Synthese (2018) 195:2387–2415
https://doi.org/10.1007/s11229-017-1435-7

S.I. : PREDICTIVE BRAINS

Content and misrepresentation in hierarchical generative models

Alex Kiefer[1] · **Jakob Hohwy**[2]

Received: 19 November 2016 / Accepted: 8 May 2017 / Published online: 22 May 2017

Abstract In this paper, we consider how certain longstanding philosophical questions about mental representation may be answered on the assumption that cognitive and perceptual systems implement hierarchical generative models, such as those discussed within the prediction error minimization (PEM) framework. We build on existing treatments of representation via structural resemblance, such as those in Gładziejewski (Synthese 193(2):559–582, 2016) and Gładziejewski and Miłkowski (Biol Philos, 2017), to argue for a representationalist interpretation of the PEM framework. We further motivate the proposed approach to content by arguing that it is consistent with approaches implicit in theories of unsupervised learning in neural networks. In the course of this discussion, we argue that the structural representation proposal, properly understood, has more in common with functional-role than with causal/informational or teleosemantic theories. In the remainder of the paper, we describe the PEM framework for approximate Bayesian inference in some detail, and discuss how structural representations might arise within the proposed Bayesian hierarchies. After explicating the notion of variational inference, we define a subjectively accessible measure of misrepresentation for hierarchical Bayesian networks by appeal to the Kullbach–Leibler divergence between posterior generative and approximate recognition densities, and discuss a related measure of objective misrepresentation in terms of correspondence with the facts.

✉ Jakob Hohwy
Jakob.Hohwy@monash.edu

Alex Kiefer
akiefer@gmail.com

1 City University of New York Graduate Center, New York, NY, USA

2 Cognition & Philosophy Lab, Monash University, Melbourne, Australia

Keywords Problem of content · Misrepresentation · Functional role semantics · Structural resemblance · Prediction error minimization · Generative model · Recognition model · Kullbach–Leibler divergence · Variational Bayesian inference · Unsupervised learning

1 Introduction

The theory that the brain is an organ for prediction error minimization (PEM) has attracted considerable attention in recent cognitive science (for discussion and review, see Clark 2013, 2016; Hohwy 2013). On this theory, the brain is engaged constantly in predicting its own sensory input, registering the error in these predictions—the prediction error—and then minimizing this error as best it can in a long term perspective, by optimizing its internal model as well as by supporting actions that intervene on the world to produce expected input.

The PEM framework offers novel insights into the mechanisms underlying human perception, action, attention, learning and other cognitive processes, and suggests surprising connections between these processes. It also seems useful for the understanding of more subtle and troublesome aspects of mental life: the sense of self and bodily presence (Limanowski and Blankenburg 2013; Seth 2013; Apps and Tsakiris 2014; Hohwy and Michael 2017), emotions (Hohwy 2011; Seth 2013; Barrett 2016), the sense of agency (Hohwy 2016a, b), the sense of presence of perceived objects (Clark 2012; Seth 2014), top-down modulation and cognitive penetrability (Farennikova 2014; Vetter and Newen 2014; Lupyan 2015; Hohwy 2017), consciousness (Hohwy 2015a, b), depersonalisation disorder (Seth et al. 2012), autism (Cruys et al. 2014; Palmer et al. 2017), and schizophrenia (Hohwy 2004; Fletcher and Frith 2009; Corlett and Fletcher 2012; Adams et al. 2015).

In addition to this wealth of literature on applications of the PEM framework, there are more theoretical discussions of what it implies for our overall conception of the mind and its place in nature (Orlandi 2014; Hohwy 2015a, b, 2016a, b; Allen and Friston 2016; Bruineberg 2016; Burr and Jones 2016; Gallagher and Allen 2016; Kirchhoff 2016; Sims 2016; Hutto 2017). Several recent discussions consider the theoretical underpinnings of the framework as they apply to concerns in the philosophy of science, philosophy of mind and epistemology (Colombo and Seriés 2012; Hohwy 2015a, b; Gładziejewski 2016; Klein 2016; Loughlin 2016; Colombo and Wright 2017; Macpherson 2017).

Overall, there is considerable excitement about the PEM framework within cognitive science, tempered by healthy skepticism (see the open access volume, Metzinger and Wiese 2017). It is an extremely ambitious approach that, in some incarnations, has very wide explanatory scope that includes not just mind and cognition but also life and morphogenesis (Friston 2013; Friston et al. 2015). The theory is accompanied by empirical process theories about implementation in terms of brain function (such as predictive coding, for the case of perceptual inference), and the jury is still out about the proposed implementation details (Bastos et al. 2012). The result is an extraordinarily fertile intellectual climate that, it seems to us, offers the possibility of genuine unification among philosophical as well as scientific theories of mind, brain, and life.

Here, we step back from recent developments and challenges and delve into fundamental aspects of representation within the PEM framework. A primary philosophical attraction of the framework is its grounding in approaches to representation derived from statistical modeling that may shed new light on old philosophical problems concerning representational content (cf. Hohwy 2013, Chs. 1, 8). The statistical modeling approach to the mind is already part of the conceptual foundations of successful approaches to artificial intelligence and machine learning focused on generative models and deep neural networks (Hinton and Sejnowski 1999; Bengio et al. 2012). This overarching idea is itself less controversial than many of the details of the PEM framework, and repays careful scrutiny in its own right. That said, the details of the PEM framework provide a setting that helps make concrete and clarify outstanding issues for the more generic approach.

In what follows, we consider how an unsupervised prediction error minimizer may meaningfully represent the changing world in which it operates. We contend that the PEM framework, like many approaches to perception and cognition in machine learning, provides the tools for a theoretically mature internalist semantics centered on the notion of structural representation, which can be used to reply to well-known arguments against a representationalist interpretation of the PEM framework. We then describe and motivate various aspects of hierarchical Bayesian inference in order to explain how such structural representations may be implemented in systems that employ it. This discussion culminates in an account of the possibility of misrepresentation, which we shall analyze in terms of the some of the formal tools of variational Bayesian inference.

2 Structural representation and the problem of content

In this section we describe and motivate the view that mental representation involves exploiting structural resemblance between an internal model and an environment in order to guide action. We consider in some depth how representational content might be assigned to parts of a structural representation under such an approach, and compare the resulting view with more established accounts of mental representation. We then argue against anti-representationalist arguments from this perspective, emphasizing the indispensable use of a similar concept of representation in the field of machine learning.

2.1 Representation as exploitable structural similarity

Recently, philosophers interested in the PEM framework and related theories have revived interest in an often neglected approach to mental representation, based on structural resemblance between a representational vehicle and what it represents (see e.g. Cummins 1994; O'Brien and Opie 2004; cf. Gładziejewski 2016). This conception of representation is clearly endorsed in early statements about how PEM systems represent the world, i.e. by recapitulating its causal structure (see e.g. Clark 2013; Hohwy 2013). A version of this view has now been considerably elaborated and defended as a theory of mental representation (Gładziejewski and Miłkowski 2017).

Here, we wish to consider how this theoretical approach to representation might handle certain traditional philosophical concerns.

One fundamental question about mental representations concerns how precisely they get their contents—that is, in virtue of what they represent what they do, or get to be "about" what they are about. One widely discussed answer to this question, advocated in various ways by (Field 1977; Block 1994; Brandom 1994; Harman 1999; Sellars 2007) and many others, is that the contents of mental representations depend in some way on their overall functional (inferential, causal, conceptual, or evidential) roles within a cognitive system. Another, advocated by (Dretske 1981; Millikan 1984, 1989; Papineau 1984; Fodor 1990) and, again, many others, identifies the contents of a representation, at least in the basic case, with (some subset of) its regular external causes, or, what amounts to the same thing, with (some subset of) the causes about which the representation carries information (Millikan 1984; Papineau 1984). Under this second umbrella we include teleological theories, which due to well-known theoretical problems for the pure causal/informational theory such as the "disjunction problem" (Fodor 1990), appeal to facts about the etiology of representations — in particular, the learning histories that gave rise to them (Dretske 1981), or the evolutionary history in virtue of which the capacity to carry information about certain things was favored by selection pressures (Millikan 1984; Papineau 1984)—to fix their contents.[1]

The structural representation (henceforth, for brevity, "S-representation") approach seems to give a clear and distinct third answer: representations get their content via structural resemblance to what they represent. A precise definition of structural resemblance will be considered later in Sect. 5, but apart from that, there are at least two obvious reasons to qualify this answer. One is that while the structured representation as a whole gets its content via resemblance, the same need not be true of its parts, which get their content via the overall structural resemblance, by being placed in correspondence with parts of the represented structure. But the part-part relations themselves are not resemblance relations.[2] Second, as is often emphasized in discussions of the structural resemblance theory, simple resemblance is not sufficient for representation in any intuitive or theoretically useful sense. Any two items resemble one another in some respect, so an account of representation based on similarity alone would be trivial. It is also clear that representation is an asymmetric relation, while resemblance is not. Both of these problems can be solved by requiring that the structural resemblance be *used* or exploited by a cognitive system in order to make cognitive functioning effective (see, e.g., Godfrey-Smith 1996; Shea 2014; Gładziejewski and Miłkowski 2017). This captures the intuitive idea that a representation (mental or otherwise) serves as a proxy or stand-in for what is represented.

