 mark.sprevak@ed.ac.uk

Predictive coding IV: The implementation level

Mark Sprevak
University of Edinburgh

9 October 2021

1 Introduction

A theory at Marr's implementation level focuses on the relationship between the algorithm and the nuts and bolts of a physical machine. When a physical system implements an algorithm, its physical states should correspond to the algorithm's abstract *inputs*, *outputs*, and *intermediate states*. Changes in those physical states should be governed by physical processes that mirror the corresponding changes between the abstract symbolic or numerical states described at the algorithmic level.[1] A theory at Marr's implementation level aims to spell out which entities are related by this relationship. It specifies which elements of the abstract algorithm's specification correspond to which elements of the implementing physical system. In the case of predictive coding, it should tell us which neural states and physical processes correspond to the numerical values and mathematical operations of the ANN. One might expect it to describe the neural states and processes correspond, for example, to the ANN's *hierarchical structure*, *prediction* and *error units*, *activation function*, and *learning rule*.

This article examines a number of approaches taken to implementation by predictive coding. Section 2 introduces some general features of predictive coding at the implementation level. Section 3, 4, and 5 explore a specific proposal about predictive

[1] This kind of mirroring condition is generally supposed to be a necessary condition for the physical implementation of any algorithm (Chalmers, 2012). It is not a sufficient condition however – for a brief review of why, see Sprevak (2018).

coding's implementation – a view that I will call the 'neocortical proposal'. Section 6 examines claims about implementation that go beyond the neocortical proposal, including the suggestion that some of the ANN's features are implemented in the non-neural body's morphology or in environmental features outside the head. Section 7 explores how predictive coding might appeal to differences between the physical implementation of different cognitive processes to explain apparent anomalies at the computational and algorithmic levels – cases where it seems that the brain is not minimising sensory prediction error. Section 8 examines a potential worry with an unbridled application of this strategy: that predictive coding's computational and algorithmic level claims may be 'immunised' against disconfirming empirical evidence. Section 9 provides a brief conclusion and review of predictive coding's overall research programme.

2 General features of a theory at the implementation level

2.1 Complexity, uncertainty, and a dilemma

Brains are, by any standard, extremely complicated physical systems. They offer up a vast array of physical states and dynamics at many spatiotemporal scales. Neurophysiological study appears to reveal a great deal of variation in the behaviour of individual neurons, synapses, and sub-cellular mechanisms. Inside the brain's constellation of swirling states and processes, no one knows exactly which are the ones responsible for the computation associated with cognition. It is not easy to distinguish between physical responses in the brain that are *functionally significant* for implementing a computation from background activities that can be safely ignored.

In contrast, computers like electronic PCs have a relatively simple internal physical structure. They are made up from a small number of identical basic components arranged in a uniform and repetitive manner. The physical states and processes that implement their computations are relatively easily identified and discriminated from background physical activity. We can say which physical states and processes implement features of the computation – e.g. electrical potentials at transistor junctions – and which states and processes – e.g. the colour of the insulation over wires, the sound of the ventilation fan – can be ignored. Electronic PCs are engineered to be comprehensible to us and to offer clear and obvious patterns for physical implementation. Our position with respect to the brain is different. There are a huge number of potential claims about implementation that one might defend about the brain. A vast number of neural responses could have a computational purpose. It should come as no surprise that there is uncertainty about exactly which claim predictive coding should defend at the implementation level. An advocate of predictive

coding might wish to hedge their bets – remain to some degree uncommitted – about the details of how their algorithm is implemented. Advocates of predictive coding correspondingly tend to take a rather guarded, cautious approach to their commitments about implementation, or at least more guarded and cautious than for their claims at the computational and algorithmic levels.[2]

However, they face a dilemma here. On the one hand, if they choose to avoid commitment to a specific proposal at the implementation level (perhaps due to uncertainty), then their algorithmic-level claims become hard to test. Evidence that confirms predictive coding's claims at the algorithmic level needs to include observations showing that the neural mechanisms and responses that actually govern behaviour conform to the proposed algorithm. However, unless one knows what that algorithm entails in terms of measurable physical changes in the brain – i.e. one adopts some specific implementation-level theory – this cannot be done. On the other hand, if they choose to adopt a specific proposal about predictive coding's implementation then, although their algorithmic-level claims become open to testing, those claims also become hostage to the fortunes of that claim about implementation. If that claim about implementation were to turn out to be false or inaccurate, then any confirmation or disconfirmation that accrued to the algorithmic-level proposal on its basis would be spurious. An advocate of predictive coding needs to tread a line between: (i) making sufficiently detailed assumptions about neural implementation to open their algorithmic-level claims to empirical test; and (ii) avoiding undue commitment to assumptions that may subsequently prove to be false or inaccurate.

It is not obvious how to navigate this dilemma. The most common approach adopted in the predictive coding literature is to accept some relatively broad, provisional assumptions about neural implementation and test algorithmic-level proposals on that basis. Of course, this opens up the unwholesome possibility that if an empirical test were to produce unwelcome results, one might preserve one's algorithmic-level theory and simply modify assumptions at the implementation level so as to fit the evidence. We will explore this risk in Section 8.

2.2 The neocortical proposal

Sections 3, 4, and 5 describe the most common broad proposal about how predictive coding is implemented in the brain – the *neocortical proposal*. This is based around the idea that relatively regular neurological structures in the mammalian neocortex, 'cortical microcircuits' – which have been long suspected to serve some

[2]For example, when discussing layers of the algorithm's predictive hierarchy, Clark (2016): 'I remain deliberately uncommitted to the correct neural interpretation of this essential functional notion of layers or levels.' (p. 313n4)

computational function – correspond to repeated elements in the ANN.[3] Section 3 examines how the neocortex might implement hierarchical layers of the ANN. Section 4 considers how prediction and error units might be implemented inside cortical areas. Section 5 considers how precision weighting, associated with the lateral connections between error units, might be implemented.

It is important to emphasise that the neocortical proposal is just one hypothesis about predictive coding's implementation. It is open to revision or even, in principle, replacement. The neocortical proposal is also underspecified in certain respects: key details regarding how some features of the ANN – e.g. its individual activation values, its activation function, and its learning rule – are implemented remain to be filled out. The neocortical proposal is also likely to be, at best, only a partial account of predictive coding's implementation. The ambition of predictive coding is to explain *all* aspects of cognitive function. The neocortical proposal, however, is silent about how predictive coding would operate in non-cortical areas of the brain.[4] Parts of the ANN may also be implemented in non-neural structures, such as the non-neural body or external environment. The neocortical proposal does not say anything about this. Finally, the neocortical proposal does not say how the ANN would be implemented in agents who do not possess a neocortex, such as birds.[5] Despite these qualifications however, the neocortical proposal has become the primary framework by which algorithmic-level claims about predictive coding have been empirically tested.

2.3 Pushing complexity down to the implementation level

At both Marr's computational and algorithmic levels, advocates of predictive coding stress the universal and unifying character of their model. A single *task* and a single *algorithm* are proposed to characterise all aspects of cognition. One might wonder how this fits with the undeniable diversity among the cognitive processes displayed by, and the cognitive tasks encountered by, different organisms, or by the same organism at different times or under different conditions. Cognitive processes and cognitive tasks are clearly not all exactly alike in every respect. At some point, predictive coding should somehow acknowledge this. It should explain, or at least provide room for explaining, not just the similarities, but also the differences

[3] See Bastos et al. (2012); Friston (2005); Friston (2009); Mumford (1992); Rao and Ballard (1999). See Douglas and Martin (2004); Harris and Shepard (2015) for a general review of the anatomical structure and potential computational function of cortical microcircuits.

[4] See Büchel et al. (2014); den Ouden, Kok and de Lange (2012); Kanai et al. (2015); Miller and Clark (2018) for proposals about how non-cortical brain structures might implement part of predictive coding's algorithm.

[5] For discussion of brain structures in birds that are homologous to microcircuits in the mammalian neocortex, see Calabrese and Woolley (2015).

between cognitive processes and tasks.

It is common for advocates of predictive coding to accommodate these differences by introducing complications and variations primarily at the implementation level. As previously observed, brains are not organised in an uniform fashion and are in no sense simple physical systems. Given the huge range of physical mechanisms that the brain affords, the brain may use *multiple physical methods* – possibly operating over different spatiotemporal scales or active under different conditions – to achieve the single computational effects described in the ANN. On such a view, one might expect a theory about the physical *implementation* of cognition to be relatively complex and heterogeneous, even if the *algorithm* being implemented and the *computational task* being solved are simple and unified.[6] The complexity and diversity displayed in cognition would reflect – not the brain implementing a collection of different algorithms or computing many functions – but that it uses a wide variety of physical processes to implement the same algorithm with the goal of computing the same function. In Section 7, we will see how an appeal to these differences at the level of physical implementation can help explain observations that might otherwise appear problematic for predictive coding, such as our inability to revise certain aspects of our generative model. To a first approximation, the predictive coding research programme tends to 'push down' complexity and variation between cognitive processes and tasks into complexity and variation at the level of physical implementation. One should aim for a relatively simple, austere, and unified theory at Marr's computational and algorithmic levels, but expect a relatively messy, complicated, and open-ended story at the implementation level.[7]

2.4 No simple mapping and ambiguous terms

A corollary to this is that predictive coding is not committed to an implementation-level theory that maps the elements of the ANN onto physical hardware in any simple or direct way. It is not committed to single ANN units being implemented by single *neurons*, connections by *synapses*, unit activation values by *neural firing rates*. The ANN provides a map of a numerical algorithm; it is in no straightforward sense a wiring diagram for the brain. Neurons, synapses, and neural firing rates of

[6] This idea – that a single computational function may be implemented by diverse neural mechanisms that operate at different timescales or are active in different contexts – is not new. See Koch (2004), pp. 471–477 for discussion of how the operation of *multiplication* could be implemented by at least five dissimilar biophysical processes in the brain.

