allow the brain to preferentially discount certain expected future errors over others. These open questions suggest that alternative formulations of predictive coding could be developed at the computational level.

Nevertheless, acceptance that the brain aims to minimise a long-term measure plays an important role in clarifying and lending plausibility to predictive coding's task description. For one thing, it allows one to understand how predictive coding could respond to the infamous 'dark room' objection. For another, it suggests that predictive coding is compatible with inferences and behaviour that tend to drive up short-term sensory prediction error, such as curiosity, exploration, and novelty seeking.

The dark-room problem is a long-standing objection to predictive coding.[13] The dark-room problem is to explain why, if predictive coding's description at the computational level is correct, a cognitive agent would not simply seek out the most predictable possible environment, such as a dark room, and remain inside for as long as possible. If the goal of cognition is to minimise sensory prediction errors, why not maximise the chances of achieving this by staying in a maximally predictable environment?

Friston, Thornton and Clark (2012) offered an initial reply to the dark-room problem.[14] Their response focuses on the idea that our generative model and prediction values are not infinitely malleable: there are limits to the kinds of predictions we

[13] See Clark (2013b), p. 193 for a statement of the problem.

[14] See also Hohwy (2013), pp. 87, 185; Clark (2016), pp. 265–268;

can generate and to how much our generative model and prediction values can be revised. These constraints, primarily due to our physical hardware, are assumed to be immune to change by learning or inference, and are called 'hyperpriors'. Human hyperpriors bias us towards making certain kinds of predictions, and not ones that are not particularly suited to life in a dark room. Although the sensory data inside a dark room might be 'easy to predict' in some disembodied sense, they might be difficult for a creature *like us* to predict. If we were a different type of creature, one that had evolved with different hard-wired biases (maybe a cave-dwelling creature), we might have no trouble in reliably generating accurate sensory predictions inside a dark room. However, we are biased to predict sensory data that come from bright, changeable environments, and so we are unlikely to minimise our sensory prediction errors inside a dark room.

This response highlights an important but as yet unmentioned point about predictive coding's task description: the problem brain faces is a *constrained* optimisation problem. The brain's objective is to minimise sensory prediction error by varying a generative model and prediction values *given* the constraints imposed by our physical hardware about how far and how rapidly that generative model and those prediction values can vary. The literature on predictive coding's computational-level proposal tends to be silent about the specific nature of these physical constraints. However, a crucial part of making the view plausible is to acknowledge that a range of constraints on the optimisation problem are implicitly there.[15]

Friston, Thornton and Clark (2012)'s reply brings to the fore an important feature of predictive coding's computational-level description, but it does not fully address the concerns that motivated the dark-room problem. For example, it does not explain why, *even relative to a constrained model*, cognitive agents like ourselves still seek out novelty and surprise. Even when we can predict a situation, we sometimes choose a more surprising alternative. In other words, cognitive agents like ourselves sometimes *prefer* novelty to predictability. How is that consistent with what predictive coding says at the computational level?[16]

An alternative reply that fares better at addressing this kind of objection is to emphasise the long-term nature of the brain's objective. The world in which we live contains both environments that are easy to predict and environments and that are hard to predict (for us). Successfully predicting our sensory inputs only where we can already do so may not, over the long term, be a good solution to the brain's

[15]We will see some constraints flow from what predictive coding says at the algorithmic level and implementation level (Sprevak, forthcoming[b], Section 2.5; Sprevak, forthcoming[c], Section 7). However, as will become clear, what predictive coding says at those levels is by no means a complete account of the relevant constraints faced by the brain in inference or learning.

[16]See also Clark (2016), pp. 265–266

problem. An agent who sequesters itself inside an easy-to-predict environment leaves itself a hostage to fortune. Unpredictable elements may intrude on the agent in ways that it has not taken the trouble to learn how to handle – light might enter the room, a stranger might enter, food supplies might run out. To guard against future surprises and an associated rise in sensory prediction error, it may be better – purely in the terms of the long-term goal of minimising sensory prediction error – to leave an environment that is easy to predict and engage in some exploration to learn a more comprehensive model of the world. Exploring environments that are harder to predict might raise current sensory prediction errors, but it is a hedge against future, possibly bigger surprises that an agent who led an entirely sheltered life would not be able to avoid. There is obviously a balance to strike here between the cost of exploring (in terms of a rise in current sensory prediction error), and its potential future pay-off (in terms of a reduction in long-term sensory prediction error). But that there is a trade-off between the value of exploration and exploitation is to be expected on any model of cognition. The important point is that what predictive coding says at the computational level allows for the possibility that a cognitive agent may sometimes prefer unpredictable environments to predictable ones. Curiosity, exploration, and novelty seeking are consistent with the brain minimising a long-term measure of sensory prediction error, even if they entail a rise in that error along the way (Schwartenbeck et al., 2013).

6 Evidence for predictive coding

Justification for predictive coding's computational-level claim often rests on one of three strategies. I call these the *case-based* defence, the *free-energy* defence, and the *instrumental-value* defence. The case-based defence considers a range of cognitive tasks and aims to show that all of these tasks can and should be described as minimising sensory prediction error. The free-energy defence shortcuts consideration of individual tasks and attempts to establish predictive coding's computational-level general claim by appeal to Karl Friston's free-energy principle. The instrumental-value defence focuses on the utility of predictive coding's task description to computational cognitive science and argues that it provides a desirable set of heuristics to make sense of, and discern patterns within, the mass of human behavioural and neural responses.