The first point shows that a form of *content holism* is entailed by the structural representation theory: the content of one part of the overall S-representation cannot be determined without simultaneously fixing the contents of the other parts. This holism

[1] We also here include Fodor's (1990) solution to the disjunction problem, based on asymmetries among nomic relations.

[2] Of course, we do not mean to rule out that parts may *also* be structured and function themselves as structural representations. Indeed, this is likely the case in hierarchically organized systems like those considered in this paper, but we lack space to consider this issue here.

seems to follow obviously from the fact that structural similarity is a relation between the whole representing system and what it represents, and not one entered into by the parts as such.

The second point, concerning the exploitability or "proxy" constraint on representation, merits careful consideration. Gładziejewski and Miłkowski (2017) unpack this constraint by requiring that structural similarity underwrite the successful operation of cognitive capacities based on S-representations. Depending on the capacity in question, successful operation may in turn be defined in terms of causal transactions with the environment (Gładziejewski and Miłkowski consider for example successful maze navigation in rats based on hippocampal maps). If this kind of case is taken as a model, the proxy constraint may be read as restricting the possible contents of any S-representation to entities and states of affairs in the environment with which its owner causally interacts.

But our cognitive capacities can also plausibly be directed at merely imaginary or hypothetical environments, as happens normally in cases of imagination and daydreaming and perhaps occasionally when radical misrepresentation of the environment occurs (cases of misrepresentation may be distinguished from non-representational failures precisely in that *the wrong thing* is represented). Importantly, we can engage with merely imagined environments cognitively, for example by reasoning about them or finding routes through them. Therefore if structural representations underlie mental representation generally, they must first and foremost be proxies for hypothetical worlds or states of affairs (defined in terms of their relevant structures).[3] In the context of capacities such as perception and action, these structures can be compared for veridicality with the actual environment, and are likely to underwrite successful causal interaction with it only insofar as they are accurate. Thus, despite the relational character of S-representation, the theory can be developed in an internalist way, which seems necessary to adequately accommodate imagination and misrepresentation.[4]

The considerations thus far suggest similarities between our S-representational and functional role approaches to mental content, which are also widely regarded as entailing content holism, and usually appeal to a "narrow" or internalist notion of content.[5] As O'Brien and Opie (2004) note, the way in which content is determined in functional role theories is also sometimes cast as relying on structural resemblance (for example, between a system of causally related vehicles and a system of inferentially related propositions—see (Fodor 1990, Ch. 1)). It is thus tempting to assimilate our structural representation proposal to the paradigm of functional role semantics.

Despite this superficial similarity, however, one might suppose that the two theories operate in very different ways. S-representation theories claim that structures represent

[3] This way of putting things may go to the heart of some of the more radical formulations of the implications of PEM-style accounts for the nature of the mind-world relation—for example, claims that perception is "controlled hallucination" (Grush 2004). Similarly, Geoff Hinton (one of the originators of contemporary models of perceptual inference involving generative models) claims in essence that the contents of mental states are hypothetical worlds (Hinton 2005).

[4] There may be ways of resisting the conclusion in the case of misrepresentation by distinguishing types of misrepresentation, as discussed in Sect. 5 below. However, the point seems difficult to sidestep with respect to imagination.

[5] It should be noted that there are 'wide' versions of functional role semantics as well; see Harman (1973).

whatever they lie in relations of (exploitable) resemblance to (in the case of interest to us, the causal structure of an organism's environment), but functional role theories do not usually claim that the network of inferential relations, to which causal transitions among vehicles stand in a relation of resemblance, is itself what is represented. Rather, the idea is that inferential roles implicitly define the contents, i.e. truth-conditions, of representations, and these conditions may themselves concern environmental states of affairs.[6] Even if inferential roles and thus contents are assigned to vehicles in a causal network on the basis of structural resemblance, the functional role approach does not require that a represented environment itself corresponds in structure to anything in the representation.

While it is true that there is no entailment in general from functional role semantics to S-representation of the environment, accepting the latter may nonetheless be consistent with an inferential role approach. To guarantee that a functional role theory of the sort just considered will entail structural resemblance between representing vehicle and environment, we need only assume that the inferential relations in terms of which contents are implicitly defined are *inductive* inferential transitions—what Sellars (1953) called "material inferences", such as that from "It's raining" to "The street is wet". Such inductive inferences keep track, in effect, of regularities in the world. Thus, the relevant inferential structure will be similar to the structure of the environment insofar as the latter is captured by the representation. And, moreover, it is still the case, as S-representation requires, that the resemblance between vehicle and environment causally explains the success of behavior based on the representation (Gładziejewski and Miłkowski 2017). We here conceive of this resemblance as mediated by a structure of inferential relations, but this just a different way of saying that fundamentally it is *possible* environments (whose causal structure is specifiable via probabilistic relations among propositions) that are represented, with representation of the immediate physical environment falling out as a special case that occurs in some (i.e. perceptual and action-involving) contexts.[7]

We suggest, therefore, that the contents of parts of a structural representation are (at least in the case of causal generative models of an environment) in effect determined by their internal functional roles. But even if one resists treating structural representation as a species of functional role semantics,[8] the preceding considerations seem to show that such representations cannot be assimilated to representations understood on the model of causal or teleological theories,[9] according to which content depends

[6] We thank an anonymous reviewer for an earlier version of this paper for pressing this crucial point, as well as the issue concerning internalism just considered.

[7] This is a decidedly Kantian reading of these ideas, but we believe it would be more procrustean to attempt to defend the opposite view according to which all representation in imagination is really somehow representation of one's physical environment. This is true even though actual sources of sensory input play an indispensable explanatory role within the PEM framework, and in fact are etiologically necessary to get any kind of representation off the ground.

[8] O'Brien and Opie (2004) distinguish strictly between functional role semantics and their preferred version of structural representation theory on the grounds that the former appeals to causal relations among vehicles while the latter appeals to physical relations. It is not obvious, however, why the latter category should preclude the former.

[9] Gładziejewski and Miłkowski (2017) draw a similar conclusion for different reasons.

ultimately on one-one (atomistic) relations between internal states and what they represent.

It should be noted that one may reject the latter view without supposing, implausibly, that the information that representations carry about environmental conditions is irrelevant to their content. A generative statistical model guided by the online minimization of prediction error, for example, is bound to carry information about the phenomena it models in its learned parameters (Hohwy 2013, and see below). This ensures that the parameters will tend to co-vary with individual causes. But if representation works fundamentally in the structural way sketched here, such covariation is one of its predictable *effects*, at least in perceptual cases, and is thus heuristically useful in identifying the contents of a representation (and perhaps in marking it as perceptual), rather than being constitutive of representation as such. The relation of indication that causal/informational and teleological semantic theories focus on is, from the perspective of S-representation, only one (non-essential) function of representations, derivable in some cases from their broader role of participating in exploitable structural resemblance.

2.2 Anti-representationalism

Before going on to consider in detail how S-representations might operate in hierarchical Bayesian systems, we address arguments against Bayesian representationalism. Such arguments would, if successful, constitute an objection to any such representationalist line of thought.

