Predictive processing, implicit and explicit

Paweł Gładziejewski

1. Introduction

The implicit/explicit distinction – as applied to internal representations and to the rules according to which they are processed – has played a significant role in the late 20th century, when the sciences of cognition saw a shift from symbolic to connectionist accounts of information processing underlying cognition. According to the classical symbolic approach, representations are explicitly encoded in strings of symbols and processed according to transition rules that are (at least sometimes) explicitly stored (Clapin 2002; Dennett 1983). Connectionist modeling introduced an altogether different view, where 'representations' are not stored in separable, localized vehicles but are rather implicit in the connection weights of neurons comprising the network; and the 'knowledge' of processing rules as implicitly embodied in the dispositions of the network to transform input vector of activation into output vectors (Haugeland 1991; Ramsey 2007, Ch. 5).

This chapter aims to put the implicit/explicit distinction into service once again to help make sense of the changing theoretical landscape of cognitive science. My focus will be on the predictive processing (henceforth PP) view of cognition that recently stirred some excitement among many philosophers and cognitive (neuro)scientists. According to PP, the brain is in the business of minimizing the prediction error, which measures the mismatch between internally predicted and actual sensory input. This single idea, the story goes, could explain perception, attention, and motor control, as well as potentially scale up to account for more off-line forms of cognition.

Two drastically different ways of understanding the commitments of PP have emerged in the literature. The 'intellectualist' interpretation views prediction error minimization as a process that rests on using an internal representation to infer the causes of sensory stimulation. Proponents of the 'radical' view argue that PP is best viewed as an enactive (non-representational and non-inferential) framework.

Here, I will propose a conciliatory reading of PP, where intellectualist and radical interpretations correspond to different ways in which predicting sensory states and computing prediction errors can be realized. In some cases (those in line with 'intellectualist' PP), the representational notions at use in PP pick out activities of an internal model that serves as an

explicit representation. The processes of updating and revising such a model count as genuinely inferential, although in a fairly minimal sense of *implicit* inference (that is, although the representations are explicit, the rules of processing are not). In other cases (those in line with 'radical' PP), the representational notions in PP pick out states merely implicit in processing biases attuned to statistics of natural scenes. Here, representational talk is arguably best read as a gloss on biased feed-forward processing that implicitly 'predicts' certain regularities in sensory input. Rather than inferring the causes of the sensory input, these implicit 'predictions' enable efficient *encoding* of the input. I will argue that this hybrid reading of PP introduces an interesting spin on the philosophical notion that explicit knowing-that presupposes and rests on a background of implicit knowing-how (Clapin 2002; Dennett 1983; Ryle 1949/2000).

The discussion is structured as follows. In section 2, I familiarize the reader with basic tenets of PP and summarize what intellectualist and radical readings of the framework amount to. Section 3 is devoted to distinguishing two types of mechanisms that often go under a single umbrella of PP, where mechanisms one sort employ an explicit representations, while mechanisms of the other sort rely on processing biases that implicitly embody 'predictions' about sensory regularities. In section 4, I outline the sense in which the mechanisms of the former, explicit-representation-using type perform genuine, albeit non-explicit (i.e. implicit or tacit) Bayesian causal inference. I end with a brief summary.

2. 'Intellectualist' and 'radical' predictive processing

2.1. Core tenets of predictive processing

To perceive the world, the cognitive system needs to reduce its uncertainty with respect to distal causes of its sensory states.[1] One possible way to achieve this is by employing a detector-based strategy, where pieces of internal machinery selectively react to relevant features of the environment (say, the presence of edges, flies, or human faces). However, complications associated with interpreting the sensory input in terms of its distal causes suggest a different solution. These complications have to do, first, with the fact that sensory states are inherently corrupted by random noise. And, second, they stem from the fact that the flow of sensory input depends on a whole manifold of interacting causes, and there are multiple possible ways of

[1] Readers already familiar with PP may wish to skip this part and proceed straight to section 2.2.

'unmixing' the input to yield a perception of a scene. Somehow, the cognitive system has to decide on just one interpretation.

Consider, then, a solution to the problem of overcoming sensory ambiguity where the causes of sensory states are *inferred* or *predicted*, rather than detected (c.f. Bogacz 2017; Clark 2013b, 2016; Friston 2005, 2010; Hohwy 2013; Rao, Ballard 1999). On this approach, the cognitive system produces a set of hypotheses about the way the sensory signals are generated by their worldly causes. The hypotheses regarding the causes of sensory states comprise a *generative model*. Technically, the generative model is defined in terms of a joint probability distribution P(H,D), where H stands for hypotheses about the distal causes 'hidden' behind the sensory states (e.g. edges, flies or faces) and D stands for 'observable' data, namely the sensory input itself (say, the retinal image). This joint distribution is equal to a product of the prior distribution P(H), which specifies how likely a given hypothesis is prior to obtaining sensory data, and the likelihood distribution P(D|H), which expresses the conditional probability of obtaining the data, given the hypothesis.[2] The generative model can thus serve as a basis for simulating the sensory input through drawing samples from the model: one takes a hypothesis from the prior distribution, and then, using the likelihood distribution, generates data that this hypothesis deems most likely.

To subserve *perception* (rather than free-flowing imagery), the hypotheses derived from the generative model ought to become answerable to the input itself (i.e. the samples 'drawn' from the environment that thfe model is trying to recapitulate). At this point, the notion of prediction error comes in. The model-based hypotheses are treated as predictions about what streams of sensory input are most likely to be obtained. The hypotheses can then be then adjusted by measuring the mismatch (for example mean squared error) between the sensory states predicted and the states actually sampled. This process enables the system to settle on a hypothesis that is most likely true, given the current sensory input. This is equivalent to performing perceptual *inference* by inversing the generative model to yield an approximate posterior distribution P(H|D) (later on I will come back to the reservation about the posterior being 'approximate'). In perceptual *learning*, the parameters of the generative model are adjusted to more effectively minimize long-term, average prediction error.

[2] For expository purposes, I am considering a simple, non-hierarchical model with just one level of hypotheses that directly predict the flow of input. Things become more complex in hierarchical models, where predictions that a hypothesis at a given level makes about sensory data are mediated by a set of hypotheses at intermediate levels of the generative model (see main text).

One can think of the statistical structure of the sensory input in terms of a set of nested patterns. Some of those patterns unfold rapidly (e.g. the arrangement of edges may change with every saccadic eye movement), while others are more invariant (e.g. ones revealing a stable 3D shape of an object), and some may go quite a bit beyond what can be immediately perceived (e.g. slight changes in average illumination due to changing seasons). Reflecting this nested structure, the generative model itself is thought of as hierarchical. Lower levels of the model predict the flow of rapidly changing sensory patterns, while higher levels predict more invariant, 'abstract' regularities. In this hierarchical setting, the job of predicting sensory input is parceled into a series of local subproblems, whereby each level L exclusively predicts the activity at the directly adjacent level L-1, and serves as input to be predicted at level L+1. In effect, this means that when brain uses the model at a given level to predict the activity at a level below, it is using a more abstract description of the causes of sensory input to predict causes at a less abstract level of description.