7 The case-based defence

The case-based defence is an abductive argument. It attempts to show that a number of tasks facing the brain – for example, during perception, decision-making, planning, motor control – can and should be thought of as instances of the single

task of minimising sensory prediction error. Some of those tasks may already have computational-level descriptions associated with them based on rival or more traditional computational research programmes. The job of predictive coding is to show that these can and should be stated as instances of minimising sensory prediction error. Behavioural and neural responses that might previously have been categorised as attempts by the brain to compute some domain-specific mathematical function should be redescribed in the manner predictive coding suggests.

Any case-based argument for predictive coding faces an obvious epistemic hurdle. Predictive coding makes a universal claim – *every* problem the brain encounters in cognition is to minimise sensory prediction error. Showing that this holds in a limited number of cases (e.g. in aspects of early vision) does not entail that it holds in other, perhaps as yet unconsidered cases (e.g. language learning). No amount of success in applying predictive coding's task description to limited domains of cognition demonstrates that in *every* case the problem the brain faces is minimisation of sensory prediction error. Nevertheless, science is rife with universal generalisations made on the back of observations about a limited number of cases. The non-demonstrative nature of such arguments is not in principle an objection to using them. However, there are clearly more and less effective ways of making such an abductive argument work.

One plausible strategy is to focus on a *diverse* range of cases – what one might hope is a *representative* sample of cases. Early work on predictive coding focused on sensory compression in the early visual system (Atick, 1992; Rao and Ballard, 1999; Srinivasan, Laughlin and Dubs, 1982). Ideally, predictive coding should seek support for its wider claim by showing that other kinds of behavioural and neural response fall under predictive coding's task description. If it can be shown that many behavioural and neural phenomena that have no obvious connection to each other, or to early vision, can and should fall under predictive coding's task description, then that would lend credence to the idea that not just in some cases, but in every case, the problem the brain faces is sensory prediction error minimisation. Example of such 'non-obvious' applications of predictive coding include music perception (Koelsch, Vuust and Friston, 2019); formation of emotions and judgements about bodily ownership (Seth, 2013); binocular rivalry (Hohwy, Roepstorff and Friston, 2008); formation of judgements about the nature of the self (Hohwy and Michael, 2017); and the perceptual, doxastic, and motor characteristics of schizophrenia and autism (Corlett and Fletcher, 2014; Fletcher and Frith, 2009; Friston, Stephan et al., 2014; Pellicano and Burr, 2012).

It is worth noting that, with respect to each individual case, a case-based argument requires one to meet two separate challenges. The first challenge is to show that the case in question *can* be described as an instance of sensory-prediction-error

minimisation. The second is to show that it *should* be described this way. The first challenge requires one to show that predictive coding's computational-level description is *consistent* with the behavioural or neural data associated with that case. The second is to show that cognitive psychology should *prefer* predictive coding's computational-level description of that data to rival or more traditional accounts. There should be some net *benefit* to adopting predictive coding's computational-level treatment of that instance of cognition – e.g. in terms of increased predictive accuracy, increased explanatory power, or some other epistemic virtue.

Predictive coding's flagship example of a 'non-obvious' application of its computational-level description is *motor control*.[17] Traditional computational approaches to cognition tend to treat perception and motor control as entirely separate problems. In perception, the task facing the brain is to use its sensory data and background knowledge to build an accurate (or an instrumentally adequate) model of the world. In motor control, the task facing the brain is to use that model, along with some set of goals or intentions, to output a sequence of motor commands that would direct muscle actuators towards accomplishing those goals or intentions. Of course, motor control might partly rely on solving the perceptual problem. Motor control problems often require an agent to first build an accurate perceptual model of the world. Rapid and complex motor control might also require online regulation by sensory predictions from a forward model (Franklin and Wolpert, 2011). However, even if the problems of motor control and perception have some degree of overlap, they remain distinct problems: the task of perception is to create an accurate model of the world; the task of motor control is to use that model to generate motor commands to fulfil goals.

According to predictive coding, perception and motor control are instances of the same problem, namely, that of minimising sensory prediction error. In perception, the brain minimises sensory prediction error by varying its generative model and prediction values to anticipate upcoming sensory input. In motor control, the brain minimises its sensory prediction error by varying its bodily position and the external world (via muscle actuators) to change its incoming sensory stream to make its internally generated sensory predictions more likely to be true. In both cases, the objective is the same – to minimise sensory prediction error. The difference lies in the method the cognitive system uses to try to achieve it. Advocates of predictive coding call the first method 'passive' inference and the second 'active' inference. Passive and active inference (perception and motor control) are claimed to be complementary strategies employed by the brain to address what is fundamentally the same problem. According to predictive coding, the task of reaching for a glass

[17] See Friston (2010), pp. 133–134; Friston, Daunizeau et al. (2010); Clark (2016), Section 4.5; Hohwy (2013), Ch. 4.

of water should be reconceptualised as the brain making the prediction that the hand is already holding the glass of water (along with all its sensory consequences), and then solving its problem – minimising its sensory prediction error – by varying its limbs and the glass to make this false sensory prediction true.[18]

Even if perceptual tasks and motor tasks *can* both be described as instances of sensory prediction error minimisation, it remains a further question whether they *should* be described this way. The justification for this second step is often not obvious. The benefits of predictive coding's proposed task description are not straightforward to calculate and they need to be estimated relative to a wide range of epistemic standards, interests, and goals in computational cognitive science. Different researchers may, with good reason, take different views about the value of the benefits on offer.[19] As we will see shortly, those benefits are also often presented as conditional on accepting other elements of predictive coding's research programme (e.g. the universal scope of its claim, or elements of its proposals at the algorithmic or implementation levels).