[7] Clark (2013a), pp. 193–194 describes a conflict between the 'Neats' and the 'Scruffies'. He suggests that predictive coding is likely to be at best only a qualified victory for the Neats: although the model of cognition offered by predictive coding at the computational and algorithmic levels is simple and unified, what predictive coding says at the level of physical implementation is likely to be disjunctive and messy.

course are likely to play a role in the implementation of predictive coding, but these physical elements need not stand in anything like a simple one-to-one relationship to the ANN's units, connections, and activation values.[8]

An unfortunate and potentially confusing feature of the predictive coding literature is that terms are sometimes used in a way that suggests that there *is* a simple mapping. Expressions such as 'hierarchical layer', 'connection', 'feedforward pathway', 'feedback pathway', 'lateral connection' may be used to refer to *either* abstract features of the algorithm *or* physical features in the brain. A 'lateral connection' might mean an element of the ANN (a weight in $\mathbf{\Sigma}$) or a physical connection (such as a synapse) between neurons. Of course, one might propose that there is a relationship between the two: one might claim that ANN 'lateral connections' are physically implemented by neural 'lateral connections'. But it is equally possible, and as we will see more likely, to say that the relationship between the two is more indirect. In principle, a lateral connection between error units of the ANN might be implemented by any number of physical relationships in the brain, and these physical relationships may have little in common with each other than their shared computational role. We will see in Section 5 that two rather dissimilar kinds of physical response – neuromodulator release and fast gamma-band synchronisation – are proposed to be among the physical resources that implement lateral connections between ANN units.

3 Implementing layers of the network

This section describes how hierarchical layers of the ANN may be implemented in the mammalian neocortex. The neocortex is organised into between 50 to 200 anatomically distinct *cortical areas*. These areas connect to each other in a relatively selective way: neurons inside one cortical area tend onto project to neurons in only a few other cortical areas. Those cortico-cortical connections also tend to be reciprocal: if neurons in cortical area A project to cortical area B, it is likely that neurons in B will project to A. The overall pattern of synaptic connectivity between cortical areas is commonly interpreted as having a hierarchical structure (Felleman and Van Essen, 1991).[9] Cortical areas are classified as 'higher' or 'lower' in the anatomical hierarchy depending on how far they are from a sensory or motor

[8] See Bogacz (2017): 'Even if the free-energy framework does describe cortical computation, the mapping between the variables in the model and the elements of the neural circuit may not be "clean" but rather "messy", i.e. each model variable or parameter may be represented by multiple neurons or synapses.' (p. 209).

[9] Although see the worries they raise about potential irregularities in the hierarchical structure (Felleman and Van Essen, 1991, p. 31). There may be also be multiple ways to divide up the neocortex into structures that are approximately hierarchical (Hilgetag, O'Neill and Young, 1996).

boundary. This distance is measured by the minimum number of synaptic steps – how many neuron-to-neuron hops – would be needed to reach the boundary. Projections from lower cortical areas (e.g. primary visual cortex, primary motor) tend to converge on targets in higher cortical areas (e.g. secondary sensorimotor areas or association areas). These higher cortical areas send reciprocal connections back to the lower areas. The synaptic pathways that go from lower to higher cortical areas – 'ascending' the hierarchy – tend to be excitatory. The synaptic pathways that go from higher to lower cortical areas – 'descending' the hierarchy – tend to be inhibitory.[10]

If the functional response of neurons is measured (using a technique like fMRI), and cortical areas are individuated in terms of their function rather than their anatomical structure, then the locations and relationships between the resulting regions tend to align closely with those of an anatomically individuated cortical hierarchy (Glasser et al., 2016). The structural anatomical hierarchy appears to coincide with a functional processing hierarchy. Functional spiking responses in different cortical areas appear to play different roles in cognitive processing and those roles appear to be related to each other in a roughly hierarchical fashion. Spiking activity in lower cortical areas generally tends to be associated with the brain tracking fine-grained features in specific sensory modalities (e.g. patches of contrast in small parts of the visual field). Spiking activity in higher cortical areas generally tends to be associated with the brain tracking abstract and large-scale features that span multiple sensory modalities (e.g. objects, faces, hands).[11]

The neocortical proposal suggests that *cortical areas* implement the functional *layers* of the ANN. The hierarchical structure of the neocortex and projections between cortical areas implement the hierarchical structure and pattern of connections between ANN layers. Lower cortical areas (closest to the sensory or motor boundaries) implement lower layers of the ANN (closest to the input, $\mathbf{x}$). Higher cortical areas (furthest from the sensorimotor boundary) implement higher layers of the ANN (furthest from input, $\mathbf{x}$). Ascending, excitatory anatomical pathways in the neocortex implement feedforward, excitatory connections between layers of the ANN. Descending, inhibitory anatomical pathways implement feedback, inhibitory connections between layers of the ANN. Neural responses in lower cortical areas implement activity in lower layers of the ANN – both are associated with tracking more fine-grained features in the sensory input. Neural responses in higher cortical areas implement activity in higher layers of the ANN – associated with tracking

[10] For a review of this pattern of neocortical connectivity, see Hilgetag and Goulas (2020); Markov and Kennedy (2013); Mumford (1992).

[11] See Ungerleider and Haxby (1994). Ricci and Serre (2020); Serre et al. (2005) give a helpful overview of the functional hierarchy for the visual cortex along with a non-predictive-coding computational model of these responses.

more abstract and large-scale features in the sensory input

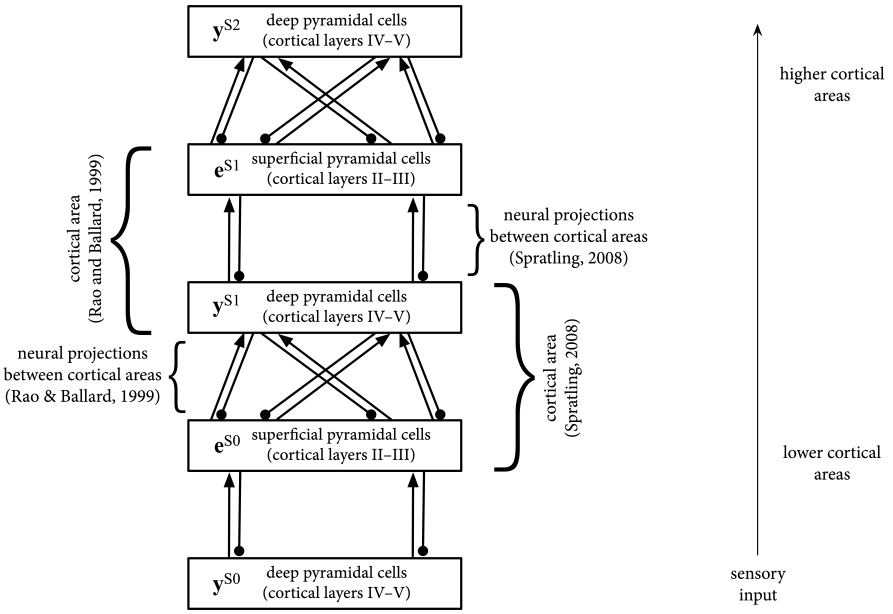

Figure 1: Neocortical proposal about the physical implementation of predictive coding.