The updating of hypotheses is regulated by estimating the relative *precisions* of prior knowledge and sensory samples. Precision here is understood formally as inversely proportional to the variance of a probability distribution. In the present setting, if the sensory input is estimated to be less precise than prior expectations, then perceptual inference proceeds conservatively, i.e. it is less perturbed by the prediction error (when trying to find your way in a heavy fog, the retinal input may prove close to useless). If the opposite is the case, the inference puts relatively more weight on sensory evidence, i.e. the prediction error (e.g. when trying to make sense of a completely new situation for which one has no prior knowledge). It is precision estimation that accounts, on this story, for attention. Allocation of attention is explained in terms of precision-based weighting of prediction errors.

This theoretical toolkit can also be employed to account for motor control. PP builds on traditional predictive theories of motor control (Grush 2004; Pickering, Clark 2014) by claiming that action *just is* a way of minimizing the prediction error (Friston 2010). Roughly, movement amounts to intervening on the world to induce sensory data that conform to some prior hypothesis about one's action (e.g. 'I am now reaching for a piece of pizza') or a sequence of actions (also called 'policy', e.g. 'I am faithful to my diet'). If successful, acting results in minimizing the prediction error. Motor control thus construed relies on a kind of self-fulfilling prophecy, where to initiate movement, the brain needs to decrease the estimated precision of current proprioceptive prediction error (regardless of its actual precision) and increase the estimated precision of the prior which is to be actualized by action (and which is false at the onset of movement).

Predictive processing (PP), then, is the view that perception, attention and action all result from using a hierarchical generative model to generate predictions and minimize their precision-weighted errors.

2.2. Intellectualist PP

The preceding description of PP made heavy use of the 'intellectualist' vision of cognition as consisting of building a model to perform inferences about the world hidden beyond the veil of sensory states. How literal should we treat this description to be? Is the brain *really* constructing models to infer the world? Or is this sort of talk merely a heuristically useful (albeit perhaps in many respects misleading) gloss on a theory that, under closer scrutiny, is much more in line with embodied/enactive/embedded views of cognition?

This is where the philosophical literature on PP breaks down into two seemingly mutually inconsistent approaches, an 'intellectualist' and a 'radical' one.[3] Space forbids me from discussing in detail the many nuanced ways in which those outlooks differ. I will restrict this general summary to two dimensions along which the distinction can be drawn. These dimensions have to do with PP's commitment (1) to *internal representations*, and (2) to the idea of perception and action being underpinned by *inference*. As will transpire later on, these two axes of the debate over PP can be connected to the implicit/explicit distinction, which applies to (explicit vs implicit) representations and to the (explicit vs implicit) rules that guide information-processing, including inferential rules.

Regarding the commitment to representation, proponents of the intellectualist PP take the *generative model* to perform the function of an internal representation (Gładziejewski 2016; Kiefer, Hohwy 2018, 2019; Williams 2017). In particular, the generative model constitutes a representation grounded in exploitable structural similarity (an S-representation), akin to cartographic maps or scale models. This essentially means that: (1) the generative model encodes a hierarchical relational structure of hidden or latent variables ('hypotheses'); (2) the capacity to minimize the prediction error depends on the degree to which the model's relational structure (including its dynamics) maps onto the nested causal structure which produces sensory states. For example, the conditional dependencies between hidden variables may map onto the

[3] This is not to deny that some attempts to peacefully marry the two interpretations have been made (see Allen, Friston 2018; Clark, 2016; Constant, Clark, Friston 2019; Dołęga 2017; Korbak 2019; Wiese, Friston, Hobson 2020).

causal relations between respective entities in the world, thus contributing to the successful prediction of the sensory input.[4] Like maps, scale models and other paradigmatic cases of representations grounded in structural resemblance, the generative model is *explicit* in that its constituents (i.e. neurally implemented hidden variables; see Section 3.2) act as separable representational vehicles that bear intentional content in virtue of their placement in a larger relational structure.[5]

As for the commitment to inference, the proponents of intellectualist PP see it as the latest incarnation of a historical lineage of theories that construe perception as unconscious inference (Gregory 1980; Helmholtz 1855). The idea is that the predictive brain updates its representations in a way that conforms to the famous Bayes rule:

$$P(H|D) = P(D|H)\,P(H) / P(D)$$

Note that this equation partially corresponds to inverting a generative model (notice the numerator on the right-hand-side) to compute the posterior P(H|D). However, exact Bayesian inference requires one to calculate the right-hand-side denominator P(D). Because the value of P(D) is conditional on all the hypotheses that could potentially explain data, Bayesian inference is computationally intractable for any hypothesis space of non-trivial size. Thus, the idea is that the brain *approximates* Bayesian inference. Here is an outline of how this may work.

Think first of an exact Bayesian inference as an incremental process where, to use a catchphrase, yesterday's posteriors become today's priors. Assume that you start with some prior which is a normal (i.e. Gaussian, bell-shaped) probability distribution. Assume further that the sensory samples actually obtained are best explained by a hypothesis which is also defined as a normal distribution (this is the likelihood). Making the assumption that we are dealing with normal distributions is mathematically crucial here, as distributions of this form can be fully described in terms of their mean and variance, which simplifies calculations. The

[4] It may be noted that similarity does not have to constitute a strict isomorphism between the two structures. Partial mapping will suffice as long as it is exploitable for the system (see Gładziejewski, Miłkowski 2017). Moreover, this view of representation is largely sympathetic to embodied and action-centric theories of cognition, as it essentially treats representations as guides of adaptive action (Gładziejewski 2016; Williams 2017).

[5] Note that there is a particular sense in which structural representations may also represent *implicitly*. Namely, some relations which are not explicitly encoded in the structure of the representational vehicle may be implicit in this structure in that it is possible to *infer* them from the relations that are explicitly represented (for discussions of 'implicit' representation understood in terms of representation that is derivable from an explicit representation, see Dennett, 1983; Ramsey, 2007, this volume).

distance between the means of both distributions (i.e. the prior and the likelihood) is the *prediction error*. A posterior is formed that updates the prior in light of the prediction error. The posterior turns into a prior in the next iteration of inference, and then a new posterior is formed in light of consecutive prediction error. Over time, this process of incrementally learning and revising your priors should reduce the average prediction error.

In PP, the brain is thought to implement an algorithm that *indirectly* maximizes the posterior probabilities of hypotheses *solely* by gradually minimizing the prediction error. That is achieved by way of applying a gradient descent on the prediction error, whereby the system 'tinkers' with the model to optimize its ability to minimize the error, eventually converging on a true posterior mean. This procedure will tend to produce results that reliably correspond to what exact Bayesian inference would yield. As Hohwy puts it, the crux is "to turn around the observation that inference leads to less prediction error, and say that, if a system is somehow able to continuously minimize prediction error, then the system will approximate Bayesian inference" (Hohwy 2020a, p. 211).