According to Nicoletta Orlandi, Bayesian theories, such as PEM, might describe how the brain "grows" structures that allow it to respond selectively to the environment, in a way that is sensitive to likely environmental causes of sensory input and even "attunes to the structure of the environment" (2014: 16), but without the internal states mediating this process deserving to be called "representations".

Orlandi argues that such internal states are, instead, merely "biases" that skew processing and operate in a causal manner:

> [I]n predictive coding accounts, priors, hyperpriors and likelihoods do not look like contentful states of any kind. They are not map-like, pictorial or linguistic representations. Having accuracy conditions and being able to misrepresent are not their central characteristics. Priors, hyperpriors and likelihoods rather look like built-in or evolved functional features of perception that incline visual systems toward certain configurations (2014: 25).

This point of view reflects a broader trend against the positing of internal representations in cognitive science (Hutto and Myin 2013), based primarily on concerns that the relevant notion of content, rooted in "indicator" or "detector", or causal covariation, relations, is too thin and liberal to be of explanatory value.

However, Orlandi does not consider the account of representation outlined above, which is in fact endorsed by most advocates of the PEM framework: that the generative model *as a whole* represents the environment in virtue of (partially) sharing the causal structure among its parts with the causal structure among events in the represented

environment. Indeed, priors, likelihoods, and such are not map-like representations, but something in the vicinity *is* the case: the overall network of hypotheses is an S-representation of the environment.[10]

Orlandi also argues (p. 19) that the causal intermediaries between proximal stimulation and high-level perceptual hypotheses are not representations because they "do not model distal or absent conditions", but are mere indicators of their proximal causes. Evidence for extra-classical receptive field effects adduced in favor of the PEM framework (see Harrison et al. 2007), which shows that what a given neural population indicates may vary contextually, complicates this picture somewhat, but neural activities may still be conceived of as indicating some combination of top-down and bottom-up proximal causes. Even so, Orlandi's description appears to beg the question against the account of representation sketched earlier, according to which these indicator relations are just strands in the broader functional roles that determine the contents of such states.

It may be objected that this is merely a representational gloss on a process that can be understood without appeal to representation (Hutto 2017). But to give this charge proper weight, at least a sketch of a non-representational explanation of the process must be given. And the alternative ecological theory endorsed by Orlandi tells precisely half the relevant explanatory story. It is best seen, not as a competitor to the representationalist PEM story, but as consistent with and in fact a part of it, insofar as either offers a genuinely explanatory account of cognition.

Orlandi relies on the framework of *natural scene statistics* (NSS): "NSS discovers the law-like or statistical regularities that [Bayesian] priors are supposed to mirror...By doing so, it makes appeal to the priors themselves secondary … Priors are explanatorily redundant in accounting for how a single percept is secured in ordinary perceptual circumstances—if such circumstances are low in noise. In such cases, the stimulus for vision constrains 'from the bottom' the perceptual hypothesis that is most plausible" (Orlandi 2016: 340).

This does not seem an optimal characterization of PEM to us. According to the PEM framework, when, for example a retinal projection gives rise to a perception of a single contour, this is because one perceptually infers to that contour as the best explanation, and the reason one does this, in turn, is that a single contour is the most likely cause given the retinal state, and one's internal model, shaped by empirical Bayes (discussed below), recapitulates the causal structure of the environment (i.e. represents its elements). Thus, the full PEM explanation, we would argue, encompasses the ecological perspective. On the other hand, without the additional insight supplied by the representational description, Orlandi's claim seems at best a promissory note for just the type of theory that PEM provides. To suppose that a retinal state gives rise to the perception of X simply because X is the most likely cause of the retinal state short-circuits the explanation.

[10] It is sometimes claimed that this notion of representation is also too liberal. The "exploitability" constraint mentioned earlier goes some way toward mitigating this. Also note that interesting, human-like cases at least are hard to come by: a system must, as a matter of empirical fact, be quite complex before it is able to structurally represent deeply hidden environmental causes.

Orlandi could perhaps respond that the full explanation, rather than requiring representation, requires only that the biases and detector-relations in the system be such that the brain is caused to activate the "single contour" hypothesis on the basis of the retinal input. But explaining how this is possible is, we think, the hard part. And it is precisely in filling in the details of the promissory note above that appeal to representation occurs in practice, as discussed next from the perspective of machine learning.

2.3 Unsupervised learning of internal models

Many of the core notions employed within the PEM framework derive from research in machine learning. Indeed, this field was among the first to offer accounts of perception based on prediction error minimization, and many of the seemingly esoteric formal notions underlying the prediction error minimization scheme are now standard textbook items in machine learning (Bishop 2007), as well as computational neuroscience (Dayan and Abbott 2001; Trappenberg 2010).

Philosophical issues about naturalizing representation are largely as remote from the concerns of machine learning researchers as they are from the work of practicing neuroscientists. But the notion of representation does play a key role in machine learning (arguably, at least as robust a theoretical role as it plays in neuroscience). It may thus prove fruitful to consider how the term 'representation' is employed in this discipline.

In general, 'representation' in machine learning (and in particular in connectionist approaches) refers to an internal state of a system that carries information, as it does throughout most of cognitive science. But the information-relations most often appealed to in connectionist theories are those *between internal states of a system*, for example the compressed information that one layer in a neural network carries about another. Internal information-relations can be thought of as capturing (probabilistic) functional/causal roles, bringing this conception of representation in line with the structural account sketched above. This is not surprising, since there is a large degree of overlap between work in machine learning and statistical modeling, and statistical models may be thought of in general as (often low-fidelity) simulations of the systems they model, representing via structural similarity (Cummins 1994).

We contend that 'representation' plays a rich and indispensable theoretical role in machine learning research, one which cannot be reduced to that of merely presenting a compelling gloss on processes that could be understood without it. In the remainder of this section, we attempt to illustrate this point by describing more fully the role that talk of representation plays in understanding connectionist accounts of perception and learning.

To date, much machine learning has focused on *supervised* learning. For example, typical artificial neural networks trained to classify images based on their content learn a mapping from image pixel intensities to a distribution over class labels. Learning proceeds by beginning with an arbitrary mapping and measuring the difference between the predicted and the correct class labels across a body of images, and adjusting the weights on the connections in the network to improve the mapping, as measured by the classification error.

This is of course implausible in itself as a model of how biological cognitive systems initially come to understand the world, since such an explicit "teacher" providing direct

feedback on the application of concepts is rarely present. Additionally, if the objective is to understand without circularity how representational content arises, then positing a supervisor will not help since the instructions given by the supervisor must themselves already be understood (lest an endless regress of supervisors is invoked). For this reason among others, *unsupervised learning*, learning that can proceed effectively without explicit, external feedback about whether the system's output is correct, is a central area of research in machine learning.

Over the past several decades, significant progress has been made on the problem of unsupervised learning, and it is now the focus of much state of the art research in the field (see, e.g., Goodfellow et al. 2014). The PEM framework is in effect a specific proposal about how unsupervised learning occurs in the brain (generalized to cover action). We may hope to make progress on problems of mental representation, then, by examining the way in which mappings between sensory input and accurate representations of the world can be learned without (in a sense) an external teacher. Such unsupervised learning seems capable of explaining from the perspective of the system itself (the "animal's perspective", in the terminology of Eliasmith 2000) how content arises.

The key assumption behind unsupervised learning schemes for models of perception is that, as Orlandi (2014), following Gibson, suggests, the driving sensory signal itself provides a very rich source of information—far richer, in terms of bits, than do the class labels used in supervised learning (Hinton and Sejnowski 1999). It is thus possible that we can understand how the brain ends up selecting hypotheses in a way that respects natural scene statistics but is mediated entirely by the proximal stimulus (plus, perhaps, genetic factors). A popular approach supposes that redundancies or correlations among the elements of the proximal stimulation—i.e., the distribution of the multidimensional sensory data itself—can be exploited to extract information about worldly causes (Hinton and Sejnowski 1999). This extraction is possible because the NSS, even those pertaining to deeply hidden causes, are implicit in the higher-order correlations among elements of the proximal stimulus.

Consider for example the information available to an unsupervised computer vision network exposed to images of natural scenes. Collections of pixels that look like edges are much more likely to occur than any given random scatter. These collections (which correspond to correlations among pixels) are correlated as well, leading to higher-order correlations that also correspond to environmental regularities. Correlations between edges define shapes likely to occur in natural images. Extrapolating, correlations between shapes may define larger features of visual objects, and visual objects can be defined in terms of those larger features. A little more speculatively, the same principle can be used to define objects in a way not tied to a particular sensory modality, in terms of correlations among representations in several modalities.