Rao and Ballard (1999) suggest that an ANN layer consists in prediction and error units that stand in a one-to-one relation to each other (e.g. units $\mathbf{y}^{S1}$ and $\mathbf{e}^{S1}$ in Figure 1 form a layer). Prediction errors are passed 'up' the ANN between layers, whereas prediction values are passed 'down'. The weights of the connections between ANN layers encode the generative model (the $\mathbf{W}$ weights). If we combine this concept of an ANN layer with the neocortical proposal about implementation, then one would predict that neocortical areas send prediction errors 'upwards' along ascending, excitatory anatomical pathways to higher neocortical areas, and prediction values 'downwards' along descending, inhibitory anatomical pathways to lower neocortical areas. The cortico-cortical connections that link different cortical areas modulate these signals and the effective synaptic strength of those connections implements the generative model. This results in one of the most frequently cited claims associated with predictive coding: *error signals* flow forwards in the brain (from lower to higher cortical areas), and *prediction signals* flow backwards (from higher to lower cortical

areas).[12]

Spratling (2017) describes an alternative way of dividing up the ANN into layers. On his proposal, an ANN layer consists in prediction and error units that are fully connected to each other via weighted connections (e.g. $\mathbf{y}^{S1}$ and $\mathbf{e}^{S0}$ in Figure 1 form a layer).[13] Prediction values are passed 'up' the ANN between layers, whereas prediction errors are passed 'down'. Unlike with Rao and Ballard's model, the connections between ANN layers shuttle prediction values and error values around without altering them and the weighted connections inside a layer encode the generative model (the **W** weights). If one combines Spratling's concept of an ANN layer with the neocortical proposal about implementation, then one would predict that neocortical areas send prediction values 'upwards' along ascending, excitatory anatomical pathways to higher cortical areas, and predictions errors 'downwards' along descending, inhibitory anatomical pathways to lower neocortical areas. The cortico-cortical connections that link different cortical areas transmit prediction values and prediction errors around the brain and do not implement the generative model; that is implemented by connections inside the cortical areas. This results in a claim that is diametrically opposed to Rao and Ballard's: *prediction signals* flow forwards in the brain (from lower to higher cortical areas), while *error signals* flow backwards (from higher to lower cortical areas).[14]

That two contradictory predictions about brain function can be derived from the same abstract numerical algorithm should encourage some degree of caution and humility when assessing the empirical content of predictive coding. It illustrates just how tightly predictive coding's predictions about cognition and brain function are indexed to the fine print of its proposal about physical implementation. Evidence that the brain implements a predictive coding algorithm only holds conditional on assumptions about which bits of the algorithm map onto which bits of neural hardware. We will explore this issue regarding the empirical content of predictive coding in more detail in Section 8.[15] For the sake of simplicity, in the next two sections I will assume that units in the ANN are grouped into layers as Rao and Ballard suggest.

[12]For examples of this, see Bogacz (2017); Clark (2013b), pp 187–188; Friston (2005); Friston (2009).

[13]See Sprevak (forthcoming[b]), Section 4.

[14]See Kok and de Lange (2015), pp. 224–225; Spratling (2008).

[15]For a helpful discussion of this issue and a wider contextualisation of the problem in cognitive neuroscience, see Teufel and Fletcher (2016), pp. 2605–2606.

4 Implementing prediction units and error units

This section describes how physical features in the brain implement the ANN's prediction and error units. Cortical areas contain millions of neurons of many different types.[16] The structure of a cortical area is usually divided into six anatomical layers (labelled I–VI). The most common neuronal cell type inside a cortical area is the pyramidal neuron, which itself comes in many different biological subtypes.[17] Pyramidal neurons are distributed primarily in layers II–V. Smaller pyramidal neurons tend to occur in the layers closer to the outer surface of the cortex (anatomical layers II–III). In the neocortical proposal, these are called the 'superficial' pyramidal cells. Larger pyramidal neurons tend to occur in layers closer to the centre of the brain (anatomical layers IV–V). These are referred to as 'deep' pyramidal cells. Superficial pyramidal cells typically send excitatory projections forwards in the neuroanatomical cortical hierarchy to deep pyramidal cells in higher cortical areas. Deep pyramidal cells typically send inhibitory connections backwards in the hierarchy to superficial pyramidal cells in lower cortical areas. The neocortical proposal claims that *superficial pyramidal cells* implement *error units* and *deep pyramidal cells* implement *prediction units*.[18]

The neocortical proposal does not suggest that there is a simple one-to-one mapping between ANN units and pyramidal neurons – each pyramidal cell does not implement exactly one ANN unit. It is hard to see how a one-to-one mapping could be plausible. First, the input–output behaviour of a pyramidal cell does not correspond any obvious way to that of an ANN unit. Second, it is unclear how a single pyramidal cell, with a behaviour that is stochastic and sensitive to thermal noise, would be capable of reliably storing and transmitting over time a continuous numerical value – which is what is required of an ANN unit. Third, individual neurons appear to be redundant to the brain's function in a way that individual ANN units are not. Individual pyramidal neurons die or change their response profile without any apparent computational side effects, whereas ANN units are often treated as non-redundant contributors to the algorithm for inference and learning; in the probabilistic interpretation, each ANN unit represents the mean value of a unique environmental variable.[19] Advocates of the neocortical proposal typically suggest that each ANN unit is implemented by a *population* of pyramidal

[16] Classifying cortical neurons into discrete biological types can be done in many different ways based on variations in their morphology, electrophysiology, connectivity, molecular biology, and/or expression of genes and proteins (Masland, 2004; Stevens, 1998; Zeng and Sanes, 2017).

[17] See Spruston (2008). There is also within-cell-type variation for each proposed type of pyramidal cell, see Cembrowski and Spruston (2019).

[18] Bastos et al. (2012); Bogacz (2017); Friston (2005); Mumford (1992).

[19] See Sprevak (forthcoming[b]), Sections 2.5, 5

neurons.[20]

According to the neocortical proposal, error and prediction units are physically distinguished from each other by the cortical layer in which they appear (superficial versus deep). But *within* a cortical layer, how are single ANN units distinguished from each other? It is possible to imagine the brain might exploit any number of its physical or functional properties here. Cortical cells within a layer might be grouped together and distinguished from others based on their physical proximity, by their connectivity, by correlations in their firing patterns, or by their neuronal subtype. In theory, the principle that determines which cells correspond to single ANN units might vary between different cortical areas or change over time. The neocortical proposal is silent about the details here. All that is proposed is that in some respect (yet to be determined), ANN units are implemented by functionally distinct neural populations. No experimental paradigm has yet attempted to probe the neural basis of predictive coding at the resolution of single prediction and error ANN units. Indeed, the evidence for the proposed laminar separation of all prediction and error units – i.e. that prediction errors (**e**) are exclusively implemented by superficial cells and prediction values (**y**) are exclusively implemented by deep cells – remains inconclusive and controversial.[21]

Let us set aside the question of how to divide cortical pyramidal cells into populations that correspond to individual ANN units, and consider a separate question: How do those neural subpopulations, wherever they are, encode the continuous numerical values associated with individual ANN units – *viz.* the e_i or y_i values? This is also left largely open by the neocortical proposal. One possibility is that the firing rate of the neural subpopulation encodes the activation level of its corresponding ANN unit. Typically, such schemes assume there is an approximately monotonic relationship between the physical quantity and the encoded number – more rapid firing encodes a higher activation value in the corresponding ANN unit.[22] An encoded e_i or y_i value might, for example, be proportional to the average firing rate, or to the log of the average firing rate. However, the neural subpopulation might not code for these values using its firing rate, but instead rely on some other physical property, such as the timing of spikes within the population, the variability among

[20]See Clark (2013b), p. 188; Clark (2016), p. 46; Friston (2005), p. 826.

[21]Kok and de Lange (2015) observe that currently there is a 'conspicuous lack of direct evidence' for superficial and deep pyramidal cells encoding prediction errors and predictions respectively (pp. 232–233). Heilbron and Chait (2018) found 'no evidence in the auditory domain' for this claimed separation. For techniques that might uncover such a separation, see discussion of functional measurement of cortical laminae with higher temporal and spatial resolution in de Lange, Heilbron and Kok (2018), pp. 773–775.

[22]Friston assumes a rate-based neural coding scheme (neural firing rates encode the numerical values of predictions and errors) in his account of predictive coding's implementation (Kanai et al., 2015, p. 11).

individual responses of the population, or the phase of its firing relative to other patterns in the brain. Alternatively, the population might encode the numerical values using a digital coding scheme, where no continuous function would take one from the magnitude of physical responses to stored values. Digital encoding is how our electronic PCs store numerical values, and it opens the door to all manner of compression schemes and efficiencies. In general, how neural populations encode the numerical values that feature in their proposed algorithms (such as, for example, how they encode the sufficient statistics of subjective probability distributions in probabilistic inference algorithms) is largely unknown and the subject of much speculation (Pouget et al., 2013). Rasmussen and Eliasmith (2013) criticise predictive coding for lack of specificity here, arguing that a lack of detail about implementation of these numerical values risks making predictive coding's algorithmic-level proposal impossible to test.

Empirical studies often refrain from making specific or particularly detailed commitments about predictive coding's physical implementation. They tend to rely on fairly broad assumptions that would be consistent with a wide range of more specific proposals about implementation. A common assumption is that, if predictive coding is correct then neural activity in deep cortical layers should be somehow correlated with prediction occurring, and neural activity in superficial cortical layers should be correlated with prediction errors occurring. This means that if one were to apply an appropriate data-analysis technique – which might involve relatively sophisticated statistical methods, careful management and curation of the data – those predictions and errors could be recovered from that neural data. This relatively minimal assumption is compatible with many specific proposals about implementation. However, it only tells us whether *experimenters* can recover prediction or error information from the neural data (perhaps by using rather complex and roundabout methods). It does not show that the brain itself uses that particular encoding scheme for storing predictions or errors. Such studies may show that neural observations are *consistent* with brains using deep and superficial layers to encode prediction and error data. However, they do not show that predictive coding offers the best or the only interpretation of that neural data.[23]

Finally, it should be stressed that predictive coding's neocortical proposal focuses on a handful of relatively broad-brush patterns in neocortical organisation. It would be a mistake to think that these patterns exhaust the structure of the neocortex, or that the features on which it relies are perfectly regular and exceptionless. Cortical biology is extremely complicated and diverse. One might hope that at least some of this complexity and diversity can be abstracted away and ignored in a compu-

[23]See Muckli (2010). For a helpful discussion of this issue, see Kok and de Lange (2015), pp 229–231.

tational account of cognition (like the colour of insulation over wires inside an electronic PC). However, it seems reasonable to leave open the possibility that at least some of that physical complexity and diversity might have a computational role, and that predictive coding's neocortical claim would need to be elaborated to accommodate it. At this stage, exactly how one should develop the neocortical proposal to accommodate the complexities and irregularities of real-world cortical organisation is unknown.[24]

5 Implementing precision weighting

This section describes how the precision weighting of error signals might be implemented in the brain. Precision weighting allows certain prediction errors to count for more than others during the prediction-error-minimisation process. At the algorithmic level, precision weighting is modelled by weighted lateral and intrinsic connections between ANN error units.[25] These connections suppress or boost the activation levels of certain error units relative to others, meaning that they have greater or lesser influence as the algorithm unfolds. The weights of the connections (the $\boldsymbol{\Sigma}$ values) control the distribution of precision weighting over the ANN's error units.