2.3. Radical PP

A common thread unifying 'radical' approaches to PP consists in the denial that representations or inferences are involved in PP, or at least in proposing a substantially deflated reading of those commitments. However, beyond that, the radical camp is much more internally theoretically diverse.

There are two general ways in which the radical view is usually unpacked in the literature. One relies on interpreting the commitments of PP through the lens of Karl Friston's Free Energy Principle (FEP; for useful discussions of FEP, see Friston 2010; Hipolito 2019; Kirchoff, Parr et al. 2018). FEP is a mathematically sophisticated way of understanding self-organization in statistical terms and it applies to any system that manages to keep itself in a non-equilibrium steady state (i.e. which avoids dispersal or keeps itself separate from its environment). From this broad and abstract perspective, PP can be seen as a particular realization or 'implementation' of the FEP. Some authors have forcefully argued that the FEP-centric way of looking at things largely reconfigures how PP itself should be understood, namely as an enactive, non-representational theory (Allen, Friston 2018; Bruineberg, Kiverstein, Rietveld 2016; Ramstead, Kirchoff, Friston 2019; however, see also more representation-friendly recent work on FEP in: Ramstead, Hipolito, Friston 2020). For the present purposes, however, I want to set the FEP aside. This is partly due to limitations of space, but also because the working

assumption of this chapter that PP can stand on its own as an account of cognitive architecture, in a broad sense of being a theory of an overarching information processing strategy that underlies (much of) cognition. PP presumably applies to only *some* self-organizing systems, namely those equipped with mechanisms whose causal architecture realizes the computations postulated by the theory (Miłkowski 2013; Piccinini 2015). PP thus construed can be held in a way that is agnostic with respect to its exact connection to the FEP (although see Hohwy 2020b).

It is the second sort of arguments for the radical reading of PP that I do want to focus on here. These arguments purport to establish that the Bayesian, representation-centric rendering of the theory misconstrues the actual workings of PP-based systems.

When it comes to the representational commitment, authors that subscribe to the radical reading deny that any of the explanatory posits of PP, including the generative model, play genuinely representational roles in the cognitive system. Kirchhoff and Robertson (2018) argue that, formally, minimizing the average prediction error is equivalent to maximizing mutual information between the internal states of the system and the states of the environment (mutual information is a measure of how well the generative model fares at reducing uncertainty about the states of the world). But this simply means that the internal states come to reliably covary or carry Shannon information regarding distal causes of sensory input. Now, most philosophers agree that carrying Shannon information is not sufficient for being a content-bearing, representational state (although see Isaac 2019). The functional profile of information carriers in cognitive theories often boils down to mere causal mediation (Hutto, Myin 2013; Ramsey 2007). On this view, then, the function of the generative model is too causal-mediator-like to count as representational (see also Hutto 2018; Orlandi 2016, 2018). Furthermore, this stance may be only apparently inconsistent with the fact that scientists and philosophers who subscribe to PP routinely employ content-involving vocabulary invoking 'hypotheses' entertained by the brain. Such contentful states may be interpreted in a deflationary way, as purely instrumental or fictional (i.e. not literally true) posits that help researchers make sense of how the internal dynamics of the cognitive system relate to the external environment (Downey 2018).

As for the commitment to inference, proponents of radical PP argue that despite Bayesian appearances, the framework does not postulate inferences in the full-bloodied sense of rational transitions between contentful states. This point has been very clearly fleshed out by Nico Orlandi (2016, 2018; see also Anderson 2017). Ever since Helmholtz (1855), the inferentialist view of perception is motivated by the 'ambiguity of the senses' argument, whereby the brain supposedly needs to rely on prior knowledge to decide between numerous possible explanations

of the sensory input. But perhaps the sensory ambiguity is oversold here. Orlandi (2016, 2018) points to empirical research on natural scene statistics, which aims to discover recurring statistical patterns in images produced by natural environments (Geisler 2008). The proposal is that this research demonstrates that the sensory input is much less underdetermined by environmental causes than is often assumed. Thus, the structure of the input signal contains enough information to significantly reduce the uncertainty about its distal causes. Rather than search through an enormous hypothesis space (even using approximate shortcuts), the brain may simply rely on the sensory signal itself, assuming that the sensory/perceptual mechanisms are appropriately sensitive or attuned to the statistics of natural scenes. This approach does not have to deny that 'priors', in some sense of the word, may play a role in disambiguating the input if needed. But priors may be realized as mere processing biases (Orlandi 2016), which nudge the feedforward processing of signals towards ecologically valid interpretations (where 'interpretation' is read non-representationally, e.g. as attunement to affordances). It would be superfluous and unjustified to consider such ecologically tuned processing inferential.

3. Explicit and implicit representation in predictive processing

3.1. Implicit 'representations' in predictive processing

Let us now zoom in on the representational commitments of PP. The aim here is to argue that there are two quite different (although related) types of information-processing mechanisms that come under the common umbrella of 'PP'. One type of mechanism merely *implicitly* 'predicts' sensory patterns and, overall, fits the radical reading of PP. The other type of mechanism relies on an *explicit* internal model of the environment, and hence fits the reading of PP favored by the 'intellectualists'.

My focus will be on PP as applied to perception. In this domain, PP overlaps with a family of algorithms traditionally dubbed 'predictive coding' (see Huang, Rao 2011; Spratling 2017). At heart, predictive coding is a computational strategy in which predictable patterns in some original signal are used to *compress* this signal by exclusively encoding the prediction errors, i.e. those parts of the original signal that do not match the predictable patterns.

Consider a seminal work on predictive coding that obtains at the very earliest stages of vision (Srinivasan, Laughlin, Dubs 1982; see also Huang, Rao 2011; Spratling 2017). Retinal ganglion cells encode a signal that is sent, through the optic nerve, for further processing upstream in the visual cortex. However, this signal itself is an output of information processing

that takes place *within* the retina, starting at photoreceptors and then mediated by layers composed of bipolar, horizontal and amacrine cells. The synaptic connections between those layers are often set up in such a way that a ganglion cell will be triggered if the light intensity at the center of its receptive filed differs from the weighted sum of the intensities of 'pixels' surrounding the center. Crucially, this process is ecologically grounded in the statistics of natural scenes. Small patches of retinal images tend to be uniform in light intensity, which means that, usually, the center of the ganglion cell's receptive field matches its surround. From this perspective, the activation of a ganglion cell signifies a *prediction error* between an ecologically grounded 'prediction' and the image sampled from the environment.