To illustrate how these questions about the benefits of predictive coding's approach might be addressed, we will switch to a simpler case: the early visual system. Two main strategies have been used to defend predictive coding's computational-level description in this context: (i) appeal to what it can explain and predict relative to more traditional computational approaches; (ii) appeal to broader theoretical virtues offered by the view (e.g. its simplicity, elegance, and unifying power).

The first set of considerations surround predictive coding's ability to predict and explain behavioural or neural responses that are generally regarded as puzzling or anomalous on other views. Traditional computational-level characterisations of the early sensory system suggest that its computational task is to function as a Gabor filter bank on retinal images and thereby extract ecologically salient stimulus features such as orientation, spatial frequency, colour, direction of motion, and disparity (Carandini, Demb et al., 2005). In formal terms, the computational task of the early visual system is to convolve a matrix of retinal data with a variety of Gabor filters to, e.g., pick out lines in the visual field of various orientation and spatial frequency. However, many responses of neurons observed in the early visual system do not fit that description (Olshausen and Field, 2005). These so-called 'non-classical' effects count as anomalies relative to the visual system's claimed

[18] As well proposing a unified account of the problem facing the brain in perception and motor control, predictive coding also suggests that the algorithms that govern perceptual and motor processing have a great deal in common (see Sprevak, forthcoming[b], Section 6.1).

[19] For the benefits of predictive coding's task description of motor control see Friston (2011), Friston, Daunizeau et al. (2010); Wiese (2017), Pickering and Clark (2014). For benefits of alternative approaches, see Kording (2007); Shadmehr and Krakauer (2008).

objective. One such 'non-classical' effect is *end-stopping*: some V1 neurons give a strong response to a line at a particular orientation in the visual field, but that response is reduced or eliminated if the line extends outside the neuron's receptive field. End-stopping is inconsistent with a simple Gabor-filter description of their computational role: a classical Gabor filter should continue to fire regardless of whether a line extends outside its receptive field. End-stopping is categorised as anomalous under traditional computational-level descriptions of the early visual system.

Predictive coding suggests that the computational task faced by the early visual system is not to perform Gabor filtering, but to contribute to minimising the brain's sensory prediction error. Under predictive coding's task description, the behaviour of the relevant neurons within V1 may be reinterpreted as signalling the difference between the current sensory input and the brain's sensory prediction (based on its statistically-informed expectations regarding likely visual input). In our environment, the statistical norm is for lines in our visual field to extend beyond the tiny regions covered by the receptive fields of individual neurons. Lines that violate this expectation are unusual and, everything being equal, should be expected to generate prediction errors. The behaviour of V1 cells when end-stopping may be reinterpreted as signalling such sensory prediction errors (Kok and de Lange, 2015; Rao and Ballard, 1999, p. 232). End-stopping, not accommodated by traditional computational-level descriptions, can potentially be accommodated under predictive coding's computational-level description.[20]

A second set of motivations for preferring predictive coding's computational-level description surround its general theoretical virtues such as its simplicity, scope, and unifying power with respect to other approaches. Arguably, even if predictive coding were to do no better than any other model at accommodating various behavioural/neural effects, these general theoretical virtues might still lead one to favour the view. As observed in Section 1, traditional computational-level approaches to cognition tend to adopt a *divide et impera* approach and assume that the brain is facing multiple computational problems. On such a view, the brain is treated as an inherently multifunctional device, not a device tuned to solve just one problem.[21] A description of human cognition at Marr's computational level would be expected to consist in a patchwork of disjoint theories describing each computational problem

[20]For other examples of non-classical effects in the early visual system that appear to be accommodated by predictive coding, see Jehee and Ballard (2009); Kok, Jehee and de Lange (2012); Hosoya, Baccus and Meister (2005); Rao and Sejnowski (2002); Muckli (2010); Kok and de Lange (2015); Spratling (2010); Alink et al. (2010); Murray et al. (2002). For alternative computational-level accounts of these non-classical effects (e.g. in terms of divisive normalisation), see Aitchison and Lengyel (2017), p. 224; Carandini and Heeger (2012); Schwarz and Simoncelli (2001).

[21]For example, see Allen (2017); Carruthers (2006); Bayne et al. (2019).

the brain faces. Stepping back from that patchwork, there need be no overarching pattern or unity. Each domain of cognition – perception, motor control, decision making, language learning – is likely to merit its own computation-level account. Predictive coding, in contrast, provides a complete, unified, and relatively simple description of the computational task the brain faces in all aspects of cognition. That, by itself, would appear to be a mark in its favour. All else being equal it is rational to prefer a simple, unifying theory (where available) over less unified alternatives:

> It is the first time that we have had a theory of this strength, breadth and depth in cognitive neuroscience … I take that property as a sure sign that this is a very important theory … Most other models, including mine, are just models of one small aspect of the brain, very limited in their scope. This one falls much closer to a grand theory. (Stanislas Dehaene quoted in Huang, 2008)

A unified computational-level theory also promises to reveal something profound about the metaphysical nature of cognition. It tells us that cognition is not a motley, a jumble of distinct phenomena; it can be characterised in terms of a single computational problem. Predictive coding identifies what the various, seemingly distinct and unrelated domains of human cognition – perception, motor control, decision making, language learning – have in common. Moreover, it appears to explain *why* they each count as instances of cognition. It potentially provides us with a criterion to judge whether new and perhaps controversial instances of cognition are genuinely cognitive.[22] It suggests that cognition is a unified and relatively simple functional kind. If a theory uncovers principles like this – that unify and simplify what otherwise appears to be a complicated and disordered domain – then, all else equal, that is reason to favour it. Knowledge about the essence of things and the patterns into which they enter is surely what science aspires to.