The physical implementation of precision weighting is one of the more open-ended and less well-understood areas of predictive coding. Naively, one might assume that the physical resources that implement precision weighting would be similar to those that implement the generative model. At the algorithmic level, both correspond to the same sort of abstract feature – weighted connections between ANN units (the weights of which are specified by the matrices $\boldsymbol{\Sigma}$ and $\mathbf{W}$ respectively). In Section 3, we saw that the weighted connections specified by $\mathbf{W}$ are implemented by the strength of synaptic projections that ascend and descend between cortical areas. One might guess that the $\boldsymbol{\Sigma}$ connections would be implemented similarly, for example, by the strength of lateral synaptic connections inside cortical areas between whichever neural subpopulations implement individual ANN error units.[26]

This may be part of how precision weighting is implemented in the brain. Aspects of precision weighting that are relatively slow to change or that change during learning may be encoded in lateral synaptic projections that allow one neural population to inhibit another. But the assumption that synaptic connectivity would be the *only* way in which precision weighting is implemented would not fit with the idea that

[24] Bastos et al. (2012) explore how some, but by no means all, of the fine-grained details of cortical physiology might fit with an account of the implementation of predictive coding.

[25] Sprevak (forthcoming[b]), Section 2.4.

[26] For example, see Bogacz (2017), p. 201.

the brain's precision weighting sometimes changes dramatically and over short time periods. Predictive coding claims that changes in the agent's attention or in their degree of uncertainty about certain hidden environmental variables depends on shifts in the brain's distribution of precision weighting over its sensory prediction errors. Such changes may occur on a millisecond timescale – much faster than the kinds of change normally associated with long-term synaptic plasticity or learning (assumed to govern **W**).[27]

Friston proposes two distinct (and likely interrelated) mechanisms as candidates for processes that implement fast changes to precision weighting:

> So how is precision encoded in the brain? In predictive coding, precision modulates the amplitude of prediction errors … This means that precision corresponds to the synaptic gain of prediction error units. The most obvious candidates for controlling gain (and implicitly encoding precision) are classical neuromodulators like dopamine and acetylcholine, which provides a nice link to theories of attention and uncertainty. Another candidate is fast synchronized presynaptic input that lowers effective postsynaptic membrane time constants and increases synchronous gain. This fits comfortably with the correlation theory and speaks to recent ideas about the role of synchronous activity in mediating attentional gain. (Friston, 2010, p. 132)

Neuromodulators are brain chemicals that have the ability to systematically change the function of a neuron in their vicinity. Examples of common neuromodulators include acetylcholine, dopamine, norepinephrine, and serotonin. Acetylcholine and dopamine are known to have many effects on cortical pyramidal neurons: they can change their intrinsic firing activity, change their threshold for firing, suppress adaptation of firing, and alter the efficacy of existing synaptic connections.[28] These effects can occur rapidly – on a timescale of milliseconds – certainly much quicker than the changes associated with long-term synaptic plasticity. Many models of cognition hypothesise that neuromodulators play a role in cognition, although their true computational function is unknown.[29] Friston argues that one of the computational functions of acetylcholine and dopamine is to selectively boost or suppress firing in the neural subpopulations that implement error units, and thus to implement precision weighting of prediction error.

Friston observes that this would create a connection between predictive coding and existing theories of attention and uncertainty. These theories already suggest that

[27] Friston (2009); Clark (2016), pp. 146–150.

[28] Hasselmo (1995).

[29] For examples of various proposals, Doya (2002); Fellous and Linster (1998); Montague, Hyman and Cohen (2004).

acetylcholine and dopamine are associated with controlling attention and tracking uncertainty.[30] That connection is somewhat complicated by the fact that those theories also tend to employ rival algorithmic-level models that do not fully agree with predictive coding on the details of the computational role that neuromodulators play. For example, the literature on rewarded-guided decision making under uncertainty interprets neural activity that is modulated by dopamine as encoding *reward prediction error*. According to Friston, it encodes the *precision of sensory prediction error*. Both proposals associate dopamine with a measure of uncertainty (broadly construed), but they disagree about the details of its computational function.[31]

The second physical mechanism that Friston proposes to implement precision weighting is fast synchronised firing. Neural spikes that arrive at the same time ('fast synchronized presynaptic input') tend to have a greater effect on downstream neurons than the same inputs would if they were to have occurred in a temporally disordered way. Synchronisation appears to 'up the gain' on a neural signal.[32]. One might imagine the effect as similar to that of a group of people pushing a heavy object in an uncoordinated fashion versus timing their pushes to move it in several big heaves. Synchronised firing in the brain can start and stop suddenly and can modulate the gain on neural responses over a timescale of milliseconds. Synchronisation may occur across a variety of firing-frequency bands, and some neural populations respond more to signals that are synchronised at some frequencies than others.[33] As with neuromodulator release, the true computational function of synchronisation is unknown and the object of much speculation. Like neuromodulator release, synchronised firing is known to have profound effects on cortical neurons. It is also correlated with changes in attention: attentional shifts tend to be associated with changes in (fast) gamma-band (30–90 Hz) synchronised firing in superficial cortical neurons.[34]

Bastos et al. (2012) suggest that superficial pyramidal cells – which are claimed to implement ANN error units – are preferentially tuned to synchronisation at

[30] See Schultz, Dayan and Montague (1997); Schultz (1998) on the role of dopamine in encoding reward uncertainty; Berridge (2007) on dopamine and salience; See Herrero et al. (2008); Klinkenberg, Sambeth and Blokland (2011) on acetylcholine and attention.

[31] See Friston (2009), p. 299 for discussion of whether dopamine encodes the 'prediction error on value' – a prediction error about reward – as proposed on models of reward-guided decision-making that use a temporal-difference computational model; or, the 'value of prediction error' – how much the brain weights a sensory prediction error in its deliberations – as proposed on his predictive coding model (Feldman and Friston, 2010; Friston, Daunizeau and Kiebel, 2009; Schwartenbeck et al., 2015). Friston claims that his approach better explains the observed experimental results regarding dopaminergic activity.

[32] Salinas and Sejnowski (2001); Chawla, Lumer and Friston (1999).

[33] Engel, Fries and Singer (2001).

[34] Fries et al. (2001); Womelsdorf and Fries (2006).

gamma-band frequencies (30–90 Hz), whereas deep pyramidal cells – which implement prediction units – are tuned to synchronisation in the slower alpha and beta ranges (<30 Hz). Gamma-band synchronisation is claimed to selectively increase the responsiveness of the cortical error units (boost their neural subpopulation's response) without affecting (amplifying or dampening) the response of cortical prediction units which are tuned to signals at lower frequencies. There is empirical data to support the idea that superficial and deep cortical pyramidal cells are differentially tuned to respond to inputs synchronised at higher and lower frequencies respectively.[35] There is also evidence that 'forwards' connections in the cortical hierarchy (originating from superficial layers and carrying error signals) and 'backwards' connections (originating from deep layers and carrying predictions) tend to carry signals with higher and lower frequencies respectively.[36]

The two proposed mechanisms for implementing precision weighting – neuromodulator release and gamma-band synchronisation – are likely to interact with each other. Release of acetylcholine, for example, appears to elicit greater gamma-band oscillations.[37] The exact nature of their interaction is unknown, although one might expect that their respective effects dominate over different (albeit overlapping) timescales – the changes in cortical neuron behaviour due to neuromodulator release are generally slower to take effect and less quick to disappear than those for gamma-band synchronisation.

The neocortical claim should be understood as proposing that neuromodulator release and gamma-band synchronisation are *among* the physical resources that implement precision weighting. It does not entail that they exhaust the neural basis of precision weighting. There may be other physical mechanisms that selectively boost and inhibit the relevant neural subpopulations to implement precision weighting. Indeed, an unlimited number of physical mechanisms, operating on different timescales and interlaced in complicated ways, may jointly function as the physical basis of precision weighting in the brain. One should not assume that a simple account of the physical implementation of precision weighting will emerge from the neocortical proposal:

> Thus while the notion of sculpting patterns of effective connectivity by means of 'precision-weighted prediction error' is simple enough, the

[35] Buffalo et al. (2011).