Note that this canonical example of predictive coding underlying early vision is in some crucial respects different from how PP is usually portrayed in the literature (including the brief exposition I provided in section 2). The processing is purely feedforward (but also includes lateral inhibition, see Srinivasan, Laughlin, Dubs 1982) rather than bidirectional. The retina cannot be meaningfully described as estimating the external *causes* of the pattern to which it is (predictively) sensitive. The prediction errors encoded in activations of ganglion cells are *not* used to correct the internal hypotheses about the causes of input. What is at stake here is more accurately described in terms of the retina producing a *sparse encoding* of the raw input image by subtracting spatial (and temporal, see Srinivasan, Laughlin, Dubs 1982) redundancies (see also Barlow 1961).[6] The resulting message is encoded in fewer bits and can be effectively 'squeezed' through the informational bottleneck of the optic nerve.

Even in this simple case, representational notions may prove useful in making sense of what the retina is doing. One may say that the retinal processing 'predicts' that input images are locally uniform in terms of light intensity, or that the ganglion cells encode errors with respect to a 'hypothesis' or 'prediction' that the center of a receptive field matches the surround. What should we make of such intentional ascriptions? I propose that the best way to read them is by appeal to a notion of 'implicit' representation (Clapin 2002; Dennett 1983; Haugeland 1991; Ramsey 2007, Ch. 5; Ramsey, this volume). Although this notion is ambiguous and has historically meant different things for different authors, for the present purposes I interpret it in a way that is nicely illustrated by the famous Dennett's classic example of a chess-playing

[6] This is not to suggest that there is a dichotomy between producing a sparse encoding of a signal and estimating or modeling its causes. Explicit generative models of causal structures that generate input (to be discussed in Section 3.2) also encode compressed information about sensory data. The point here is that in the case of the retina, creating a sparse code of the input signal is the sole computational goal which is achieved without modeling the causes of the signal.

computer program to which one attributes the desire to take the queen early (see also Ramsey, this volume). Although there is no localized, separable, causally active internal vehicle that bears this content, the overall dispositional pattern of the program's behavior embodies the desire *implicitly*. To generalize, the hallmark of implicit representations in the present sense is the lack internal structures where separable items act as vehicles that encode intentional contents. Instead, on this notion, (implicit-)representational ascriptions are grounded in the fact that a given information-processing system is wired in a way that allows it to embody dispositions to behave or respond 'as if' it (explicitly) represented the world to be certain way.

In the context of PP in particular, the notion of implicit representation would cover cases where the cognitive machinery is set up so as to allow the organism to be attuned to or appropriately responsive to relevant statistical regularities, without explicitly representing those regularities or the causal structures that subserve them (see Orlandi 2016). So, to get back to the retina example, the 'knowledge' or 'predictions' one may want to attribute to the retina are implicitly embodied in the capacity or disposition to turn raw light intensity distributions into sparse messages. This disposition is tuned to natural statistics of retinal images, and hence the retina implicitly embodies 'knowledge' or 'prediction' of those statistics. Of course, the retina has this disposition in virtue of its feedforward synaptic wiring. But it would be misleading to say that this wiring is explicitly encoding a model or a representation of how the signals are generated.[7]

A worry may be raised that the retina is a special case and that the example hardly generalizable beyond the very periphery of perceptual processing.[8] However, it seems that the case is generalizable to a non-trivial degree, and that it is just one illustration of a larger recurring strategy where, due to the hard-wiring of the nervous system, purely feedforward processing implicitly 'predicts' the natural statistics of the input (for an extensive review, see Teufel, Fletcher 2020). Other examples may include the findings that in the primary visual cortex, neurons tuned to the cardinal orientations outnumber those tuned to oblique orientations, echoing natural orientation statistics (Li, Peterson, Freeman 2003); or that neurons in the visual cortex tuned to particular object categories tend to have their receptive fields predictively biased

[7] This claim is consistent with the idea that there is also an entirely *non-semantic* meaning of 'prediction' (as a matching between two non-semantic, physical signals) on which the activities of neighboring 'surround' bipolar cells in the retina predict the activity of 'center' cells, and the activity of the output ganglion cell is determined by the degree to which this prediction is 'accurate' (see also Anderson, Chemero 2013; Cao 2020).

[8] I am indebted to Alex Kiefer for raising this point.

towards locations usually occupied by the objects belonging to those categories (Kaiser, Quek et al. 2019). Akins (1996) provides an intriguing case of more imperative (desire-like) 'predictions' of this kind, whereby the variable distribution of thermoreceptors on different parts of the body (e.g. the scalp vs the hands) implicitly embodies preferences regarding how cold/hot the organism can allow those parts to become.

I submit that the understanding of representational states as merely implicit in feedforward processing biases fits the radical view of PP much better than the intellectualist one. The line between implicit representationalism and non-representationalism is thin. We may treat intentional talk in such contexts as aiming to highlight how the internal processing relates to the statistics of natural scenes. For example, the center-surround antagonism of the retinal ganglion cells makes sense once you consider what those cells are doing in the context of natural image statistics. However, this sort of connection to the environment is arguably not *literally* representational (Downey 2018), and may be more parsimoniously cashed out in terms of ecological embeddedness of perceptual/sensory mechanisms (Orlandi 2016, 2018). Relatedly, it has been forcefully argued that the notion of implicit 'representation' does not even pick out representations in an explanatorily valuable sense of the term (Ramsey 2007, Ch. 5; see also Ramsey, this volume), which, in line with the radical view, invites a purely non-representational reading.

3.2. Predictive processing and explicit representations

The way PP is usually understood in the literature today originates from a significant extension of vanilla predictive coding, put forward by Rao and Ballard (1999), and subsequently refined by Friston (2005; see also a precursor to this approach in Dayan, Hinton et al. 1995). I want to argue that this development constitutes a shift from implicit PP to a version that relies on an *explicit* model of the environment.

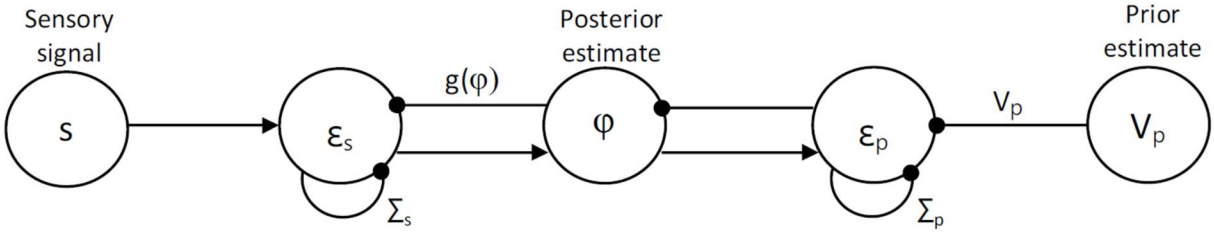

Figure 1. The bidirectional predictive coding scheme, adapted from Bogacz (2017). Arrows signify excitatory connections, and the lines ended with circles signify inhibitory connections. See main text for details.