8 The free-energy defence

Pursuit of the case-based defence of predictive coding is likely to be long and fraught. It requires engaging with the details of many specific cognitive tasks and showing that their distinctive effects – of which there may be many – are captured or recaptured on predictive coding's task description. A case-based defence has no obvious stopping point. A defender of predictive coding faces a potentially endless sequence of battles: there will always be more tasks, more behavioural and neural effects to consider. It is not obvious when enough cases – or a diverse enough selection of cases – will have been considered to justify the conclusion

[22] For predictive coding as a potential 'mark of cognitive', see Clark (2017); Kirchhoff and Kiverstein (2021); Ramstead et al. (2021).

that not just *some* tasks, but *every* task faced by the brain, is sensory prediction error minimisation. The free-energy defence aims to shortcut all this. It attempts to establish predictive coding's computational-level claim in one fell swoop by appealing to general properties shared by all cognitive (or living) systems. Friston (2010) presents a defence of predictive coding along these lines based on his 'free energy' formulation of predictive coding. Friston proposes that the task faced by the brain is that of *minimising free energy*. Minimising free energy can be shown, under appropriate further assumptions, to be equivalent to the task of minimising sensory prediction error.

Free energy is a mathematical quantity that appears inside classical thermodynamics, statistical mechanics, and information theory. Friston's claim is that there is a relationship between two distinct applications of the mathematical abstraction of free energy: *variational* free energy and, what I will call, *homoeostatic* free energy.[23] Variational free energy is an information-theoretic quantity predicated of agents who engage in probabilistic inference. If a probabilistic reasoner minimises their variational free energy, this can be shown to be equivalent to them approximating Bayesian inference (see Sprevak, forthcoming[d], Section 1). Under further assumptions, minimising sensory prediction error can also be shown to be equivalent to minimising variational free energy (see Sprevak, forthcoming[d], Section 2). 'Homoeostatic' free energy applies the same formal construct to a different set of properties. Unlike variational free energy, it is not (or at least, not directly) associated with the subjective probabilities that feature in probabilistic or Bayesian inference. Rather, it is associated with the objective probability of the macroscopic physical state the agent is in given its physical environmental conditions. Minimising homoeostatic free energy is associated with the agent's survival within a narrow band of macroscopic physical state types ('being alive'). According to Friston, these two types of free energy – homoeostatic free energy and variational free energy – are connected. Agents who minimise their homoeostatic free energy – who survive and maintain homeostasis – also minimise their variational free energy (and hence, given certain assumptions, minimise their sensory prediction error).

Friston is clear that neither variational nor homoeostatic free energy is the same as thermodynamic free energy. Thermodynamic free energy measures the useful mechanical work that can be extracted from a physical system. It is usually defined in terms of that system's ability to exert macroscopic mechanical forces on its surroundings – its energy that is 'free' to perform mechanical work. This is normally formalised as a difference between the physical system's internal energy and its thermodynamic entropy (its internal energy that is 'useless' for work). Having a reserve of thermodynamic free energy is generally a good thing for a cognitive or living

[23] Friston does not use these terms. He refers to both as 'variational' free energy.

creature: a surplus of thermodynamic free energy is a prerequisite for it to be able to move or act in the world. *Minimising* thermodynamic free energy would make little sense as a rule for cognition or survival. Friston is explicit that his principle – that all cognitive/living systems aim to minimise their homoeostatic/variational free energy – is not meant to be somehow a consequence of, or a principle about, thermodynamic free energy. He justifies his free-energy principle not on thermodynamic grounds, but on what he calls 'selectionist' grounds: all cognitive/living creatures strive to minimise their homoeostatic free energy because if they did not, they would tend to die off and hence be less likely to reproduce or to be observed by us.[24] Friston suggests that the only connection between thermodynamic free energy and his notion of free energy is their shared mathematical form.[25]

In outline, the logic of the free-energy defence of predictive coding is as follows. Its starting point is the observation that all cognitive (and living) creatures face the problem of surviving and maintaining homeostasis. That task, according to Friston, can be formalised as the problem of minimising a particular free-energy measure (what I have called homoeostatic free energy). Friston claims that minimising homoeostatic free energy entails that the creature also minimises a second free-energy measure associated with the creature's subjective probabilistic guesses (variational free energy). Minimising variational free energy, given certain further assumptions (detailed in Sprevak, forthcoming[d], Section 2), entails that the creature minimises its sensory prediction error. Hence, cognitive and living creatures, because they face the problem of survival and maintaining homeostasis, face the problem of minimising sensory prediction error.

There is much to unpack here.

First, the argument relies on a tight connection between homoeostatic and variational free energy. However, the nature of, and justification for, that connection is not obvious. Homoeostatic free energy pertains to how well the creature maintains its physical state within the narrow band associated with survival and homeostasis in the face of actual and possible perturbations from a changing physical environment. Living creatures change their microscopic physical state all the time. When they do so, they risk undergoing a fatal phase transition in their macroscopic physical state. When living systems resist this tendency – when they survive and maintain homeostasis – they minimise their homoeostatic free energy. Minimising homoeostatic free energy involves the creature trying to arrange its macroscopic state so as to avoid being overly changed by likely environmental physical transitions (Friston, 2013; Friston, Kilner and Harrison, 2006; Friston and Stephan, 2007). In contrast, variational free energy is predicated of an agent's subjective probability distribu-

[24] Friston and Stephan (2007), pp. 419–420, 451; Friston, Kilner and Harrison (2006), p. 85

[25] See Friston and Stephan (2007), p. 419.

tions. It measures how far the agent's probabilistic guesses depart from the optimal guesses of a perfect Bayesian observer armed with the same evidence.[26] According to Friston's formulation, the brain's task is to minimise variational free energy and so approximate an ideal Bayesian reasoner in inference. Minimising variational free energy makes the sensory data stream less surprising (in the Shannon sense), and therefore tends to drive down the agent's sensory prediction error (granted certain additional assumptions, see Sprevak, forthcoming[d], Section 2).