A hierarchically organized system can exploit these correlations by explicitly marking them at each level and using these marks as data for the next level up, a process which at each level is easy to implement in a neural network using simple rules for synaptic plasticity such as Hebbian learning rules (see, e.g., Bogacz 2017). The activities of neurons with receptive fields sensitive to first-order correlations can then be modeled by higher layers, and this process can be extended recursively, yielding a

hierarchical stack of filters from raw data (i.e. hierarchical *representation learning* or feature learning (Goodfellow et al. 2016)).

Even though activities in each layer are only immediately determined by activities in proximal layers, the overall system of dependencies among states recapitulates the system of dependencies among properties in the source of the training data (i.e. the world), and thus the system's states are genuine structural representations. Importantly, such learning schemes have been shown empirically to yield intuitively meaningful features in artificial networks, from oriented edge filters to higher-order features corresponding to the presence of objects in images (Rao and Ballard 1999; Bengio et al. 2012; Le et al. 2012), which can be used effectively for image classification based on semantic categories even though such classification was not an explicit goal of learning (Hinton 2007).

Crucially, as a rule, this bottom-up "feature grabbing" in unsupervised learning systems is made possible only by simultaneously learning a *generative model* that can be used to drive the units top-down, implementing a generate-and-refine cycle that, as we will see, can be interpreted in terms of successive rounds of hypothesis-testing. To take three prominent examples: the Helmholtz Machine (Dayan et al. 1995) learns an explicit generative model in its top-down connections; the "contrastive divergence" algorithm (Carreira-Perpiñán and Hinton 2005) used to train Restricted Boltzmann Machines accomplishes the same thing (increasing the probability of generating things similar to the data and decreasing the probability of generating arbitrary "fantasies") by adjusting symmetrical weights between layers; autoencoders (Bengio 2009) minimize the "reconstruction" error between input and output, which amounts to improving a generative model of the input.

Thus, a learning goal that makes sense for unsupervised systems is to attempt to learn to match the statistics in the input using a generative model (this same goal can be described also in terms of energy minimization, or minimization of the Shannon description length of the inputs—see, e.g., Dayan et al. 1995).[11] Even simple unsupervised learning schemes not usually thought of as involving generative modeling tend to work in similar ways. The k-means clustering algorithm, for example, can be seen as a variant of the expectation-maximization algorithm, which improves a generative model by minimizing free energy (Neal and Hinton 1998; Friston 2005).

It is not obvious how such talk of "models", "features", "representations", and so on, whose descriptive efficiency, dimensionality and other features can be measured, could be understood in non-representational terms, and indeed it would be difficult even to state one of the core achievements of unsupervised learning models (showing that useful representations can be learned even without explicit feedback about their correct application) without appeal to representation. Over the next several sections, we supplement the foregoing by describing how generative models can be learned via Bayesian inference in the hierarchically organized systems posited in the PEM framework.

[11] There is thus reason to think that perceptual systems employ generative models based on considerations about learning alone, in addition to the considerations about contextually biased interpretation of stimuli mediated by extra-classical receptive field effects.

3 From prediction error minimization to content

The PEM framework assumes that mental representation is a matter of probabilistic perceptual and active inference, precision optimization and complexity reduction, through which a system infers the hidden causes of its sensory input, in a process similar to that described above from the perspective of machine learning. Much of this is discussed in philosophical terms elsewhere (Hohwy 2013; Clark 2016). In this section we briefly review the main tenets of the framework and tie the discussion more closely to Bayesian inference. We close by addressing an early criticism of the PEM framework as equivocating on 'prediction'.

3.1 Bayesian inference and sensory evidence

Starting with a set of hypotheses $h_i, ..., h_n$, based on an overall model of what the worldly causes of sensory input might be, a system needs to infer the most likely hypothesis given the input (evidence), e, it receives. This can be captured in Bayesian terms, where the posterior probability of a hypothesis $p(h|e)$ is proportional to the product of the prior probability of the hypothesis $p(h)$ and the probability of the evidence given the hypothesis, or the likelihood, $p(e|h)$. Selection among two hypotheses, h_1 and h_2, then depends on the ratio $p(e|h_1)p(h_1)/p(e|h_2)\mathrm{p}(h_2)$.

Here, the likelihood term captures the notion of prediction error. If the predictions for the sensory evidence e generated by h_1 are much better—less erroneous on average—than those generated by h_2, then $p(e|h_1) > \mathrm{p}(e|h_2)$. In this way, the likelihood varies inversely with the prediction error. To consider a simple example, if h_1 is 'a dog is present' and h_2 is 'a sheep is present', we should expect the system in question to be able to assess $p(e|dog)$ vs. $p(e|sheep)$ (as well as the priors $p(dog)$ vs. $p(sheep)$). This suggests that the system must be able to discern patterns in e pertaining to dogs and sheep.[12] If e is a particular auditory input, a_1 (for example, captured as a time series of spikes in auditory cortex), then $p(a_1|dog)$ might be higher than $p(a_1|sheep)$. The sound might be a bark, which is better predicted by the hypothesis that a dog is present than the hypothesis that a sheep is present. We elaborate further on this by setting inference in a hierarchical context in Sects. 3.2 and 3.3 below.

It is possible to recast Bayesian inference to bring out more clearly the role of prediction errors in learning. We will assume that the probability distributions involved are normal Gaussian distributions that can be described in terms of their sufficient statistics, namely their mean and variance (or precision, which is the inverse of variance, cf. Mathys et al. 2014). Likelihoods and priors can thus be described using two numbers. The prediction is given by the mean of the prior, m, and the prediction error by the difference between m and the mean of the current sensory input, e.

Given this setup, how should m be updated to yield the new posterior estimate in the light of e? The system has a spectrum of possibilities at its disposal, from ignoring the prediction error altogether (not changing m) to changing m by the entire

[12] Though of course the system need not *begin* by representing such distinctions, for reasons discussed in Sect. 2.3.

prediction error, $e - m$. Bayes mandates a particular strategy: update relative to how much is already known (the prior) and how much is being learned (the likelihood). We can capture this by assigning a Bayesian *precision weight*, $\pi_l/\pi_p + \pi_l$, to the prediction error. This weight will ensure that the prediction error influences inference according to (i) how certain the system already is—the weight decreases as the prior precision π_p increases—and (ii) how much is being learned: the weight increases with the likelihood precision π_l.

For example, if there is a state of near total ignorance, then the prior precision is close to 0 and the weight on the prediction error will increase so that the posterior will largely follow *e (modulo* π_l). Since the posterior becomes the prior for the next round of inference, the prior precision should increase over time, up to the limit set by the variance of the evidence. This displays how Bayesian inference, updating the prior step-by-step, depends crucially on learning from prediction errors as well as on how the priors evolve over time as the old posterior becomes the new prior for the next step of inference. Simply by doing Bayesian inference, the prediction error will decrease, as accuracy improves and evidence is accumulated at each step.

3.2 Bayesian inference in a changing world

If we assume a simple universe with causes that do not interact, so that sensory input is a linear function of the causes, the type of Bayesian inference discussed above should suffice to arrive at a correct model. The rate of learning in such a universe would decrease as evidence is accumulated over time and the model's predictions approached perfect accuracy. In the actual world, however, causes tend to interact with one another and with sources of noise to create variable, nonlinear changes in the sensory input. This may cause increases in prediction error when the world changes in ways not captured by the model, even if other aspects of the world have been modeled perfectly.

Suppose for example that after having spent some time learning the true location of a sound through many steps of inference, an agent begins to move around the source of the sound such that the existing prior for its location begins to generate unexpected prediction error. The system as described so far would tend to ignore this change in the environment due to a high-precision prior that would treat the new prediction errors as outliers. As a result, real change in the environment, in this case the agent's own actions, will go unnoticed. In essence, the model is now underfitted: it is too simple to accurately represent the world.