Consider an extremely simple, informal illustration of this extended predictive coding, which I adopt from Bogacz (2017), who in turn draws on the model proposed by Friston (2005). Imagine a simple organism trying to perceptually estimate the value of some hidden environmental variable, *V*. Assume that *V* is the size of an item of food, which the organism needs to estimate while only having direct access to a sensory signal, *s* (corresponding to, say, a state of a light receptor). This signal depends causally on *V*, such that *s* is a function of *V*, g(*V*). Suppose further that the organism is equipped with a prior expectation V_p about what the most likely size of a food item is before obtaining any input. Given some technical details which I will leave out here, the job faced by the organism is then equivalent to computing P(*V*|*s*), that is, a posterior distribution over *V*, given *s*.

Importantly, in this model, instead of establishing the whole posterior *distribution*, the organism attempts to estimate a determinate value φ, which encodes the most likely value of *V*, given the current sensory signal, *s* (see Fig. 1). The idea is that the system may start with a random guess, and then engage in a recursive trial-and-error procedure (gradient descent) to gradually bring φ closer to the mean of the posterior distribution. This process relies on minimizing the prediction error. The estimate φ generates an inhibitory prediction signal that is a function of φ, g(φ). This prediction is propagated downwards and compared to *s*, giving rise to a prediction error ε_s, which measures the difference between *s* and the prediction. The error is weighted according to the variance Σs of the sensory signal, and propagated up the hierarchy to allow revision of φ. Simultaneously, the updating of φ is driven by a *prior* estimate V_p. Again, this is realized by minimizing the prediction error ε_p, i.e. the difference between φ and V_p, weighted by the variance of the prior, Σp (see Fig. 1). The upshot is that φ is constrained both by the incoming sensory signals and the prior, weighted by their respective reliabilities.

At the implementational level, the variables *s*, φ, ε_s and ε_p correspond to occurently active nodes and may be implemented in the activity of individual neurons or neural populations. The prior V_p acts as a model parameter and is encoded in the strength of a synaptic feedback connection between a 'tonically active' neuron and the node encoding φ (Bogacz 2017; see also Friston 2005). The values of Σp, Σs similarly act as parameters and are implemented in self-inhibitory synaptic connections of the error nodes.

Now, what this sort of scheme and the retinal predictive coding have in common is that they both rely on the idea that predictions (in a broad sense) are used to compute prediction errors

to efficiently encode data.[9] However, there are crucial differences as well. The model just discussed involves a bidirectional information flow that includes feedback projections. Under this regime, predictions correspond to descending inhibitory signals. The ascending prediction error signals are not the end-point of processing, but allow the correction of internal estimates.

Most importantly for the present purposes, on the scheme discussed above, the descending predictions are based on the estimates of the *cause(s)* of the input signal. The prior assumptions about the causal structure generating the sensory input are encoded in V_p.[10] The 'best guess' about the current state of the environment is encoded in φ. The relation between φ and the prediction signal g(φ) encodes the functional relations between the worldly causes and their sensory effects. Because prior assumptions about the most likely causes of the sensory input and about how these causes generate the input correspond to specifiable components of the neural architecture, we may say that the *generative model* is explicitly encoded. Furthermore, the idea that this model constitutes an explicit *representation* gets traction once we recognize that, functionally, it fits the 'job description' of a representation (Ramsey 2007). Like many things we pretheoretically regard as representations, it relies on structural similarity, allows learning through error correction, guides adaptive action (in realistic models which encompass both perception and action) and can potentially underpin more off-line types of cognition (see Gładziejewski 2016 for an extensive defense of this view). [11]

As should be clear from the preceding discussion, the idea that PP postulates an explicit internal representation of the environment is very much in line with the intellectualist reading of PP. Crucially, however, the point here is *not* to argue whether the cognitive system engages in PP by implicitly 'predicting' natural scene statistics (rehabilitating the radical view of PP) or

[9] In cases like the one just discussed, the internal model of how the sensory data are generated (see main text) constitutes an informationally parsimonious encoding of the data (Huang, Rao 2011).

[10] The example assumes that a single variable is sufficient to describe the environment. In more realistic scenarios, prior knowledge would encompass multiple interrelated (also hierarchically) hidden variables mirroring, to some biologically affordable degree, the complex causal dependencies in the environment. In those more sophisticated cases the idea that (explicit) PP involves *structural* representations becomes clearer.

[11] Note that in our example, φ will covary with or carry Shannon information about the sizes of food items. It would be manifestly wrong, however, to draw from this a conclusion that the whole set-up boils down to using detectors or causal mediators. In explicit PP, the covariance is established through predictive, error-corrected *modeling* of the environment. More generally, what is decisive in figuring out whether representations are involved in a given story is the mechanics of *how* the informational relationship is established, e.g. through modeling vs detector-like causal dependence (see Kiefer, Hohwy 2017).

by storing an explicit model of the environment (rehabilitating intellectualism). Instead, it is perfectly sound to think that the cognitive system could rely on *both* strategies. Not only could these strategies peacefully co-exist, but they may functionally complement each other in interesting ways.

In fact, some empirical results and theoretical considerations give leverage to the idea that intellectualism and radicalism about PP can be married along those lines. In visual perception, a rapid (150ms) feedforward phase of processing is followed by a second phase, carried out using feedback projections (Kreiman, Serre 2020). From the PP perspective, the initial feedforward sweep could correspond to 'implicit' predictions, and the onset of the 'feedback' phase would demarcate the point at which explicit generative model(s) kicks in. This would also suggest a functional division of labor, as we know that while the initial phase enables quick object recognition, the feedback phase is required to interpret the meaning of complex visual scenes by flexibly relating them to background knowledge (see Kreiman, Serre 2020 for an extensive discussion).[12] Relatedly, it has been argued that the 'predictions' implicit in feed-forward processing track context-invariant regularities in the sensory input, and the job of mechanisms corresponding to (what I construe as) explicit PP is to flexibly track context-dependent sensory patterns (Teufel, Fletcher 2020; see also the distinction between 'structural' and 'contextual' priors in Seriès, Seitz 2013).[13]

Philosophically, this discussion sheds new light on the old idea – due to Ryle (1949/2000), and raised in the cognitive-scientific context by Dennett (1983) – that explicit knowledge-that rests on a background of abilities implicitly embodied in practical 'know-how'. The conciliatory view of PP vindicates this overall approach in the following way. Orlandi (2016) notes that a challenge to intellectualist PP lies in the sheer vastness of the space of all the prior hypotheses that the cognitive system could bring to bear when interpreting the sensory input. How does the system constrain this space to only consider a subset of (relevant) hypotheses? The present view suggests that much of the computational or epistemic load is distributed to implicit machinery. This way, the explicit machinery is freed from the need to learn and

[12] It needs to be stressed that, for now, this is a speculative interpretation of empirical results that requires further research. It may be the case that the distinction between the feedforward sweep and the feedback phase does not map nicely onto the distinction between implicit predictions/representations (embodied in feedforward processing biases) and the explicit generative model. I am grateful to Alex Kiefer for pointing this out.