[36] See Bosman et al. (2012). Gamma-band synchronisation is also proposed as the brain's way of solving the 'binding problem' – how representations in distant parts of the cortex get bound together into a single percept (Engel, Fries and Singer, 2001; Engel and Singer, 2001; Singer, 1999). It is not clear how the proposed 'long-range' synchronisation between distant neural populations for binding fits with predictive coding's proposal about 'short-range' synchronisation between subpopulations of error units inside a single cortical area.

[37] Buhl, Tamás and Fisahn (1998); Börgers, Epstein and Kopell (2005).

> [physical] mechanisms that implement such effects may be multiple and complex, and they may interact in important but as yet under-appreciated ways. (Clark, 2016, p. 149)

6 Beyond the neocortical proposal

One might wonder about whether the neocortical proposal is the full story about the implementation of predictive coding. Do non-cortical brain regions implement aspects of the algorithm? Do physical resources outside the brain – parts of the agent's non-neural body or technological resources in the environment – implement elements of the algorithm? It is not unusual for predictive coders to suggest that the neocortical proposal only describes one part of the implementation of predictive coding. Resources that lie outside the neocortex or outside the brain may also contribute to the algorithm.

The motivation for going beyond the neocortical proposal comes partly from within the neocortical proposal itself. In the previous section, we saw that the neocortical proposal suggests that *diverse* physical processes in the brain implement precision weighting. These processes might include neuromodulator release, gamma-band synchronisation, some combination of the two, and other mechanisms as well. While the exact mixture of physical resources that implement precision weighting is uncertain, the general idea is that a *single* formal element of the ANN need not be implemented by a *single* physical type of resource.

A similar point could be made about what the neocortical proposal says for the implementation of the generative model. The neocortical proposal claims that the generative model (the **W** matrix) is implemented by *effective synaptic connectivity* between cortical areas.[38] However, the term 'effective synaptic connectivity' does not name a single biological property. It rather denotes a functional relationship concerning how activity in one neural population tends to influence activity in another. This relationship could be physically realised in any number of specific biological changes in the molecular make-up of synaptic junctions, in the post-synaptic cell, in the pre-synaptic cell, or in the biochemical environment surrounding a synapse. The neocortical proposal is silent about how effective synaptic connectivity is achieved in the brain; it only requires that *some* physical change takes place such that firing activity in one neural population has a greater/lesser chance of causing firing in the second population. Like with precision weighting, the neocortical proposal allows for the possibility that diverse physical resources physically implement the generative model (the **W** matrix).

[38] Friston (2011b), p. 14.

Taking this idea further, one might wonder whether the physical relationships that realise, for example, effective synaptic connectivity need to be restricted to those in the immediate vicinity of the synapse. In principle, the **W** weights could be encoded by *whatever* physical features systematically change the effective synaptic connectivity between neural populations. On this reading, all manner of physical characteristics in the brain, body, and environment could qualify as 'part' of the physical implementation of the generative model. Gross anatomical features of the brain (e.g. different degrees of myelination), the spatial distances between cortical areas (such that closer areas are more likely to influence each other by spreading activation), the physical constraints on the speed of transmission of neural depolarisations, differing levels of metabolic support afforded to neural cells by non-neural cells – all of these can, in principle, change effective synaptic connectivity and thus could be claimed as part of the implementation of the generative model:

> … our basic evolved structure (gross neuroanatomy, bodily morphology, etc.) may itself be regarded as a particularly concrete set of inbuilt (embodied) biases that form part of our overall 'model' of the world (Clark, 2016, p. 175)

The same sort of reasoning applies to the implementation of the ANN connection weights associated with precision weighting (the $\boldsymbol{\Sigma}$ matrix). There is no reason why only physical processes that take place in or around the neural populations that implement error units (such as neuromodulator release, gamma synchronisation in presynaptic input) should implement precision weighting. In principle, *any* physical process that systematically changes the gain of the relevant neural subpopulations is a candidate for an implementation of precision weighting. Kanai et al. (2015) explore how *subcortical* neural activity – responses in the pulvinar nuclei in the thalamus – systematically changes the firing of populations of superficial pyramidal neurons via corticothalamic loops and hence changes the response of error units. Activity in these loops is already known to correlate with changes in attention. Clark (2016) proposes that external physical resources – mechanisms that lie entirely outside the brain – might perform a similar function. He claims that a key feature of human cognition is that it exploits non-neural bodily and environmental resources to systematically change the weighting of the brain's sensory prediction errors: external symbols and public language conjure up 'artificial contexts' that boost the weight of some sensory prediction errors (pp. 282–284),[39] our cultural practices and social institutions lend certain sensory prediction errors extra importance (pp. 275–279), and our reliance on technology such as laptops and smartphones directs our brain to correct for certain sensory prediction errors in preference to others as well as making certain aspects of the incoming sensory stream more (and occasionally

[39] See also Lupyan and Clark (2015).

less) predictable (pp. 260–262).[40]

It is worth considering that there might be a dynamic element to all this too. The precise mixture of physical resources that implement any given formal element of the algorithm (e.g. a specific **W** weight) could conceivably change over time. In principle, this might allow for more efficient utilisation of whatever physical resources – neural, bodily, and environmental – happen to be available to the cognitive system at that moment. An analogy might be drawn with 'cloud computing' paradigms on the Internet. In cloud computing, multiple physical devices are scattered around the world and activity across their various physical components implements a single distributed computation. The exact mix of physical resources inside these devices that implement the computation may change over time to suit the demands of the task and which physical resources happen to be free. Despite these variations in its physical basis, which may occur while the computation is running, the computation can proceed smoothly so long as at each moment each physical part plays the appropriate role and interacts with its fellows in the right way. In a similar fashion, our cognitive system might employ different physical resources at different times to implement formal features of the algorithm, rebalancing the mixture of physical resources across the brain, body, and environment based on current demands and availability.[41] This suggests that the full story of predictive coding's physical implementation may be extremely complex and hard to fathom. A simple Rosetta-stone-style description of predictive coding's implementation – that says that *this* formal element of the ANN is always implemented by *this* neural response – might be unrealistic. The physical implementation of predictive coding may instead be an idiosyncratic matter that varies depending on an individual's specific circumstances and available physical resources.

7 Using implementation to explain anomalies

The possibilities discussed in the previous section introduce new degrees of freedom into predictive coding's overall model of cognition. In this section, I will consider how this might allow predictive coding to accommodate behavioural or psychological phenomena that might otherwise appear puzzling or as potential counterexamples to its algorithmic or computational claims. The freedom in question concerns possible variations in the hardware that implements formally indistinguishable elements of the computation. Different physical resources that

[40] See also Clark (2017).

[41] See Clark (2016) on 'transient assemblies' of neural and environmental resources in cognition (pp. 150–151, 256–260). For more on how the physical states that implement a cognitive computation may shift depending on task demands, see the hypothesis of 'cognitive impartiality' in Clark (2007); Clark (2008), Ch. 6; and studies by Weis and Wiese (2019).

play exactly the same formal role in the computation may have markedly different physical characteristics. These physical characteristics can result in the cognitive system producing responses that depart from what one might expect if one were to assume that every physical component behaved in exactly the same way in all respects and that the only properties that matter to cognition are those that algorithmic or computational levels describe. An advocate of predictive coding may point to these *implementation-level differences* – variations in the domain of physical hardware – to explain anomalies with respect to what one might expect from the algorithmic or computational level accounts.

Two respects in which physical resources that implement the same formal element are likely to vary are how *much* they can change during the algorithm, and how *rapidly* they can change. In the idealised world of predictive coding's mathematical algorithm, formal elements like ANN connection weights were assumed to be capable of an unlimited amount of change (in principle, they can take any real-valued number), and each connection weight is assumed to change at the same rate (during the operation of the learning algorithm). In the concrete implementation of the algorithm in the brain, this may not be true. Some physical resources that implement ANN connection weights – whether they are specific parameters of the generative model or precision weightings over error units – may be harder or slower to change than others. Some might correspond to relatively fixed features of bodily morphology that are not open to revision during learning. These differences among the physical resources that implement formally indistinguishable elements may account for why a cognitive system might find it harder to change, say, certain parameters of its generative model.

One way to illustrate the point is to revisit the analogy with cloud computing. In such a computation, different physical devices distributed across the Internet may be treated as formally identical (as indistinguishable 'processing' or 'memory' units), but some may run faster than others. Some processing units (physical CPUs, GPUs) may have a faster clock speed or access to higher bandwidth channels; some physical memory units (RAM, solid-state devices, hard disks, magnetic tape) may be slower to respond or more stable over time. These implementation-level differences may be deliberately ignored at the level of the specification of the algorithm: all that matters to the algorithm is that certain operations take place in a timely enough fashion to not throw off the next step in the algorithm. Consistent with this however, may be variation in how the algorithm is physically implemented. One might need to appeal to these differences at the implementation level to explain patterns in the real-world behaviour of the system. So called 'implementational details' can have highly tangible effects. It may matter a great deal to me if, while waiting for an important message, I suffer a delay in receiving my emails, even though what is responsible for the delay is not some malfunction of the retrieval algorithm, but that

the relevant subroutine happens to be implemented on that occasion on slightly slower or less responsive hardware.