The right response to this kind of underfitting would be to increase the learning rate (that is, give the prediction error more weight) so that new patterns can be learnt. In general, in a changing and uncertain world, the learning rate needs to be variable in a way that is responsive to such changes. To see how this adjustment of the learning rate can be governed by Bayesian norms, we return to the hierarchical filters implemented in neural net models (see Sect. 2.3), which can be interpreted in a Bayesian way—i.e., *hierarchical* Bayes.

In a hierarchical model, various causes and their patterns of interaction can be modelled in terms of their sufficient statistics, and the learning rate can be controlled

dynamically by adjusting expected precision. For example, such a model can represent both the source of a sound (the dog, say) and a further cause interacting with the dog (the master taking the dog for a walk). Now the system can anticipate changes in the pattern of sensory input, and thereby decrease prediction error. For example, the system will anticipate when the master tends to take the dog for a walk and hence when the pattern of barks is likely to change location; the learning rate can then go up (in effect weakening the influence of the prior for the dog's location) when the master is around so the dog's location can be estimated, and go down again when the dog is in the pen.

By judiciously varying the learning rate underfitting can be avoided, since now real changes are accurately picked up. Conversely, by decreasing the learning rate, some cases of *over*fitting can be avoided, since this will help prevent accidental changes in the prediction error from influencing the posterior too much. The link from cause, v_1, to effect e, is now modulated by a further cause, v_2, operating at a longer time scale. Of course the modulating link from v_2 might itself be modulated by yet another cause v_3 (e.g., the master finally gets permission to go away on holiday such that contrary to expectation the dog is not taken for walks).

The types of modulating causes can vary in many different ways. For example, the precision of the sensory input might periodically deteriorate (during rush hour, traffic noise on the nearby road adds a source of noise to e, making it hard to accurately assess the location of the dog). In this case, the learning rate should decrease rather than increasing even though some of the e-pattern may be indistinguishable from the pattern emerging when the dog is taken for a walk. This introduces further problems for inference, as the system may need to discern whether the change in e is caused by a changing mean or a changing precision.

This complicated causal structure in the world needs to be recapitulated in the Bayesian system's internal model, necessitating a hierarchy of interacting levels. By optimising means and precisions over time, the system should be able to vary its learning rate and update its posterior in Bayes-optimal fashion. That is, prediction error will always be weighted according to how much, in a global perspective, is already known, and how much is being learnt in a particular situation, given learned expectations for precision (Mathys et al. 2014). In such a system, each inference occurs within a represented context informed by long-term patterns in the sensory input. The accuracy of the inferences will tend to increase over the long term, even in a changing and uncertain world, because now interacting, modulating hidden causes are taken into account and the system can update its learning as new and unexpected things begin to happen.

Of course, prediction error will fluctuate as the system encounters volatile and changing environments. There may be scenarios where irreducible uncertainty increases, and there will inevitably be periods where, given the knowledge of the system, it will make inferences that are rational given its prior beliefs but which a better-informed observer would be inclined to say are false. This may happen during learning of volatility in the environment, for example as the system begins to model the changes in precision during dusk, dawn and full daylight—the system may rationally but falsely infer that a dog is a sheep. Mistakes cannot be entirely eliminated, given that even extremely likely hypotheses may be false on some occasions, as dis-

cussed later. But in the long-term average, systems engaged in hierarchical Bayesian inference should be able to keep their prediction error within narrow bounds.

Earlier, we discussed the ambition in machine learning of developing an unsupervised system for causal inference. Systems like those described here need only the sensory evidence and their own (subjective) priors that are refined over time to drive inference of the hidden causes. This suggests that a hierarchical Bayesian system could in principle learn representations in an unsupervised way. However, the term 'unsupervised' is, taken at face value, not quite appropriate. Though not guided by a supervisor's knowledge of the causes, inference is "supervised" or constrained by prediction error and by knowledge stored in the model itself subsequent to previous inferences, both of which stem from the world (since the priors are not set subjectively for each round of inference but extracted over time through successive inferences—this is known as *empirical Bayes*). Bayesian inference thus delivers rational supervision that stems from the sensory signal itself and ultimately from the causes in the world being modeled, since these shape the sensory signal and its evolution over time.

3.3 Content in Bayesian hierarchies

We conclude this section by explicitly relating the hierarchical Bayesian PEM architecture to the notion of S-representation outlined earlier. Broadly, hierarchical predictions tie individual hypotheses like 'There is a dog' or 'There is a sheep' to distinct spatiotemporal patterns of sensory attributes, and to hypotheses at other levels in the hierarchy.

In the hierarchical PEM framework as described by Friston (2005), the influence of units encoding hypotheses on the prediction error units at lower levels implicitly represents the likelihoods, and also implement the predictions (a bark will be more likely given the 'dog' hypothesis according to the model just in case activation of the 'dog' hypothesis causes a top-down prediction for characteristic auditory signals at lower levels of the model). The influence on the hypotheses at a given level from error units at the same level implements the prior for those hypotheses, which is also influenced by the hypotheses at the level(s) above, since these influence the error units (the likelihood $p(h_2|h_1)$, where h_1 is a higher-level hypothesis relative to h_2, in part determines the prior for h_2).

The resulting network of causes yields states with internal causal roles in virtue of which those states stand in for environmental entities and states of affairs via structural resemblance of the two causal networks. The prior probability of the "dog" hypothesis may, for example, depend on the represented setting (rural vs. urban; daylight vs. dusk). Concepts thus emerge in a linked network of sensory attributes and progressively more abstract ways of grouping them, and hypotheses are tested by passing messages (predictions and prediction errors) up and down in the hierarchy.

One challenge to the PEM framework that we would like to address here concerns the sense in which terms like 'prediction' are being used. In particular, it has been suggested (Anderson and Chemero 2013) that PEM trades on an ambiguity between contentful, personal-level predictions (i.e., about the weather or what will happen this evening) and the "predictions" that occur when, e.g., the activities of one neuronal

population are anticipated by those of another, as in predictive coding theories (e.g., Hosoya et al. 2005).

As our examples have illustrated, the predictions in question can be understood in neural terms—a prediction is a top-down signal that feeds into a lower cortical area and is compared with activity at that level. This is a "vehicular" description—couched in terms of the vehicle of representation. But each of these predictions also has a content, determined by its place in the overall network of causes. Thus, the predictions referred to in a description of the neural network architecture are also predictions about the world. Predictions can be picked out by their local anticipatory functional role within the brain, but their contents depend on their place in the broader functional architecture. Importantly, the content-based and vehicle-based descriptions of a prediction are systematically related. A given neural vehicle could in general not have the specific content it has without playing the specific causal role it plays, including its local causal role of probabilistically anticipating nearby neural activity.

Accordingly, the term 'prediction error' can also be understood in this dual light. Prediction errors are implemented by the state of the error units at each level of the hierarchy, which depend on local features of the top-down and driving sensory signals, and the way in which they interact. But from the perspective of S-representation, each such error is also an error in a genuine representational sense. Here caution is due, because there are two senses in which a top-down prediction may be considered to be (genuinely) "in error". First, it may fail to entirely "explain away" the sensory signal at the lower level, in which case there is a mismatch between the predicted content and the content arrived at at the lower level via approximate inference (see the following section for more detail). Or, it may fail to match the way things really are in the world.

Crucially, both of these sorts of errors depend, as such, on the states' having full-fledged truth-evaluable content. One person can contradict another only if both genuinely say something, i.e. if the speech-acts of both are contentful representations. And what each person says may contradict a true description of things, that is, fail to "correspond with reality". In short, we can in general assess representational error either by direct appeal to a mismatch between the content of a representation and the way things really are, or indirectly by comparing the content of one representation to another presumed to be true (as is done in practice). After discussing approximate inference in detail in the next section, we will return to this theme and see how it can be made more precise for the representations considered here, using the formal apparatus of Bayesian inference.

So far, we have seen that, if we begin with a system capable of hierarchical, empirical Bayesian inference, we get a plausible first step in allowing a resolutely intrinsic, unsupervised (or "world-supervised") perspective on the system, within which contents, conceived of in terms of causal roles similar to those of environmental causes, can arise.

4 Approximate inference through prediction error minimization

The PEM framework goes beyond a straightforward commitment to hierarchical Bayesian inference in that it focuses on ways in which Bayesian inference may be

approximated in biologically plausible ways via long-term prediction error minimization. In this section, we explain how such approximate inference works, first informally and then in slightly more formal terms. The latter discussion sets the stage for our conception of misrepresentation in Sect. 5.