Homoeostatic free energy and variational free energy have certain features in common. They are both information-theoretic quantities and they are both measured over probability distributions. However, they are not the same. Homoeostatic free energy is measured over the *objective* probability distributions of macroscopic physical states that could occur; variational free energy is measured over the *subjective* probability distributions entertained by an agent about what could occur. Homoeostatic free energy is defined over the chances of various possible (fatal) physical states of the agent occurring in response to environmental changes; variational free energy is defined over the subjective probabilistic guesses the agent might make. The respective probability distributions might involve different sets of events, the distributions might have different shapes, and they each involve different types of probability (subjective and objective). There might, for various reasons, be a correlation between the two types of free energy, but it is not obvious that minimising one entails minimising the other.[27]

To see this more clearly, consider the tight relationship already mentioned between minimising variational free energy and Bayesian inference. An agent who minimises its variational free energy *ipso facto* approximates an ideal Bayesian reasoner. In many circumstances, a Bayesian agent would be well placed to survive and maintain homeostasis. But the precise nature of the connection between *being Bayesian* and *maximising one's chances of physical survival and homeostasis* is far from obvious. A non-Bayesian agent might live in a 'irrationality friendly' environment that maintains its homeostasis and physical integrity, even if it does not update its subjective probability distributions that represent its environment according to Bayesian norms. Conversely, an ideal Bayesian reasoner might live in a 'rationality hostile' physical environment that changes so rapidly and dramatically that it fails to survive or maintain homoeostasis, even if it updates its subjective probability distributions quickly and accurately according to Bayesian norms. Bayesian reasoning is plausibly related to survival, but it is not obvious in what sense it would guarantee it. Currently, the exact nature of the relationship between Friston's two measures of

[26] See Sprevak (forthcoming[d]), Section 1 for the connection between variational free energy and Bayesian inference.

[27] For further discussion of this point, see Sprevak (2020), pp. 602–604.

free energy – homoeostatic and variational – is unclear and the subject of ongoing analysis.[28]

At least two other aspects of the free-energy defence invite further scrutiny.

First, the predictive coding research programme aims to defend a universal claim: *every* task that the brain faces can and should be described as minimisation of sensory prediction error. Survival/homoeostasis is clearly one task faced by the brain, and an important one. If the internal logic of the free-energy defence is correct, then because the brain faces that task it also faces the task of minimising sensory prediction error. But it is not obvious how this reasoning is meant to generalise. Plausibly, our brains face other challenges that may be unrelated, or even in tension with, our long-term survival or homoeostasis – e.g. problems of mate selection, fulfilment of social roles, or arbitrary challenges set in the classroom or wider social environments. It is not clear how the free-energy defence is intended to handle these cases. The free-energy defence appeals to a formal connection between survival/homoeostasis and minimising sensory prediction error, but it is largely silent about how problems that do not (or do not obviously) improve the chances of our long-term survival/homoeostasis are meant to be related to minimising sensory predictive error. Even if the internal logic of the free-energy defence is correct, it is unclear how it supports the claim that every aspect of cognition is sensory prediction error minimisation.

Second, recall that the case-based defence required one to show not only that every problem faced by the brain in cognition *can* be described as sensory prediction error minimisation, but also that it *should* be described that way. The free-energy defence appears to only speak to the first issue. It attempts to establish a formal relationship between the task of survival/homoeostasis and the task of minimising sensory prediction error. However, even if such a connection were to exist, it would say nothing about the merits of one task description over the other. In order to address that issue, one would need to go beyond a purely formal equivalence between the task descriptions and consider the *value* of predictive coding's proposed description with respect to the wider standards, interests, and goals in cognitive neuroscience. *Why* should we describe the task facing the brain as sensory prediction error minimisation, even if, as the free-energy defence suggests, we could? That part of the argument remains to be made and it is likely to depend, at least in part, on an examination of the benefits offered by predictive coding's computational-level description to specific cases of interest to cognitive neuroscience. This suggests that the free-energy defence may not be able to entirely shortcut the exigencies of the case-by-case defence.

[28]For discussion of this point, see Bruineberg, Kiverstein and Rietveld (2018); Colombo and Wright (2021); Sprevak (2020).

9 The instrumental-value defence

The instrumental-value defence has a different character from the previous two. This third strategy for defending predictive coding helps to explain an otherwise puzzling phenomenon: the widespread adoption of predictive coding's computational-level claim in cognitive neuroscience despite what we have seen as the view's current relatively slender epistemic support. According to the instrumental-value defence, predictive coding should be interpreted, not as a passive claim that awaits confirmation, but as a *way of thinking* – an assumption that researchers may adopt in order to help organise data, guide experimental design and interpretation, and formulate further, more specific hypotheses for testing. Predictive coding's computational-level claim provides a novel way to systematise behavioural and neural data. It constrains the way one might group behavioural and brain responses into psychologically relevant, computationally-defined capacities, and the kinds of experimental and control conditions one might design. Furthermore, if one understands predictive coding as a package that includes proposals at Marr's algorithmic and implementation levels, it provides a rich set of heuristics to guide and inspire claims about the formal methods and neural mechanisms that underlie those computational capacities. The focus in the previous two sections was on whether predictive coding gets the computational-level description of the brain *right* or *wrong* (or whether it does better than alternatives). But one might equally well ask the prior question of how one should come up with a computational-level description *at all*. Scientific work here can potentially benefit from what predictive coding says, even if uncertainty remains about the view's ultimate epistemic standing.