[13] Note, however, that the invariant-and-inflexible vs contextual-and-flexible division may be oversimplified, as there are demonstrations of flexible adaptation to changing input statistics taking place even at the earliest, retinal stage of visual processing (Hosoya, Baccus, Meister 2005).

represent certain prior assumptions (that small patches of the retinal input tend to be uniform; that vertical and horizontal edges dominate visual scenes; that clouds are usually found in the upper side of the visual field; that it is more dangerous to expose one's head to cold than it is to expose one's hands, etc.). Instead, much of what is 'known' or 'predicted' about the environment is implicit in feedforward information-processing mechanisms. Speculatively, perhaps the abstract priors regarding space, time, or objecthood that structure perception are (at least in part) implicitly embodied in this way. If true, this would dovetail with the Kantian reading of the implicit/explicit distinction in cognitive science (Clapin 2002).

4. Implicit Bayesian inference in predictive processing

Are PP-based cognitive systems inferential? What would it mean to say that they are? Here again I want to recommend a conciliatory approach, on which predictive cognitive systems are partially inferential. In particular, they are inferential to the degree to which they use explicit generative models. Intellectualists about PP get this part of the story right. However, the part of the story that encompasses computational strategies where feedforward processing implicitly embodies predictions is non-inferential, in line with claims made by the 'radicals'.

To put this discussion in context, the questions surrounding the status of Bayesian inference in cognitive science reach beyond PP. PP is just one incarnation of a larger trend of employing Bayesian probability theory to model cognitive phenomena (c.f. Griffiths, Kemp, Tenenbaum 2008). According to one influential criticism, Bayesian cognitive modeling confounds descriptive use of Bayesian mathematics with an explanatory use (Bowers, Davis 2012; Colombo, Seriès 2012; Jones, Love 2011). Bayesian models typically aim to establish that human performance on some cognitive task overlaps with an optimal Bayesian solution. However, the mere fact that subject's response patterns can be captured using Bayesian formalism does not any way guarantee that the *internal causal mechanisms* responsible for those patterns perform or implement Bayesian inference.

We may situate this against an even wider philosophical background, as the issue at hand relates to a distinction between behavior that *conforms* to a rule and behavior that is *guided* by a rule (Davies 2015). Quine (1970) once used this distinction to argue against Chomskyan linguistics. He pointed out that the fact that linguistic behavior conforms to a set of syntactic rules (to which a subject has no conscious access) does not imply that these rules genuinely guide this behavior. In fact, one might posit a completely different set of unconscious rules which correspond to the very same set of behaviors, and, as long as we rely on behavioral data

alone (which was Quine's assumption), there will be no principled way to decide which rules guide the behavior. The aforementioned criticism of modern Bayesian models in cognitive science boils down to raising a similar point: the fact that task performance conforms to the Bayes rule does not imply that it is guided by it.

A natural way out of this problem for anyone subscribing to a Bayesian-inferential view of cognition (including those who buy into intellectualist PP) would consist in trying to spell out conditions under which a given system is *guided* by the Bayes rule. Arguably, these conditions should have to do with whether the Bayesian inferences postulated by a cognitive model map onto or correspond to causal transactions within a cognitive mechanism responsible for a given explanandum phenomenon (Miłkowski 2013; Piccinini 2015). This, however, is easier said than done, as one needs to steer the account so that it avoids specifying, on the one hand, conditions that are too restrictive or, on the other hand, excessively liberal ones. An exemplary approach belonging to the former category would require the Bayes rule to be *explicitly* (even if without conscious awareness of the subject) encoded or stored in the mechanism, and have each particular instance of inference 'consult' and the follow this rule. Because exact Bayesian inference is intractable, such an account would be utterly naïve.

An example of the opposite, too-relaxed sort of approach would consist in simply *equating* predictive coding, in all its forms (see Spratling 2017), with Bayesian inference. A position of this kind would treat even the predictive coding scheme employed by the retina as an instance of Bayesian inference. This would be gratuitous to say the least, as retinal processing hardly aims to estimate the most likely hypotheses about the causes of sensory stimulation. Its function, to repeat, is to produce a sparse encoding of the raw image registered by photoreceptors. In fact, predictive coding is best seen as an 'algorithmic motif', which may serve several different computational goals, Bayesian inference being just one (Aitchison, Lengyel 2017). To generalize this point, whenever the cognitive system 'predicts' states of affairs simply by virtue of the implicit attunement of feedforward processing to natural scene statistics, the usage of hefty cognitivist notions like 'inference' will be at least deeply problematic. Proponents of radical PP are right in this respect.

How about the cases like the one discussed in section 3.2, where the prediction error minimization relies on an explicit model of the environment? This is where an argument can be made that, in line with intellectualist PP, genuine (approximate) Bayesian inferences sometimes guide the cognitive system. Of course, an explicit inscription of the Bayes rule is nowhere to be found in Figure 1. However, the proposal here is that although the *representations* at play are explicitly encoded, the *rule* of processing (i.e. the Bayes rule) is itself implicit in the causal

transitions between the representational states. In other words, although the updating of representational states merely conforms to Bayes rule (without explicitly following it), the states undergoing the updating are states of components of an internal causal mechanism, and so there is a clear sense in which the rule guides perception, rather than merely describes it. That causal transitions conform to Bayes rule should already be clear from the preceding discussion in sections 2.2 and 3.2. Two considerations are crucial. First, the computational goal of a mechanism like the one present in Figure 1 corresponds to the computational objective of Bayesian (perceptual) inference, i.e. it consists in estimating the most likely causes of the sensory signal. Second, as long as the output posterior estimate (φ) relies on prediction error minimization, with the prior estimate (V_p) and the current sensory samples (s) serving as inputs, the result will tend to approximate the true posterior distribution that exact Bayesian inference would yield.[14]

Before I conclude, I want to briefly sketch out another, perhaps more revealing sense in which implicit inference is at play in PP. What I have in mind here is inspired by a solution proposed by Evans (1981) to the aforementioned problem raised by Quine. Evans put forward an account of *non-behavioral* evidence that could determine whether the behavior of a given system merely conforms to a rule or is guided by it. The account relies on the notion of 'causal systematicity'. Put simply, according to this view, the rule is guiding the system as long the behavioral manifestations of said rule have a common causal origin in the system.

Take an illustrative example (Davies 1995, 2015). Suppose that a person's behavior conforms to a rule according to which if a word starts with 'b', then its pronunciation starts with the sound /B/. We intend to establish whether the person's behavior is also guided by this rule. To achieve this, the proposal goes, we need to verify whether particular behaviors that conform to this rule are causally mediated by a single state of the subject. That is, under the hypothesis that the /b/-B rule guides the system, the particular instantiations of behavior conforming to this rule count as manifestations of a complex disposition underpinned by a single internal causal state. If, however, we found out that that the person's brain stores a long look-up table such that each instantiation of the rule-conforming behavior is generated by a different causal state, the claim that the person is guided by the 'b'-/B/ rule would be falsified.