4.1 Prediction error minimization and inference

Above, it was noted that Bayesian inference over time will tend to minimise average prediction error. As samples come in, the system will settle on the best overall estimate. New samples will disperse around this estimate since there will always be levels of irreducible noise, and thus always some prediction error, even after a long series of inferences. Some errors will be small (sitting close to the estimated value) and some will be larger (outliers). On average and over time, however, the error will tend towards a minimum value determined by the level of irreducible noise. Notice that here the long-term perspective is critical: prediction error may fluctuate and fail to be minimized in the short term.

This relationship between Bayesian inference and long-term average prediction error minimization can be turned around. If Bayesian inference minimizes prediction error, then it seems reasonable to expect that a system that is able to minimize long term average prediction error is able to approximate Bayesian inference. To be sure, this reversal needs to be handled with care since without further assumptions it is not guaranteed that a prediction error minimizer will in fact end with the estimates that would have come from applying Bayesian inference explicitly. But if Bayesian inference is the optimal way to improve a model and thus reduce the errors in its predictions, a system employing such a model must approximate the results of Bayesian inference in proportion to its ability to minimize prediction error.

This is important because a biological system such as the brain is more likely to be able to minimize prediction error than it is to apply Bayes' rule explicitly. To minimize prediction error in the simple case of a single neural representational unit encoding a probability distribution, neuronal populations just need to be tuned to some types of input and their variance so that the mechanism can match the sensory input with the expected mean and variance of the distribution. This matching can occur either by changing the expectations over time so that they better approximate the statistics of the perceptual input (this is perceptual inference), or the system can act in the world to change its sensory input to fit with the expectations—this is action or "active inference" (so-called because action will minimize prediction error, and minimizing prediction error is approximating Bayesian inference).

We have already discussed (in Sect. 3.2) how effective minimization of prediction error in a world such as ours requires a hierarchically organized system that scales up from singular representational units to harbour expectations of sensory input at increasing time scales, as well as expectations about how levels of uncertainty change over time, and about how causes in the world interact in ways that, under Bayes, would change the learning rate. The system must also learn how to balance perceptual and active inference, and must be able to manage this balance over the long term. This involves building up expectations for how the prediction error minimization

rate is affected by certain learning rates in perceptual inference VS certain types of behaviours. Further, the system must learn to adjust the complexity of its model so that it is neither underfitted nor overfitted, each of which leads to increased prediction error.

Over time, a system with these capacities should be able to approximate hierarchical Bayesian inference by employing a number of strategies, which minimize prediction error and thereby increase the joint probability of its model and the sensory input it receives. We then have an informal account of how a system that minimizes long term average prediction error can be expected to approximate Bayesian inference, that is, to progress towards approximating Bayesian models that represent the world. Crucially, the PEM framework shows how this can be done using biologically plausible, neuronal mechanisms.

4.2 Approximate variational inference

So far we have considered informally the relationship between prediction error and inference. Parts of this argument can however be considered from a slightly more formal perspective, borrowed from statistical physics and now adopted in machine learning (Bishop 2007). Below we offer a condensed version of the core ideas, leading to the general notion of variational Bayes that sits behind the ideas we have just gone through.

Return first to exact inference. The aim is to infer the probability of the hypothesis h given the evidence e, $p(h|e)$. We know from Bayes' rule that $p(h|e) = p(e|h)p(h)/p(e)$ but rather than trying to tackle this directly we can adopt an approximate "recognition" probability distribution (or density), $q(h)$, and try to make it approximate the true posterior, $p(h|e)$. This is useful because computing the latter directly may require solving very complex integrals in the marginal probability function for the denominator, which is often intractable.

The measure of success for such a strategy is the Kullbach-Leibler divergence, or relative entropy, between $q(h)$ and $p(h|e)$, $\mathrm{KL}(q(h)||p(h|e))$. The KL divergence is either 0 or positive and, roughly, measures the difference in average information between two probability distributions (that is, for the discrete rather than continuous case, the entropies are subtracted and parts of the expression are reorganized as a log ratio such that $\mathrm{KL}(p||q) = \Sigma p(i)\log(p(i)/q(i)) = -\Sigma p(i)\log(q(i)/p(i))$. There is a twist in that, while entropy or average information of a distribution p over a set of events is a sum of the log probabilities of the events weighted by the probability of each event occurring under p, the KL divergence between p and q, $\mathrm{KL}(p||q)$ privileges p, in that it weights both sets of log probabilities for events i by $p(i)$. This allows the log probabilities, when subtracted, to be expressed as the single log ratio).

Once we use the fact, from probability theory, that $p(h|e) = p(e,h)/p(e)$, it is relatively simple to show that $\mathrm{KL}(q(h)||p(h|e)) + \Sigma q(h)\log p(e,h)/q(h) = \log p(e)$. Here $\log p(e)$ is the log probability of e, known as the surprise or model evidence. Since we are considering $p(h|e)$, the evidence e is given and this means that $\log p(e)$ is a fixed negative value (negative because it is the log of a value between 0 and 1) and this in turn constrains what happens on the left side of the expression. Specifically,

since e is given, we can only vary $q(h)$ and as we do that in the second term, which is always negative, we must be controlling KL, which is never negative (this relation between the terms is easy to verify with a simple set of numbers). The second term is thus a lower bound on the log probability of the evidence, often written as L, so that the overall expression can be simplified to KL + L = log $p(e)$. This shows that we can approximate the true posterior just by manipulating $q(h)$ to maximize the function for L, $\Sigma q(h) \log p(e, h)/q(h)$, rather than attempt to engage in exact inference.

The burden then moves on to considering how $q(h)$ can be used to perform this job. Variational Bayes suggests that this can be done by dealing with the parameters of the model, h, one by one, assuming that the other parameters are known even though in fact they are not (Bishop 2007; Bogacz 2017). One might imagine that the recognition density $q(h)$ "wobbles" around and begins to approximate the true posterior, $p(h|e)$, during such variational Bayesian inference. We can view the earlier more informal discussion of prediction error minimization as an instantiation of variational Bayes (properly speaking for probability density functions and under the (Laplace) assumption that the probability densities are Gaussian); for discussion see (Friston 2010).

The technical literature here is immense and mathematical but there is a key philosophical point that is relatively easy to bring across. The brain is proposed to approximate Bayesian inference by adopting some parameterised internal model h and engaging in variational Bayes on $q(h)$. Essentially, since we are both approximating the posterior p($h|e$) and improving the lower bound on the (log) probability of the evidence, this means changing parameters so as to maximise the joint probability of the hypothesis h and the evidence e. The processes outlined as part of the PEM framework above spell out how maximizing this joint probability can be done via perception, action, attention, and model selection. In principle, variational Bayes can be conducted in a number of different ways—PEM spells out one of these, which, in particular, gives a role for action. Predictive coding, which is often mentioned in these debates and was assumed above in our discussion of predicting sensory evidence, is one process theory for the perceptual inference part of PEM.

Given the conceptual tools just introduced, we can return again to unsupervised learning and formulate a perspicuous way of thinking about how it is possible: as priors and likelihoods of hypotheses are mutually adjusted in light of prediction error, a reliable channel of information transmission is set up between neural populations encoding sensory input and higher-level representations—an approximate recognition model.[13] In the other direction, a reliable channel is also constructed from those high-level representations back down to the sensory input layers—the generative model. Since sensory input drives a signal up through the hierarchy, which reaches the highest levels, and then those high-level representations send signals back down through the hierarchy to the lowest levels, we can think of the overall network as learning a mapping from sensory input, through high-level representations of causes, back onto sensory input. Hierarchical message-passing and revision in light of prediction error can then

[13] It should be stressed that the construction of the reliable information channel and the development of meaningful representations are constitutively related. These are two ways of describing the process whereby causal structure (which representation exploits) is set up within cortical hierarchies.

accomplish the kind of unsupervised learning we have been seeking. The supervisor is, as we mentioned earlier, literally the world, in the form of the proximal stimulus at the sensory input channel, caused by states of affairs in the world.