Appeal to variations at the implementation level can help to explain a range of behavioural and psychological anomalies for predictive coding. One source of such anomalies is perceptual illusion. Perceptual illusions are cases in which our cognitive system fails to minimise a sensory prediction error and continues to fail to do so apparently regardless of how well evidenced the error is (or how heavily the cognitive system attempts to weight it). An inability to make a sensory prediction error 'go away' by the usual means is what makes perceptual illusions robust, stable, reproducible phenomena. In the Müller–Lyer illusion, two straight lines of equal length are estimated to be of different lengths. This is reflected both in our conscious experience of the lines and the subpersonal estimates and responses generated by our brain.[42] No matter how many times one sees the lines, no matter how much one knows about how the illusion works, no matter how many times one might measure the lines with a ruler and verify their length, no matter how one distributes one's spatial attention over the lines, and no matter how high the stakes for the cognitive system to correct for that error, one's perceptual system still seems biased to represent them as different lengths. In the limit, one might be *morally certain* – willing to bet one's life and the lives of all one's descendants – on the proposition that the lines are the same length, yet one's perceptual system still seems to stubbornly represent them as different lengths. Additional evidence, background knowledge, shifts in attention, and so on can affect the strength of the Müller–Lyer illusion.[43] But these factors are not enough to make the illusion disappear, as they would in a normal case of misperception, or to bring about veridical perception of the lines.

What is significant here is not that a false sensory prediction occurs, but that the cognitive system seems oddly unable to correct that error. What the Müller–Lyer illusion appears to show is that certain assumptions that make up our generative model are to a certain degree *inflexible*, or at least remarkably resistant to revision. *Prima facie*, this does not fit with what predictive coding says at the algorithmic level (or with its claim about perception being a form of Bayesian inference). An appropriate sequence of weighted prediction errors *should* in principle be able to update the ANN to stop making a false sensory prediction, even if the cost might be to change the generative model in ways that would cause it to start making false predictions about other cases. There is nothing in predictive coding's algorithm to suggest that an appropriate stream of prediction errors would be incapable of changing the prediction values or revising parameters of the generative model. But we – at least, adult humans – cannot seem to make this happen. An input stream of

[42] Bruno and Franz (2009); Tudusciuc and Nieder (2010); Weidner and Fink (2007).

[43] See Qiu et al. (2008); Weidner and Fink (2007).

relevant, weighted prediction errors seems powerless to revise the model.

The stubborn nature of these sensory prediction errors only presents a puzzle, however, if one assumes the idealised, infinitely flexible model of predictive coding's algorithm. In the abstract world of the mathematical algorithm, every element of the generative model can be revised arbitrarily far in light of incoming data. If one complicates the model by noting that the ANN is implemented in finite physical resources that may have different physical characteristics and be more or less amenable to change, then the observed lack of flexibility in the generative model becomes less surprising. Indeed, given the constrained and finite nature of any physical implementation, one should expect that a real-world implementation of the generative model (and precision weighting) would not have the kind of flexibility possessed by the abstract model. Moreover if one assumes, as suggested in the previous section, that the physical implementation of predictive coding consists in a mix of different physical elements, one should expect that different aspects of the generative model to have different constraints imposed by their physical nature on how easily and how far they can be modified. Certain parameters inside the generative model may correspond to synaptic connections that are open to change (albeit to a finite degree) by learning; others may correspond to gross anatomical features that cannot be modified after development. Explanation of the persistence of errors in the Müller–Lyer illusion may thus be pushed down to the implementation level. The behavioural and psychological profile we observe – that certain sensory prediction errors associated with the illusion seem incapable of being minimised – is explained, not in terms of some characteristic of the formal algorithm, but as a consequence of the specific physical implementation of the algorithm in a complex and varied mix of physical resources.[44]

Lupyan (2015), Clark (2016, pp. 199–201), and Hohwy (2013, p. 141) also discuss the Müller–Lyer illusion, but they have a slightly different issue in mind. Their aim to explain why any false sensory prediction occurs *at all*. They suggest that the false prediction is generated by assumptions that allow the cognitive system to generate numerous true sensory predictions in the context of three-dimensional scenes (see proposals by Gregory, 1963; Howe and Purves, 2005). They argue that the prediction errors observed in the Müller–Lyer illusion case are sufficiently rare in realistic ecological settings that any failure to minimise them does not conflict with the assumption that the brain's overall goal is to minimise long-term sensory prediction error. In other words, the false prediction that the lines are different lengths should be understood as a 'short-term' or 'local' error of the kind discussed in Sprevak (forthcoming[a]), Section 5. However, even if this is correct, it would

[44]See Yon, de Lange and Press (2019) for examples of physical features in the brain that might explain what they call 'evidence-resistant' predictions.

not explain why it is *hard to change* the relevant assumption when errors do start coming in. This resistance to revision is what the algorithm fails to explain. What is proposed above is that it should be explained at the implementation level: the hardware that encodes the assumption that Lupyan et al. describe is less open to change than others. In line with what they suggest, the relevant hardware perhaps does not need to change in normal ecological settings – past evolutionary forces may have calcified the assumption by encoding it in relatively fixed physical resources.

Kirchhoff and Kiverstein (2019, pp. 88–90) argue that the persistence of the Müller–Lyer illusion should be explained in terms of the distribution of precision weighting. They claim that the prediction errors associated with the illusion are systematically assigned a low precision weighting, and so they are less likely to be revised: i.e. low precision weighting explains why the illusion persists. This might be a literally correct description of the situation, but it raises the question of why the precision weightings are set this low and why it is so difficult to change them. The weightings do not appear to change, or at least not enough to eliminate the error, in response to considerations that in other contexts would be sufficient to radically shift them: e.g. certainty that a prediction error has been made, shifts in attention, changes in reward. If the prediction error associated with the illusion is being discounted by the algorithm via precision weightings, it being discounted in a puzzlingly inflexible way. The algorithm provides no explanation for this: in principle, all lateral and intrinsic connections between ANN error units are as malleable as each other. One seems to be thrown back on the idea that there may be physical differences in how precision weighting is physically implemented in the brain that these are responsible for the observed differences in the illusion case. In other words, a return to the same basic strategy sketched in this section: explain the persistence of the illusion by appeal to differences among the physical resources that implement the formal model.

8 Constraining the empirical content of predictive coding

Introducing these extra degrees of freedom in predictive coding's model offers dangers as well as opportunities. The previous section described some of the advantages of letting the implementation of predictive coding be complex, diverse, and open-ended. However, it is easy to slip from this into treating predictive coding's physical implementation as completely unconstrained. On such a view, the implementation of a given element of the formal model (e.g. an ANN connection weight) on any given occasion could potentially be *anything* provided it fulfils the role required of it by the algorithm. No further constraints are placed on the nature of the physical resource that implements a component of the ANN. If one's assumptions about the physical implementation of predictive coding are this liberal,

then potentially any sequence of physical states that a biological system undergoes over time could be mapped to some sequence of features of the ANN. In effect, *whatever* an organism happens to do on any given occasion could be treated as the implementation of some appropriate element of the formal model, and hence as an instance of predictive coding.

Such thinking can lead one to very strange places, as Clark and Friston describe:

> We are built to breathe air through our lungs, hence we embody a kind of structural 'expectation' of staying (mostly) above water – unlike (say) an octopus. Some of our action tendencies are likewise built-in. The reflexive response to touching a hot plate is to draw away. This reflex amounts to a kind of bedrock 'expectation' of avoiding tissue damage. In this attenuated sense every embodied agent (even a bacterium) is, just as Friston (2012) claims, already a kind of surprise-minimizing model of its environment. (Clark, 2016, p. 263)

> … each organism represents a hypothesis or model that contains a different set of prior expectations about the environment it inhabits (Friston, 2011a, p. 90)

It is hard to see how these claims could be empirically tested. Whatever the system happens to do, it is treated as engaged in 'prediction' about sensory input. Any physical resource that it happens to employ on any occasion (including the resource of *having lungs*) is mapped onto it having an appropriate assumption in its generative model. No matter what physical behaviour is observed, that behaviour is treated, in that context, as an implementation of some or other aspect of the formal model. In short, the physical implementation of the ANN consists in *whatever physical resources that the system happens to deploy on any given occasion*.

Our computational claims about physical systems in science and engineering are not normally like this. When we say that our electronic PCs implement an algorithm, what we mean is that *a small number of specific electrical circuits inside the PC* implement that algorithm. We can empirically verify this claim – we can check whether the PC is running that algorithm – by examining the pattern of physical activity inside those circuits. If we discover that the pattern of physical activity in those circuits does not conform to the algorithm, then we would say that the device does not implement the algorithm; if we discover that it does conform to the algorithm, we would say that it does implement the algorithm. However, if one were to permit that *any* physical activity in and around the device could implement any aspect of the algorithm at any given moment, then one would not be able to conduct such tests. There would be no specific empirical content associated with the claim that the device implements the algorithm. Checking the electrical activity

would be little more than a pantomime, not a meaningful test of implementation. For if the electrical activity were not to conform to the algorithm, it would still be consistent with any number of other patterns of activity, perhaps highly idiosyncratic and context-dependent ones, implementing the algorithm. If one refuses to place any limitations at all on which physical resources do and do not implement the algorithm, it is hard to see how the algorithmic-level proposal could be subject to empirical constraints. Any observed sequence of physical states could, in principle, be treated as consistent with the formal model, for any sequence of physical states could be treated as implementing the relevant aspect of the prediction-generating machinery on that particular occasion.