It is worth stressing that individuating the mass of human behavioural and neural responses into discrete, well-defined computational capacities is hard. Cognitive neuroscientists do not have an agreed methodology to do this. Formulating a computational-level description of the brain usually requires adopting some broad theoretical orientation about the overall purpose of the brain's activity. It is not obvious where an empirically minded researcher should look to for inspiration or guidance here. Traditionally, folk psychology has provided one possible source of inspiration. Someone might, for example, start by assuming that the brain is trying to use roughly 'belief'-like states and 'desire'-like states to produce outcomes that satisfy what it represents as desired. Bringing this general framework to bear on empirical data might motivate a researcher to formulate more specific hypotheses about particular kinds of belief-like and desire-like states inside the brain, the relationships between them, the processes that transform them, and how sensory and behavioural responses update those beliefs and fulfil those desires.[29]

[29] See Machery (forthcoming), Section 1.1.

Machery (forthcoming) describes an alternative source of inspiration that might lead a researcher to a different set of more specific, testable hypotheses about the computational tasks the brain faces and its underlying computational capacities, states, and mechanisms. He argues that one feature of evolutionary psychology is that, irrespective of its other epistemic properties, it provides a potentially valuable set of discovery heuristics. Some of these speak directly to the problem of coming up with hypotheses at Marr's computational level. For example, the 'forward-looking' heuristic suggests that our computational capacities can be identified by looking at the problems that were encountered by our ancestors that regularly bore on their fitness.[30] Hypotheses about the computational capacities that our brains have today can be inferred from the problems faced by our evolutionary ancestors (Cosmides and Tooby, 1989). Hypotheses about our computational capacities arrived at in this fashion of course need to be empirically confirmed. But even in advance of securing epistemic support, it may make sense to accept a framework like evolutionary psychology (or folk psychology) *pro tem* as a discovery heuristic, in order to make the problem of task description tractable at all.

Predictive coding could potentially play a similar role for cognitive neuroscience. It suggests that neural and behavioural responses should be organised around the central notion that those responses are all attempts by the brain to minimise long-term, precision-weighted sensory prediction error. Even if the evidential basis for that claim is relatively slight, it may still function as a useful discovery heuristic to guide design of experiments, measurement, and as a means of generating more specific, testable proposals about physical responses.

For example, Fletcher and Frith (2009), inspired by predictive coding's computational-level claim, hypothesise that a range of positive symptoms of schizophrenia – including hallucinations, delusions, abnormal saliences in perception, disturbances in low-level motor functioning – should be categorised as instances of a single, unitary dysfunction in the computational ability to minimise precision-weighted sensory prediction error. They go on to propose that this dysfunction is unwritten by both a single computational mechanism and a single physical basis, again prompted by predictive coding's claims at those levels.[31] Such work suggests novel experimental designs that might attempt to dissociate the relevant factors in schizophrenia, probe how they might be quantitatively affected by manipulating sensory prediction errors, and explore analogues of schizophrenia in healthy subjects with designs that induce similar effects on sensory prediction error.[32] Corlett and Fletcher (2014) describe how predictive coding could function

[30] ibid.

[31] ibid., pp. 53–55; Corlett, Frith and Fletcher (2009).

[32] For example, see Fletcher and Frith (2009), p. 55–56.

as a discovery heuristic for clinicians to find new therapeutic interventions for patients (including pharmacological treatments). The idea that the brain aims to minimise its sensory prediction error might function as the starting point for any number of theoretical, experimental, and therapeutic developments.

In contrast to both the case-based defence and the free-energy defence, the focus here is not primarily on truth, but on predictive coding's utility. The relevant kind of utility should be understood as broader than merely a concern with achieving a narrowly instrumental outcome. Cognitive neuroscience needs to make assumptions regarding the overall purpose of brain activity in order to make any sense of activity in the brain and behaviour. Those assumptions need to come from somewhere. It is reasonable to assume that any candidate source of such assumptions should be understood to be uncertain and exploratory. Predictive coding provides one among many possible sources (and one distinct from folk psychology or evolutionary psychology). Its sheer novelty and boldness is undoubtedly an attraction. It allows us to see familiar behavioural and neural responses in a new light and group them together in different ways from previous research programmes.

It should be clear that using predictive coding in this way – as a heuristic to guide discovery rather than a claim that passively awaits confirmation – does not somehow magically confer justification on the view. Merely believing something to be true does not make it so. Justification for predictive coding only accrues if it can predict and explain better than alternative theoretical approaches.[33] The instrumental-value defence does not undercut the need to gather conventional empirical evidence to confirm predictive coding. However, it does explain why someone might be rational to accept what predictive coding says now, even in advance of such evidence being obtained.