What makes this more than a simple loop, or a matter of slavishly imitating the sensory input just received, is the fact that higher-level representations (a) receive input from (and output signals to) more than one sensory modality, thus integrating multiple sources of sensory information (forming a joint generative model over wider slices of the input channel), (b) tend to compress information from sensory input channels, so that succinct descriptions of sensory input (and motor output) result, and (c) are already active given the context established by previous experience over several time scales, so that information is integrated over time (one of the core features of PEM systems). Because of these facts—and relatedly, the articulation of the loop into distinct stages of hierarchical processing—the system is able to generalize and make predictions not just about the current input but about temporally surrounding inputs, and even inputs in the distant future, via refinement of its internal representations.

5 Representation and misrepresentation

The brain is proposed to be an organ for prediction error minimization. This proposal is, as we have discussed, best understood in terms of a system that by adjusting various parameters minimizes long-term average prediction error and thereby approximates Bayesian inference. The result of this process of long-term prediction error minimization can be cast in terms of minimization of the KL divergence between a recognition distribution (or density) $q(h)$ and the distribution for the true posterior under the generative model. Inference of the states of the world given by $q(h)$ is correct when $q(h)$ corresponds to what Bayesian inference would yield, i.e. when $\mathrm{KL}(q(h)||p(h|e)) = 0$. In this section we propose a measure of misrepresentation based on this idea.

5.1 Misrepresentation and KL divergence

In an early computational model of Bayesian perceptual inference, Hinton and Sejnowski (1983) propose the KL divergence between a neural network's generative distribution over sensory input states and the distribution over sensory states determined by the external environment as "an information theoretic measure of the discrepancy between the network's internal model and the environment" (p. 451). This KL divergence is thus a candidate measure of *misrepresentation*, but for reasons discussed below is limited as a measure appropriate to structural representation. After discussing these limitations, we adopt a similar proposal set specifically in the context of our above treatment of hierarchical, structured representations characteristic of hierarchical generative and recognition models in PEM systems.

Before we begin this discussion, we address two preliminary points. One concerns the definition of structural similarity. For present purposes, we do not need a very precise definition, but at least for comparisons of probabilistic causal models such as Bayesian networks, an appropriate measure would be one that takes into account both the topological properties preserved across two graphical models and the degree

of similarity in dependencies between similarly situated nodes. We are not aware of a single measure that explicitly combines both properties, but methods for structural comparison of various types of graphical models (including "inexact matching" that is weaker than isomorphism or homomorphism) are an active area of ongoing research, and there are many options to choose from (see, e.g., Gallagher 2006 for a survey). For concreteness, we can provisionally require a strict homomorphism from one graph to the other as a prerequisite for any degree of structural similarity, with the degree depending on the difference between the joint distributions determined by the conditional dependencies, as discussed below.

The second preliminary point concerns the fact that *a priori*, there are at least two sorts of possible misrepresentation in systems that employ S-representations: mismatch in structure (which, as the previous paragraph suggests, is likely a graded notion) and improper registration of the representation with the represented system (as happens for example when one locates oneself on a map incorrectly, or, in the case of interest at present, infers a false hypothesis within a generative model).[14] The account of misrepresentation offered here directly captures the first type of misrepresentation, whose relation to the second type is discussed shortly.

With these preliminaries out of the way, consider first whether Hinton & Sejnowski's measure makes sense as a rendering of the intuitive account of structural misrepresentation sketched so far. On that account, representation is a matter of partial structural resemblance, and misrepresentation occurs when there is some divergence between the structure of the model and the structure of the thing modeled, against the background of an overall exploitable similarity sufficient to ground content.[15]

Prima facie, Hinton & Sejnowski's measure seems not to quantify misrepresentation in this sense, because it only concerns the sensory input states and is agnostic about the structure of the model. It may be supposed that an agent that can predict the sensory input must understand the structure of the environment, but two very different networks of causal factors could, in principle, give rise to the same distribution over observed variables (sensory input states). In real cases, the chance of being able to generate the same distribution with an alternative structure of causes is negligible even if intelligible. But the divergence at the sensory periphery still seems at best a proxy for misrepresentation in the intuitive sense.

We suggest instead taking the divergence between $p(h|e)$ and the approximate recognition distribution $q(h)$, which tracks prediction error in PEM systems, as a measure of misrepresentation. On this proposal, as long as $\mathrm{KL}(q(h)||p(h|e)) > 0$, the inferred state of the world given by $q(h)$ is a misrepresentation. For example, if $q(h)$ says that the most probable state of the world conditional on some particular sensory input is that the dog is going for a walk, and this differs from the true posterior, which

[14] We thank an anonymous reviewer for pressing us to clarify this distinction, which is also drawn by Gładziejewski (2016), who uses the "X"-on-a-map example referred to here.

[15] Clearly, if each generative model represents only the hypothetical world whose causal structure is isomorphic to it, there can be no misrepresentation, and thus arguably no genuine representation either. We need an independent standard of comparison to define misrepresentation, but note that this target need not be the actual world: it could be one specified by a fictional description, for example.

might say it is most probable under these sensory conditions that the dog is in the pen, there is misrepresentation.

This proposal has the merit of focusing not just on the model's predictive accuracy with respect to input, but on the probabilities of the various hypotheses conditional on the sensory input. Elsewhere, Hinton and Sejnowski take the same divergence to be a measure of "the difficulty of performing inference with the model" (Hinton and Sejnowski 1999: xiv). Crucially, since conditional dependencies among hypotheses determine the structure of the model, this measure also tracks degree of structural representation, and therefore dovetails with the account of representation outlined earlier.

There is further work to be done to show that this a satisfactory account of misrepresentation, however. As noted in Sect. 3.3, representations can be assessed for error in at least two ways: by comparison with other representations or by comparison with the thing represented. The notion of structural representation by statistical models can in principle accommodate both of these: models may be structurally compared both with other models and with the modeled phenomenon. Thus, $\mathrm{KL}(q(h)||p(h|e))$ compares, in effect, the recognition model and generative model. So, the proposal thus far would be that a comparison, using the KL divergence, of generative and recognition models suffices as an account of misrepresentation.

There may seem to be a problem with this proposal, in that the generative model is implemented by "backward" (top-down) synaptic connections and the recognition model is implemented by "forward" (bottom-up) connections (in addition to the lateral connections shared by both in Friston's (2005) model). If distinct structures (captured in terms of graphical models) determine distinct contents, the two models would then seem to represent very different things. But the fact that the recognition model is the inverse of the generative model (i.e. a mapping from sensory states to causes) ensures that, in a well-trained system that minimizes prediction error, these models *will* share structure despite the inversion (similarly, perhaps, my visible surface and that of my mirror-image share structure despite the latter's being reflected around the vertical axis).[16] Though the two models factor the joint distribution over hypotheses and evidence differently, they can still be expected to share structure to the extent that these joint distributions converge.

Consider next that minimizing $\mathrm{KL}(q(h)||p(h|e))$ guarantees at best that no misrepresentation occurs in a *relative* sense: $q(h)$ does not misrepresent the world as it is represented by $p(h|e)$ (or, $q(h)$ does not misrepresent the hypothetical world represented by $p(h|e)$). To the extent that the probability of the hypotheses under the models differ, the mapping defined by the recognition model will fail to invert the generative model, so prediction error and thus relative misrepresentation are more likely. But with respect to the external world, this accuracy of representation is only as valuable as the accuracy of the posterior distribution under the generative model itself, which will be improved over the course of learning using empirical Bayes.

[16] In Friston's model (2005), only the top-down connections introduce nonlinearities, but this is just a further way in which the two models diverge while still sharing structure. The whole point, of course, is that the recognition model is a (to some degree crude) approximation of the posterior under the generative model.

A true measure of accurate structural representation of an environment would of course compare the structure of causes of sensory input in the generative model with the actual structure of environmental causes. Assuming that real-world causal relations can be understood probabilistically, we could measure the true degree of misrepresentation by measuring the difference between the actual conditional distributions that environmental causes determine over one another (including sensory input states), call this $c(h)$, and the generative distribution $p(h|e)$. $\mathrm{KL}(p(h|e)||c(h))$ is thus a promising measure of objective misrepresentation.[17]

Of course, from a Bayesian perspective, uncertainty is ineliminable, and even the most likely hypotheses may on some occasions turn out to be false. There presumably will have been some determinate set of causes that actually contributed to producing sensory evidence on any given occasion, and the model only avoids misrepresenting the world if inference selects the corresponding hypotheses, which it can be only asymptotically likely to do as probabilities assigned to actually true and false hypotheses approach 1 and 0, respectively. The best we can hope for is a generative model that assigns high probability to the true causes most of the time.