[14] There are other relevant considerations that I leave out here for the sake of space. These have to do with the fact that representational transitions in mechanisms of this kind should naturally exhibit truth-preservation and tend to maximize coherence among representations, which are traditional hallmarks of genuinely inferential processes (see Kiefer 2017 for an extensive discussion).

Note now that in (explicitly representational) PP, the error-minimizing brain gradually uncovers the hidden causal structure that produces the sensory input (Clark 2013b; Friston 2005; Kiefer, Hohwy 2017; Hohwy 2013). Over time, through learning (or iterated perceptual inference), the raw stream of sensory stimulation is 'unmixed' into a generative model comprising latent variables that capture the complex, non-linear interactions between the worldly causes of the sensory input. The structure of the model can be expressed as a Bayesian network (see Danks 2014 for an extensive discussion of cognitive representations construed as Bayesian networks). That is, the model can be regarded as a network comprising (1) nodes that encode the values of latent variables[15], and (2) edges which encode dependencies between those variables. On this view, each single node – simply in virtue of its position within a larger relational structure – systematically encodes a whole range of inferential connections[16] between itself and other nodes (see also Clark 2013a; Kiefer 2017, 2019; Williams 2018). In addition, the network is productive in that, in principle, the system could endogenously tinker with the values of certain variables to run counterfactual simulations that do not have to correspond to any previous perception (see also Clark 2013a; Kiefer, Hohwy 2017; Williams 2020).

What this amounts to is that whenever the cognitive system brings a given latent variable to bear on a cognitive process – be it on-line in perceptual inference or in counterfactual imagery – this process will be mediated by a single node in the network. This cognitive story becomes causal once we allow the generative model to be explicitly encoded in the system so that the nodes correspond to internal causal states or components of a causal mechanism (although the exact implementational details may turn out quite messy). Hence, we may expect the inferential transitions attributable to the system to map onto causal transitions in the system (see also Kiefer 2017, 2019). Seen this way, the predictive cognitive system – to the extent that it encodes an explicit generative model of the environment – embodies causal systematicity that allows it to count as inferentially guided rather than merely inferentially describable.

5. Conclusions

[15] Importantly, the contents encoded by the nodes do not have to correspond to personal-level conceptual or propositional contents. Because the nodes in Bayesian networks are 'atomic' and lack a constituent syntactic structure, doubt has been cast over whether these networks (and PP in general) can capture the 'rich' systematicity of human thought (Williams 2018; see Kiefer 2019 for an attempt to answer this challenge).

[16] Note, however, that the claim is not that the system performs Bayesian inference simply in virtue of the fact that it encodes a Bayesian network (Bayesian networks do not have to undergo Bayesian updating, see Danks 2014).

In this chapter, I set out to show that the implicit/explicit distinction in cognitive science – in particular, as regarding the nature of representation and the rules of information processing – can illuminate our understanding of predictive processing. I proposed that depending on the details, PP may come in two flavors: (1) explicitly representational and implicitly inferential, or (2) implicitly representational and non-inferential. This distinction at least partially maps onto the 'intellectualist' and 'radical' readings of PP, respectively. The upshot is that 'intellectuals' and 'radicals' may both get *part* of the story right, in that there is some theoretical and empirical traction to the claim that the brain uses both aforementioned strategies to deal with sensory uncertainty. This position also restores the classical idea that explicit cognition is built on top of implicit know-how. In the present, PP-based guise, this means the computational strain on explicit, Bayesian mechanisms is reduced by the fact that those mechanisms process signals already 'filtered' by expectations implicitly embodied by feedforward processing biases.

Acknowledgments

I thank Alex Kiefer for his insightful critical remarks on the previous version of this chapter. My work on this paper was kindly supported by the National Science Center in Poland (grant no. 2019/33/B/HS1/00677).

References

Allen, M., Friston, K. J. (2018). From cognitivism to autopoiesis: Towards a computational framework for the embodied mind. *Synthese*, 195, 2459–2482.

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

Akins, K. (1996). Of sensory systems and the "aboutness" of mental states. *The Journal of Philosophy*, 93(7), 337–372.

Anderson, M. L. (2017). Of Bayes and bullets. In T. Metzinger, W. Wiese. *Philosophy and Predictive Processing*. MIND Group. Available online at: https://predictive-mind.net/papers/of-bayes-and-bullets.

Anderson, M.L., Chemero, T. (2013). The problem with brain GUTs: Conflation of different senses of "prediction" threatens metaphysical disaster. *Behavioral and Brain Sciences*, 36, 204–205.

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

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, 76, Part B, 198–211.

Bowers, J. S., Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, 138(3), 389–414.

Bruineberg, J., Kiverstein, J., Rietveld, E. (2016). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, 195, 2417–2444.

Cao, R. (2020). New labels for old ideas: Predictive processing and the interpretation of neural signals. *Review of Philosophy and Psychology*, 11, 517–546.

Clapin, H. (2002). Tacit representation in functional architecture. In: H. Clapin (ed.). *The Philosophy of Mental Representation* (pp. 295–312). Oxford: Oxford University Press.

Clark, A. (2013a). Expecting the world: Perception, prediction and the origins of human knowledge. *The Journal of Philosophy*, 110(9), 469–496.

Clark, A. (2013b). Whatever next? Predictive brains, situated agents and the future of cognitive science. *Behavioral and Brain Sciences*, 36, 181–204.

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

Colombo, M., Seriès, P. (2012). Bayes in the brain—On Bayesian modelling in neuroscience. *The British Journal for the Philosophy of Science*, 63(3), 697–723.

Constant, A. and Clark, A. and Friston K. J. (2019). *Representation Wars: Enacting an Armistice through Active Inference.* Preprint available at: http://philsci-archive.pitt.edu/16125/

Dayan, P., Hinton, G.E., Neal, R.M., Zemel, R.S. (1995). The Helmholtz machine. *Neural Computation*, 7(5), 889–904.

Danks, D. (2014). *Unifying the Mind: Cognitive Representations as Graphical Models*. Cambridge (MA): The MIT Press.

Davies, M. (1995). Two notions of implicit rules. *Philosophical Perspectives*, 9, 153–183.

Davies, M. (2015). Knowledge (explicit, implicit and tacit): Philosophical aspects. In: J.D. Wright (ed.). *International Encyclopedia of the Social & Behavioral Sciences* (pp. 74-90). Oxford: Elsevier Ltd.

Dennett, D. (1983). Styles of mental representation. *Proceedings of the Aristotelian Society*, 83, 213–226.