Clark suggests that we should pull back from an unconstrained approach to predictive coding's implementation and distinguish between physical states that *properly* implement the algorithm and those that merely *could*, in some attenuated sense, be mapped onto elements of the formal model:

> If my skin heals after a cut, it would be misleading to say that in some structural, embodied fashion I 'predict' an unbroken membrane. Yet it is only in this strained sense that, to take another example, the shape of the fish could be said to embody expectations concerning the hydrodynamics of seawater. Perhaps we should allow that in some very broad sense, fish-y 'surprisal' is indeed partially determined by such morphological factors. *Our focus, however, has been on suites of entwined predictions issued by a neurally encoded generative model* – the kind of process implemented, if [predictive coding] is correct, by iterated exchanges of prediction and prediction error signaling in asymmetrical bidirectional cascades of neuronal processing. Consequently, I do not think we ought properly (without scare quotes) to speak of all these bedrock adaptive states and responses as themselves amounting to structurally sedimented (Friston says 'embodied') predictions or expectations. (Clark, 2016, p. 264–265, emphasis mine)

This raises the question of on what basis we should draw the distinction between physical resources that 'properly' implement the algorithm and those that merely implement it in scare quotes. Clark suggests that we do this in terms of resources that 'set the scene' for the prediction-error minimisation process versus those that 'more explicit[ly]' run the algorithm (ibid.). But that distinction itself seems hazy and with a questionable empirical basis. Physical resources that count as 'setting the scene' for some investigators may be classified as 'principal players' by others, and vice versa. The resources in question both fulfil the formal role required of them by the algorithm. Who is to say which are the principal players and which are not? In the case of the quotation above, why should only the neural/neocortical activity fall

into the foreground? Clark's own discussion of the role of external technology in setting precision weighting tends to blur this distinction – on his view, it is unclear whether a piece of the external environment should be understood as merely setting the scene for the assignment of 'true' neural precision weighting or whether it itself sets precision weighting in some extended implementation of the algorithm (Clark, 2016, pp. 260–262). It is simply not obvious how one should distinguish between physical resources that properly implement the algorithm from those that also play the role, but only do so to set the scene.[45]

Of course, one could settle the issue by fiat.[46] For example, one might say that the 'proper' implementation of a prediction value, y_i, consists, exclusively, in the average neural firing rate of a deep layer of pyramidal neurons. No other physical activity in the brain, body, or environment implements prediction values. Adopting this restriction opens the door to empirical testing. One would be able to look at the state of deep layer pyramidal neurons to see if they conform to the algorithm's prescriptions. But why accept a restriction like this on predictive coding's implementation? Why think that only a single type of physical state – or even a small number of physical state types – correspond to a single formal state in the model? Attempts to artificially restrict the implementation base quickly run up against the kinds of considerations raised in Sections 6 and 7 which motivated a liberal, open-ended, unconstrained approach towards predictive coding's physical implementation. In general, it is not obvious how predictive coding should reconcile two opposing forces: (i) permitting the implementation to be complex, idiosyncratic, and varied in ways that we do not yet understand; and (ii) imposing some constraints on which physical states do and do not implement the model in order to render the view empirically testable.

This brings us back to the dilemma about implementation first described in Section 2.1. On the one hand, predictive coding faces pressure to allow the cognitive system to use an unconstrained set of physical resources in and around the brain to implement its formal model. The pressure comes not only from observation of the sheer complexity of the brain and our current uncertainty about which neural processes are functionally significant to cognition, but also from the reasonable expectation that the physical implementation of predictive coding is likely to be extremely complicated and varied. On the other hand, predictive coding faces pressures to ensure that the physical implementation of its formal model is somehow constrained. Without restrictions on the physical resources that implement the model, it is impossible to test the model – to bring evidence about observed physical activity to bear either for or against the model. Without constraints, any observation

[45] Roskies and Wood (2017) draw this distinction in terms of physical elements that are more 'active' or 'passive' during the prediction process, but the nature of this distinction is again unclear.

[46] This is arguably how it is done for electronic PCs.

will be compatible with the model, as it is compatible with a suitably complicated and qualified story about physical implementation. There is currently no agreement about how to resolve this dilemma. That makes empirically testing predictive coding's model – outside the context of some artificially restricted mapping like the neocortical proposal – extremely difficult.

9 Conclusion

Pinning down what predictive coding actually says is hard. The view can take (sometimes radically) different shapes in different hands. This shifting ground is normal in cutting-edge science, and it is a sensible way for the scientific community to explore the contours of a new view. But it can be frustrating for philosophers. It is reasonable to wonder what predictive coding really is and is not committed to. In this series of papers, I have tried to sketch the bare bones of the view. That sketch is incomplete in many ways and the research programme is rapidly changing. What I have said is also likely to be contentious, at least for some advocates of the view. However, my aim has been only to get the rough shape of the view on the table, and to convey a sense of its potential attractions and challenges.

In the present paper, I have argued that bringing neural and behavioural evidence to bear on predictive coding's research programme requires making non-trivial assumptions at Marr's implementation level – about which neural (and perhaps also extra-neural) properties map to which numerical components of the algorithm. Only relative to some specific mapping can one check whether the abstract processes described actually occur in the physical world and whether they drive behaviour in the way suggested. I outlined a popular and influential theory about predictive coding's implementation – the neocortical proposal. This suggests that long-known anatomical structures in mammalian neocortex implement predictive coding's ANN. However, the neocortical proposal is a relatively broad-brush theory at Marr's implementation level – many important details regarding the ANN's implementation are omitted. It is also normally understood as offering only a partial account of the implementation of predictive coding. Moving from a broad-brush, partial theory of predictive coding's implementation to a full theory – one that would allow for uncontentious definitive confirmation or disconfirmation of predictive coding's claims – remains problematic however, and faces not just empirical challenges but also conceptual ones.

Predictive coding's research programme is an alliance of three claims at Marr's computational, algorithmic, and implementation levels. There is scope for committing to one of these claims but not others – *unbundling* the research programme. There is scope for developing the details of the claims in many different ways – *forking* the

research programme. Finally, there is scope for reigning in how much of cognition and behaviour the resulting model aims to describe and explain – *weakening* the research programme. In a sense, it is purely a semantic matter which view out of this constellation one ends up calling 'predictive coding'. What I have described here is a version that aims to connect all three claims together tightly, develop them in a way that aims to be relatively simple at the computational and algorithmic levels, and tries to cover all (or as much as possible) of cognition and behaviour. To my mind, this represents the sort of iteration of the view that best expresses the initial promise to provide a computational model of human cognition that is comprehensive, unifying, and complete. What we have seen however, that even in this ideal case what is currently in hand is more of an aspiration than a theory – one that remains to be articulated in relevant details and securely connected to standard forms of empirical evidence.

Bibliography

Bastos, A. M., W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries and K. Friston (2012). "Canonical microcircuits for predictive coding". In: *Neuron* 76, pp. 695–711.

Berridge, K. C. (2007). "The debate over dopamine's role in reward: the case for incentive salience". In: *Psychopharmacology* 191, pp. 391–431.

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

Börgers, C., S. Epstein and N. J. Kopell (2005). "Background gamma rhythmicity and attention in cortical local circuits: A computational study". In: *Proceedings of the National Academy of Sciences* 102, pp. 7002–7007.

Bosman, C. A., J.-M. Schoffelen, N. Brunet, R. Oostenveld, A. M. Bastos, T. Womelsdorf, B. Rubehn, T. Stieglitz, P. De Weerd and P. Fries (2012). "Attentional stimulus selection through selective synchronization between monkey visual areas". In: *Neuron* 75, pp. 875–888.

Bruno, N. and V. H. Franz (2009). "When is grasping affected by the Müller–Lyer illusion? A quantitative review". In: *Neuropsychologia* 47, pp. 1421–1433.

Büchel, C., S. Geuter, C. Sprenger and F. Eippert (2014). "Placebo analgesia: A predictive coding perspective". In: *Neuron* 81, pp. 1223–1239.

Buffalo, E. A., P. Fries, R. Landman, T. J. Buschman and R. Desimone (2011). "Laminar differences in gamma and alpha coherence in the ventral stream". In: *Proceedings of the National Academy of Sciences* 108, pp. 11262–11267.

Buhl, E. H., G. Tamás and A. Fisahn (1998). "Cholinergic activation and tonic excitation induce persistent gamma oscillations in mouse somatosensory cortex *in vitro*". In: *Journal of Physiology* 513, pp. 117–126.

Calabrese, A. and S. M. N. Woolley (2015). "Coding principles of the canonical cortical microcircuit in the avian brain". In: *Proceedings of the National Academy of Sciences* 112, pp. 3517–3522.

Cembrowski, M. S. and N. Spruston (2019). "Heterogeneity within classical cell types is the rule: Lessons from hippocampal pyramidal neurons". In: *Nature Reviews Neuroscience* 20, pp. 193–204.