10 Conclusion

In its boldest form, predictive coding proposes that the only computational problem that the brain faces is to minimise its long-term, prediction-weighted sensory prediction error. It is natural to wonder what would happen if one were to qualify this claim.[34] Perhaps predictive coding describes some, but not all, problems the brain faces. One might imagine a variety of ways in which the scope of its ambition at the computational level might be reigned in. At the more modest end would be the relatively trivial claim that in early vision one thing the brain does is to minimise its sensory prediction error. At the more speculative end would be the revolutionary claim that this is the only problem that the brain aims to solve. Advocates of predict-

[33] See Machery (forthcoming), Section 3.2 for a similar point regarding evolutionary psychology.
[34] See Clark (2013b), pp. 200–201.

ive coding might wish to allow for the possibility that their view will fall between these two extremes. It is worth noting however, that to the extent to which the scope of the view is reduced, its unifying power is also compromised. If predictive coding is to fulfil its original promise of offering a grand unifying theory, the research programme should aim to deliver as broad and comprehensive a theory of brain function as possible.

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.

Alink, A., C. M. Schwiedrzik, A. Kohler, W. Singer and L. Muckli (2010). "Stimulus predictability reduces responses in primary visual cortex". In: *Journal of Neuroscience* 30, pp. 2960–2966.

Allen, C. (2017). "On (not) defining cognition". In: *Synthese* 194, pp. 4233–4249.

Atick, J. J. (1992). "Could information theory provide an ecological theory of sensory processing?" In: *Network: Computation in Neural Systems* 3, pp. 213–251.

Bayne, T., D. Brainard, R. W. Byrne, L. Chittka, N. Clayton, C. Heyes, J. Mather, B. Ölveczky, M. Shadlen, T. Suddendorf and B. Webb (2019). "What is cognition?" In: *Current Biology* 29, R603–R622.

Bogacz, R. (2017). "A tutorial on the free-energy framework for modelling perception and learning". In: *Journal of Mathematical Psychology* 76.198–211.

Bruineberg, J., J. Kiverstein and E. Rietveld (2018). "The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective". In: *Synthese* 195, pp. 2417–2444.

Carandini, M., J. B. Demb, V. Mante, D. J. Tolhurst, Y. Dan, B. A. Olshausen, J. L. Gallant and N. C. Rust (2005). "Do we know what the early visual system does?" In: *Journal of Neuroscience* 25, pp. 10577–10597.

Carandini, M. and D. J. Heeger (2012). "Normalization as a canonical neural computation". In: *Nature Reviews Neuroscience* 13, pp. 51–62.

Carruthers, P. (2006). *The Architecture of the Mind*. Oxford: Oxford University Press.

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

Clark, A. (2013b). "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.

— (2017). "How to knit your own Markov blanket". In: *Philosophy and Predictive Processing*. Ed. by T. Metzinger and W. Wiese. Frankfurt am Main: MIND Group. DOI: 10.15502/9783958573031.

Colombo, M. and C. Wright (2021). "First principles in the life sciences: The free-energy principle, organicism, and mechanism". In: *Synthese* 198, S3463–S3488.

Corlett, P. R. and P. C. Fletcher (2014). "Computational psychiatry: a Rosetta Stone linking the brain to mental illness". In: *The Lancet Psychiatry* 1, pp. 399–402.

Corlett, P. R., C. D. Frith and P. C. Fletcher (2009). "From drugs to deprivation: a Bayesian framework for understanding models of psychosis". In: *Psychopharmacology* 206, pp. 515–530.

Cosmides, L. and J. Tooby (1989). "Evolutionary psychology and the generation of culture, part II: Case study: A computational theory of social exchange". In: *Ethology and Sociobiology* 10, pp. 51–97.

Feldman, H. and K. Friston (2010). "Attention, uncertainty, and free-energy". In: *Frontiers in Human Neuroscience* 4, pp. 1–23.

Fletcher, P. C. and C. D. Frith (2009). "Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia". In: *Nature Reviews Neuroscience* 10, pp. 48–58.

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

Friston, K. (2003). "Learning and inference in the brain". In: *Neural Networks* 16, pp. 1325–1352.

— (2005). "A theory of cortical responses". In: *Philosophical Transactions of the Royal Society of London, Series B* 360, pp. 815–836.

— (2009). "The free-energy principle: a rough guide to the brain?" In: *Trends in Cognitive Sciences* 13, pp. 293–301.

— (2010). "The free-energy principle: A unified brain theory?" In: *Nature Reviews Neuroscience* 11, pp. 127–138.

— (2011). "What is optimal about motor control?" In: *Neuron* 72, pp. 488–498.

Friston, K. (2013). "Life as we know it". In: *Journal of the Royal Society Interface* 10, p. 20130475.

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

Friston, K., J. Kilner and L. Harrison (2006). "A free energy principle for the brain". In: *Journal of Physiology (Paris)* 100, pp. 70–87.

Friston, K., J. Mattout and J. Kilner (2011). "Action understanding and active inference". In: *Biological Cybernetics* 104, pp. 137–160.

Friston, K. and K. E. Stephan (2007). "Free-energy and the brain". In: *Synthese* 159, pp. 417–458.

Friston, K., K. E. Stephan, P. R. Montague and R. J. Dolan (2014). "Computational psychiatry: the brain as a phantastic organ". In: *The Lancet Psychiatry* 1, pp. 148–158.

Friston, K., C. Thornton and A. Clark (2012). "Free-energy minimization and the dark-room problem". In: *Frontiers in Psychology* 3, pp. 1–7.

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

Hohwy, J. and J. Michael (2017). "Why should any body have a self?" In: *The Body and the Self, Revisited*. Ed. by F. de Vignemont and A. Alsmith. Cambridge, MA: MIT Press, pp. 363–392.