Dołęga, K. (2017). Moderate predictive processing. In T. Metzinger, W. Wiese (eds),

Philosophy and Predictive Processing. Available online at: https://predictive-mind.net/papers/moderate-predictive-processing.

Downey, A. (2018). Predictive processing and the representation wars: a victory for the eliminativist (via fictionalism). *Synthese*, 195, 5115–5139.

Evans, G. (1981). Semantic theory and tacit knowledge. In: S. Holtzman, C. Leich (eds.). *Wittgenstein: To Follow a Rule* (118–137). London: Routledge and Kegan Paul.

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

Friston, K.J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Friston, K.J., Wiese, W., Hobson, J.A. (2020). Sentience and the origins of consciousness: From Cartesian duality to Markovian monism. *Entropy*, 22(5), 516.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, 59, 167–192.

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, 193, 559–582.

Gładziejewski, P., Miłkowski, M. (2017). Structural representations: causally relevant and different from detectors. *Biology & Philosophy*, 32, 337–355.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 290, 181–97.

Griffiths, T.L., Kemp, C., Tenenbaum, J.B. (2008). Bayesian models of cognition. In R. Sun (ed.). *Cambridge Handbook of Computational Cognitive Modeling* (pp. 59–100). Cambridge: Cambridge University Press.

Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences*, 27, 377–96.

Haugeland, J. (1991). Representational genera. W. Ramsey, S. Stich, and D. Rumelhart (eds.). *Philosophy and Connectionist Theory* (pp. 61–89). Hillsdale, NJ: Lawrence Erlbaum.

Helmholtz, H. (1855). Über das Sehen des Menschen (85–117). In *Vorträge und Reden von Hermann Helmholtz*. 5th ed. Vol.1. Braunschweig: F. Vieweg.

Hinton, G.E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, 11(10), 428–434.

Hipolito, I. (2019). A simple theory of every 'thing'. *Physics of Life Review*, 31, 79–85.

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

Hohwy, J. (2020a). New directions in predictive processing. *Mind and Language*, 35(2), 209–223.

Hohwy, J. (2020b). Self-supervision, normativity and the free energy principle. *Synthese*, DOI: 10.1007/s11229-020-02622-2.

Hosoya, T., Baccus, S.A., Meister, M. (2005). Dynamic predictive coding in the retina. *Nature*, 436, 71–77.

Huang, Y., Rao, R.P.N. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2, 580–93.

Hutto, D. (2018). Getting into predictive processing's great guessing game: Bootstrap heaven or hell? *Synthese*, 195, 2445–2458.

Hutto, D., Myin, E. (2013). *Radicalizing Enactivism: Basic Minds without Content*. Oxford (MA): The MIT Press.

Isaac, A. M. (2019). The semantics latent in Shannon information. *The British Journal of Philosophy of Science*, 70(1), 103–125.

Jones, M., Love, B. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 34, 169–188.

Kaiser, D., Quek, G. L., Cichy, R.M., Peelen, M.V. (2019). Vision in a structured world. *Trends in Cognitive Sciences*, 23(8), 672–685.

Kiefer, A. (2017). Literal perceptual inference. In T. Metzinger, W. Wiese (eds). *Philosophy and Predictive Processing*. MIND Group. Available online at: https://predictive-mind.net/papers/literal-perceptual-inference.

Kiefer A. (2019). *A Defense of Pure Connectionism*. Unpublished dissertation. The Graduate Center, City University of New York. DOI: 10.13140/RG.2.2.18476.51842.

Kiefer, A., Hohwy, J. (2017). Content and misrepresentation in hierarchical generative models. *Synthese*, 195, 2387–2415.

Kiefer, A., Hohwy, J. (2019). Representation in the Prediction Error Minimization framework. In: S. Robins, J. Symons, P. Calvo (eds.). *The Routledge Companion to Philosophy of Psychology* (pp. 384–409). London: Routledge.

Kirchoff, M.D., Parr, T., Palacios, E., Friston, K.J., Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, 15(138), 20170792.

Kirchhoff, M.D., Robertson, I. (2018). Enactivism and predictive processing: A non-representational view. *Philosophical Explorations*, 21, 264–281.

Korbak, T. (2019). Computational enactivism under the free energy principle. *Synthese*. DOI: 10.1007/s11229-019-02243-4.

Kreiman, G., Serre, T. (2020). Beyond the feedforward sweep: feedback computations in the visual cortex. *Annals of the New York Academy of Sciences*, 1464 (1), 222–241.

Li, B., Peterson, M.R., Freeman, R.D. (2003). Oblique effect: A neural basis in the visual cortex. *Journal of Neurophysiology*, 90(1), 204–217.

Miłkowski, M. (2013). *Explaining the Computational Mind*. Cambridge (MA): The MIT Press.

Orlandi, N. (2016). Bayesian perception is ecological perception. *Philosophical Topics*, 44, 255–278.

Orlandi, N. (2018). Predictive perceptual systems. *Synthese*, 195, 2367–2385.

Piccinini, G. (2015). *Physical Computation: A Mechanistic Account*. Oxford: Oxford University Press.

Pickering, M.J., Clark, A. (2014). Getting ahead: Forward models and their role in cognitive architecture. *Trends in Cognitive Sciences*, 18(9), 451–456.

Ramstead, M.J.D., Friston, K.J., Hipolito, I. (2020). Is the Free-Energy Principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy*, 22(8), 889.

Quine, W. V. O. (1970). Methodological reflections on current linguistic theory. *Synthese*, 21, 386–398.

Ramsey, W. (2007). *Representation Reconsidered*. Cambridge: Cambridge University Press.

Ramstead, M.J.D., Kirchoff, M.D., Friston, K.J. (2019). A tale of two densities: active inference is enactive inference. *Adaptive Behavior*. DOI: 10.1177/1059712319862774

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

Ryle, G. (1949/2000). *The Concept of Mind*. Chicago: University of Chicago Press.

Seriès, P., Seitz, A. R. (2013). Learning what to expect (in visual perception). *Frontiers in Human Neuroscience*, 7. DOI: 10.3389/fnhum.2013.00668

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition*, 112, 92–97.

Srinivasan, M. V., Laughlin, S. B., Dubs, A. (1982). Predictive coding: A fresh view of inhibition in the retina. *Proceedings of the Royal Society of London, B*, 216(1205), 427–459.

Teufel, C., Fletcher, P. (2020). Forms of prediction in the nervous system. *Nature Reviews Neuroscience*, 21, 231–242.

Ullman, T.D., Spelke, E., Battaglia, P., Tenenbaum, J.B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, 21(9), 649–665.

Williams, D. (2017). Predictive processing and the representation wars. *Minds and Machines*, 28, 141–172.

Williams, D. (2018). Predictive coding and thought. *Synthese*, 197, 749–1775.

Williams, D. (2020). Imaginative constraints and generative models. *Australasian Journal of Philosophy*. DOI: 10.1080/00048402.2020.1719523.