Chalmers, D. J. (2012). "A computational foundation for the study of cognition". In: *Journal of Cognitive Science* 12, pp. 323–357.

Chawla, D., E. D. Lumer and K. Friston (1999). "The relationship between synchronization among neuronal populations and their mean activity levels". In: *Neural Computation* 11, pp. 1389–1411.

Clark, A. (2007). "Curing cognitive hiccups: A defense of the extended mind". In: *The Journal of Philosophy* 106, pp. 163–192.

— (2008). *Supersizing the Mind*. Oxford: Oxford University Press.

— (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.

— (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). "Busting out: Predictive brains, embodied minds, and the puzzle of the evidentiary veil". In: *Noûs* 51, pp. 727–753.

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

Den Ouden, H., P. Kok and F. P. de Lange (2012). "How prediction errors shape perception, attention, and motivation". In: *Frontiers in Psychology* 3, p. 548.

Douglas, R. J. and K. A. C. Martin (2004). "Neuronal circuits of the neocortex". In: *Annual Review of Neuroscience* 27.419–451.

Doya, K. (2002). "Metalearning and neuromodulation". In: *Neural Networks* 15, pp. 495–506.

Engel, A. K., P. Fries and W. Singer (2001). "Dynamic predictions: Oscillations and synchrony in top-down processing". In: *Nature Reviews Neuroscience* 2, pp. 704–716.

Engel, A. K. and W. Singer (2001). "Temporal binding and the neural correlates of sensory awareness". In: *Trends in Cognitive Sciences* 5, pp. 16–25.

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

Felleman, D. J. and D. C. Van Essen (1991). "Distributed hierarchical processing in the primate cerebral cortex". In: *Cerebral Cortex* 1, pp. 1–47.

Fellous, J.-M. and C. Linster (1998). "Computational models of neuromodulation". In: *Neural Computation* 10, pp. 771–805.

Fries, P., J. H. Reynolds, A. E. Rorie and R. Desimone (2001). "Modulation of oscillatory neuronal synchronization by selective visual attention". In: *Science* 291, pp. 1560–1563.

Friston, K. (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.

— (2011a). "Embodied Inference: or "I think therefore I am, if I am what I think"". In: *The Implications of Embodiment (Cognition and Communication)*. Ed. by W Tschacher and C Bergomi. Exeter: Imprint Academic, pp. 89–125.

— (2011b). "Functional and effective connectivity: A review". In: *Brain Connectivity* 1, pp. 13–36.

— (2012). "A free energy principle for biological systems". In: *Entropy* 14, pp. 2100–2121.

Friston, K., J. Daunizeau and S. J. Kiebel (2009). "Reinforcement learning or active inference". In: *PLoS ONE* 4, e6421.

Glasser, M. F., T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith and D. C. Van Essen (2016). "A multi-modal parcellation of human cerebral cortex". In: *Nature* 536.171–178.

Gregory, R. L. (1963). "Distortion of visual space as inappropriate constancy scaling". In: *Nature* 199, pp. 678–680.

Harris, K. D. and G. M. G. Shepard (2015). "The neocortical circuit: Themes and variations". In: *Nature Reviews Neuroscience* 18.170–181.

Hasselmo, M. E. (1995). "Neuromodulation and cortical function: Modeling the physiological basis of behavior". In: *Behavioural Brain Research* 67, pp. 1–27.

Heilbron, M. and M. Chait (2018). "Great expectations: Is there evidence for predictive coding in auditory cortex?" In: *Neuroscience* 389, pp. 54–73.

Herrero, J. L., M. J. Roberts, L. S. Delicato, M. A. Gieselmann, P. Dayan and A. Thiele (2008). "Acetylcholine contributes through muscarinic receptors to attentional modulation in V1". In: *Nature* 454, pp. 1110–1114.

Hilgetag, C. C. and A. Goulas (2020). "'Hierarchy' in the organization of brain networks". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 375, p. 20190319.

Hilgetag, C. C., M. A. O'Neill and M. P. Young (1996). "Indeterminate organization of the visual system". In: *Science* 271, pp. 776–777.

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

Howe, C. Q. and D. Purves (2005). "The Müller–Lyer illusion explained by the statistics of image–source relationships". In: *Proceedings of the National Academy of Sciences* 102.1234–1239.

Kanai, R., Y. Komura, S. Shipp and K. Friston (2015). "Cerebral hierarchies: predictive processing, precision and the pulvinar". In: *Philosophical Transactions of the Royal Society of London, Series B* 370, p. 20140169.

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

Klinkenberg, I., A. Sambeth and A. Blokland (2011). "Acetylcholine and attention". In: *Behavioural Brain Research* 221, pp. 430–442.

Koch, C. (2004). *Biophysics of Computation: Information Processing in Single Neurons*. Oxford: Oxford University Press.

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.

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.

Lupyan, G. and A. Clark (2015). "Words and the world: Predictive coding and the language-perception-cognition interface". In: *Current Directions in Psychological Science* 24, pp. 279–284.

Markov, N. T. and H. Kennedy (2013). "The importance of being hierarchical". In: *Current Opinion in Neurobiology* 23, pp. 187–194.

Masland, R. H. (2004). "Neuronal cell types". In: *Current Biology* 14, R497–R500.

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

Montague, P. R., S. E. Hyman and J. D. Cohen (2004). "Computational roles for dopamine in behavioural control". In: *Nature* 431, pp. 760–767.

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.

Pouget, A., J. M. Beck, W. J. Ma and P. E. Latham (2013). "Probabilistic brains: Knows and unknowns". In: *Nature Neuroscience* 16, pp. 1170–1178.

Qiu, J., H. Li, Q. Zhang, Q. Liu and F. Zhang (2008). "The Müller–Lyer illusion seen by the brain: An event-related brain potentials study". In: *Biological Psychology* 77, pp. 150–158.

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.

Rasmussen, D. and C. Eliasmith (2013). "God, the devil, and the details: Fleshing out the predictive processing framework". In: *Behavioral and Brain Sciences* 36, pp. 223–224.

Ricci, M. and T. Serre (2020). "Hierarchical models of the visual system". In: *Encyclopedia of Computational Neuroscience*. Ed. by D. Jaeger and R. Jung. Springer. DOI: 10.1007/978-1-4614-7320-6.

Roskies, A. L. and C. C. Wood (2017). "Catching the prediction wave in brain science". In: *Analysis* 77, pp. 848–857.

Salinas, E. and T. J. Sejnowski (2001). "Gain modulation in the central nervous system: Where behavior, neurophysiology, and computation meet". In: *The Neuroscientist* 7, pp. 430–440.

Schultz, W. (1998). "Predictive reward signal of dopamine neurons". In: *Journal of Neurophysiology* 80.1–27.

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

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

Serre, T., M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman and T. Poggio (2005). "A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex". In: *AI Memo* 2005-036. URL: https://dspace.mit.edu/handle/1721.1/36407.

Singer, W. (1999). "Neuronal synchrony: A versatile code review for the definition of relations?" In: *Neuron* 24, pp. 49–65.

Spratling, M. W. (2008). "Reconciling predictive coding and biased competition models of cortical function". In: *Frontiers in Computational Neuroscience* 2, pp. 1–8.

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

Sprevak, M. (2018). "Triviality arguments about computational implementation". In: *Routledge Handbook of the Computational Mind*. Ed. by M. Sprevak and M. Colombo. London: Routledge, pp. 175–191.

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

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

Spruston, N. (2008). "Pyramidal neurons: Dendritic structure and synaptic integration". In: *Nature Reviews Neuroscience* 9, pp. 206–221.

Stevens, C. F. (1998). "Neuronal diversity: Too many cell types for comfort?" In: *Current Biology* 8, R708–R710.

Teufel, C. and P. C. Fletcher (2016). "The promises and pitfalls of applying computational models to neurological and psychiatric disorders". In: *Brain* 139, pp. 2600–2608.

Tudusciuc, O. and A. Nieder (2010). "Comparison of length judgments and the Müller–Lyer illusion in monkeys and humans". In: *Experimental brain research* 207, pp. 221–231.

Ungerleider, L. G. and J. V. Haxby (1994). "'What' and 'where' in the human brain". In: *Current Opinion in Neurobiology* 4, pp. 157–165.

Weidner, R. and G. R. Fink (2007). "The neural mechanisms underlying the Müller–Lyer illusion and its interaction with visuospatial judgments". In: *Cerebral Cortex* 17, pp. 878–884.

Weis, P. P. and E. Wiese (2019). "Problem solvers adjust cognitive offloading based on performance goals". In: *Cognnitive Science* 43, e12802.

Womelsdorf, T. and P. Fries (2006). "Neuronal coherence during selective attentional processing and sensory–motor integration". In: *Journal of Physiology (Paris)* 100, pp. 182–193.

Yon, D., F. P. de Lange and C. Press (2019). "The predictive brain as a stubborn scientist". In: *Trends in Cognitive Sciences* 23, pp. 6–8.

Zeng, H. and J. R. Sanes (2017). "Neuronal cell-type classification: challenges, opportunities and the path forward". In: *Nature Reviews Neuroscience* 18, pp. 530–546.