Hohwy, J., A. Roepstorff and K. Friston (2008). "Predictive coding explains binocular rivalry: An epistemological review". In: *Cognition* 108, pp. 687–701.

Hosoya, T., S. A. Baccus and M. Meister (2005). "Dynamic predictive coding by the retina". In: *Nature* 436, pp. 71–77.

Huang, G. T. (May 2008). "Is this a unified theory of the brain?" In: *New Scientist* 2658, pp. 30–33.

Jehee, J. F. M. and D. H. Ballard (2009). "Predictive feedback can account for biphasic responses in the lateral geniculate nucleus". In: *PLoS Computational Biology* 5, e1000373.

Kirchhoff, M. D. and J. Kiverstein (2021). "How to determine the boundaries of the mind: A Markov blanket proposal". In: *Synthese* 198, pp. 4791–4810.

Koelsch, S., P. Vuust and K. Friston (2019). "Predictive processes and the peculiar case of music". In: *Trends in Cognitive Sciences* 23, pp. 63–77.

Kok, P. and F. P. de Lange (2015). "Predictive coding in the sensory cortex". In: *An Introduction to Model-Based Cognitive Neuroscience*. Ed. by B. U. Forstmann and E.- J. Wagenmakers. New York, NY: Springer, pp. 221–244.

Kok, P., J. F. M. Jehee and F. P. de Lange (2012). "Less is more: Expectation sharpens representations in the primary visual cortex". In: *Neuron* 75, pp. 265–270.

Kording, K. (2007). "Decision theory: What "should" the nervous system do?" In: *Science* 318, pp. 606–610.

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.

Machery, E. (forthcoming). "Discovery and confirmation in evolutionary psychology". In: *The Oxford Handbook of Philosophy of Psychology*. Ed. by J. Prinz. Oxford University Press.

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

Miller, M. and A. Clark (2018). "Happily entangled: prediction, emotion, and the embodied mind". In: *Synthese* 195, pp. 2559–2575.

Muckli, L. (2010). "What are we missing here? Brain imaging evidence for higher cognitive functions in primary visual cortex V1". In: *International Journal of Imaging Systems and Technology* 20, pp. 131–139.

Mumford, D. (1992). "On the computational architecture of the neocortex. II The role of cortico-cortico loops". In: *Biological Cybernetics* 66, pp. 241–251.

Murray, S. O., D. Kersten, B. A. Olshausen, P. Schrater and D. L. Woods (2002). "Shape perception reduces activity in human primary visual cortex". In: *Proceedings of the National Academy of Sciences* 99, pp. 15164–15169.

Olshausen, B. A. and D. J. Field (2005). "How close are we to understanding V1". In: *Neural Computation* 17, pp. 1665–1699.

Pellicano, E. and D. Burr (2012). "When the world becomes 'too real': a Bayesian explanation of autistic perception". In: *Trends in Cognitive Sciences* 16, pp. 504–510.

Pickering, M. J. and A. Clark (2014). "Getting ahead: Forward models and their place in cognitive architecture". In: *Trends in Cognitive Sciences* 18, pp. 451–456.

Ramstead, M. J. D., M. D. Kirchhoff, A. Constant and K. Friston (2021). "Multiscale integration: beyond internalism and externalism". In: *Synthese* 198, S41–S70.

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

Rao, R. P. N. and T. J. Sejnowski (2002). "Predictive coding, cortical feedback, and spike-timing dependent plasticity". In: *Probablistic Models of the Brain: Perception and Neural Function*. Ed. by R. P. N. Rao, B. A. Olshausen and M. S. Lewicki. Cambridge, MA: MIT Press, pp. 297–315.

Schwartenbeck, P., T. FitzGerald, R. J. Dolan and K. Friston (2013). "Exploration, novelty, surprise, and free energy minimization". In: *Frontiers in Psychology* 4, pp. 1–5.

Schwarz, O. and E. P. Simoncelli (2001). "Natural signal statistics and sensory gain control". In: *Nature Neuroscience* 4, pp. 819–825.

Seth, A. K. (2013). "Interoceptive inference, emotion, and the embodied self". In: *Trends in Cognitive Sciences* 17, pp. 565–573.

Shadmehr, R. and J. W. Krakauer (2008). "A computational neuroanatomy for motor control". In: *Experimental Brain Research* 185, pp. 359–381.

Shagrir, O. (2010). "Marr on computational-level theories". In: *Philosophy of Science* 77, pp. 477–500.

Shagrir, O. and W. Bechtel (2013). "Marr's computational level and delineating phenomena". In: *Integrating Psychology and Neuroscience: Prospects and Problems*. Ed. by D. Kaplan. Oxford: Oxford University Press.

Spratling, M. W. (2010). "Predictive coding as a model of response properties in cortical area V1". In: *Journal of Neuroscience* 30, pp. 3531–3543.

— (2017). "A review of predictive coding algorithms". In: *Brain and Cognition* 112, pp. 92–97.

Sprevak, M. (2020). "Two kinds of information processing in cognition". In: *Review of Philosophy and Psychology* 11, pp. 591–611.

— (forthcoming[a]). "Predictive coding I: Introduction". In: *TBC*.

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

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

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

Srinivasan, M. V., S. B. Laughlin and A. Dubs (1982). "Predictive coding: A fresh view of inhibition in the retina". In: *Proceedings of the Royal Society, Series B* 216, pp. 427–459.

Wiese, W. (2017). "Action is enabled by systematic misrepresentation". In: *Erkenntnis* 82, pp. 1233–1252.