But note that because the internal model is "calibrated" to the environment at the sensory periphery, a model that avoids structural misrepresentation is also highly likely to avoid misrepresentation of the second type discussed earlier, which occurs, simply, when hypotheses are selected whose truth-conditions are not in fact satisfied. Thus in the case of pairs of generative and recognition models, a measure of structural misrepresentation serves also as a measure of *average* misrepresentation in the sense of false hypothesis selection.

The upshot is that we have two measures of (average) misrepresentation. The first is a direct, "external" measure $\mathrm{KL}(p(h|e)||c(h))$, which an outside observer could in principle determine but which may not be very useful in practice due to the difficulty of determining c. The second, $\mathrm{KL}(q(h)||p(h|e))$, measures relative, "internal" misrepresentation. Unlike the external measure, it is accessible from the "animal's perspective" and so can be exploited in learning. As we have shown above, learning is supervised by the world—by the truth—and therefore this internal measure is very likely over time to track the external one.

We therefore propose that PEM comfortably allows a new and coherent account of misrepresentation that fits well both with an intuitive functional-role account and with the notion of representation as exploitable structural similarity. The current recognition model, $q(h)$, assigns probability to various contentful hypotheses, and is an approximate inversion of the generative model $p(h|e)$. The divergence between these two models measures misrepresentation. The accuracy of the generative model is measured in terms of a KL divergence between it and the true states of affairs, $c(h)$. This proposal has the advantage of showing how a structural, functional-role account of representational content can capture the requirement that misrepresentation be relative to actual represented states of affairs, while also showing how the system can

[17] The distribution c of course specifies the relevant natural scene statistics. NSS are therefore important in principle to understanding Bayesian models of perception, as Orlandi claims (see discussion in Sect. 2.3), even if such models are interpreted in representational terms.

avoid misrepresentation without the need for supervised learning, by improving the generative model via hierarchical inference.

5.2 Misrepresentation, ignorance, and model fitting

The foregoing account of misrepresentation in PEM systems and related systems that implement unsupervised learning seems to us to amount to real progress on the problem of content and misrepresentation, on the assumption that we are such systems. No doubt there are objections and opportunities for refinement of the proposal as presented here, but it seems to us sufficient to establish a serious representationalist interpretation of the PEM framework. In this section we seek to defend and elucidate the proposal by considering the sense in which overfitted and underfitted models misrepresent, and how this issue interacts with learning.

A potential issue with our proposal is that models are refined during learning, and are necessarily underfitted to the data early on. Assuming we start from scratch or from a minimal set of evolutionarily endowed priors, neither the recognition nor the generative model will really earn their names at the very earliest stages, since on the S-representation account adopted here, world-related contents only emerge once a system of largely veridical representations arises, which in PEM systems occurs once the system has been refined significantly via (approximate) inference.

Relatedly, traditional accounts of learning that appeal to hypothesis-testing (Fodor 1975) encounter difficulty in explaining where the hypotheses come from. PEM solves this by supplementing the hypothesis-testing story with the machinery of unsupervised representation learning. In effect, there is a smooth transition between early learning that defines the hypothesis space and subsequent learning that refines the model. Rather than an innate stock of concepts, we need only the innate representational capacity of a statistical model. The blurred distinction between constructing the hypothesis space and refining the probabilities of various hypotheses in light of evidence may seem philosophically troubling, but it is coherent given that structural resemblance is a graded notion. We may suppose that any model trained via causal interaction with the world begins by representing it rather indefinitely, and ends with a set of relatively determinate hypotheses.

This graded transition between inchoate representation at the earliest stages and fairly determinate content at the later stages suggests that the processes of representation learning (i.e. development of representations of sensory causes) and learning in the ordinary sense (i.e. learning what is true about the world) are not fundamentally distinct in kind. This is perhaps surprising, because the latter involves updating one's model in light of error, and one is tempted to say that at early stages, models do not *misrepresent* the world in the sense of inaccurately representing it, but only fail to represent it as determinately as could be done. This seems to amount merely to *ignorance* (or agnosticism), and intuitively there is a distinction between ignorance and misrepresentation (e.g. false belief).

The intuitive distinction between ignorance and misrepresentation seems to be at play in framing the disjunction problem for causal theories of content. Standard treatments of this problem assume that, for example, a 'dog-or-sheep' hypothesis is *true*

of a sheep (since a sheep is either a dog or a sheep), while the 'dog' hypothesis is not. Though the disjunction problem as such does not arise for the structural representation proposal, simple models may lack the structure to mark such distinctions as that between dogs and sheep and may invoke the same hypothesis to explain sensory input caused by both. We may thus consider what our proposal entails about such "disjunctive" hypotheses. There turn out to be reasons to suppose that from a global perspective, the application of disjunctive contents implies error.

Consider first that in viewing a sheep under degraded perceptual conditions, the 'dog-or-sheep' hypothesis is likely to minimize prediction error just as well as the 'sheep' hypothesis. It thus seems that they should involve the same degree of misrepresentation, measured in terms of KL divergence. However, this is true only when considered in the short-term perspective. Whereas both hypotheses explain away the current evidence, there may be a difference in long-term prediction error minimization. In particular, having the ability to represent dogs and sheep rather than dogs-or-sheep will help explain away non-linear evolutions of the sensory input due to interactions between causes.

This point can be seen more intuitively from the perspective of model-fitting. The model harbouring the disjunctive hypothesis rather than distinct 'dog' and 'sheep' hypotheses ignores some of the genuine causal structure of the world and therefore can be said to be underfitted—it averages out differences between dogs and sheep and thereby underestimates the precision of the sensory input, because it fails to model a modulating cause (the deteriorating viewing conditions). This makes it a case of ignorance, since the model is as yet unable to model the relevant full, deep hierarchical structure of the causes of its sensory input.

The contrast here is overfitting, in which parameters are fitted to noise, as would happen in a (counterfactual) world in which a distinction between dogs and sheep does not reflect any genuine underlying causal structure, but spurious regularities happened to occur in the sample on which learning was based. Crucially, both underfitted and overfitted models will fail to predict the sensory input as well as an optimally fitted model, and thus will on average increase their (internal and external) KL divergences relative to the optimal case, even if this difference is temporarily obscured in degraded perceptual conditions.

Thus, if we take the KL divergence proposal as a guide, we should in fact expect that there is no deep difference between ignorance and (structural) misrepresentation. Deploying the 'dog-or-sheep' hypothesis will after all involve misrepresentation. On reflection, this is not surprising given a structural account of representation: on that account, as stressed earlier, content is determined holistically for all representations in the system, and while it is accurate to classify a dog as either a dog or a sheep, it is to some extent inaccurate to classify anything as a dog-or-sheep, at least in our world, because the latter implies a conceptual scheme that takes the world to be simpler than it is (at a particular level of abstraction) and thus misrepresents it.

This can be seen in more detail by considering the probability of some visible feature f (for example, size or face shape) conditional on various hypotheses. The likelihood, $p(f|dog)$ will of course differ from $p(f|sheep)$ for most features. We may suppose for the sake of argument that the objective probability of f taking on a certain value given that the thing in the environment is either a dog or a sheep, $p(f|dog\ or\ sheep)$, is a

bimodal distribution. A system that harbors both the 'dog' and 'sheep' hypotheses may in principle accurately represent this distribution, though in perceptual inference the disjunction of hypotheses would not be selected because hyperpriors, we may assume, dictate that animals in the environment are determinately members of one species or another. But a system that instead harbored a 'dog-or-sheep' hypothesis would, at least on the PEM account sketched above, attempt to model this bimodal distribution using a single Gaussian (perhaps as geologists did for jade before leaning that jade is either jadeite or nephrite). This would necessarily lead to a misrepresentation of the true probabilities for f given the disjunction of hypotheses. In short, systems that entertain disjunctive hypotheses are not equivalent to systems that are capable of entertaining disjunctions between hypotheses.

There is less to be said about overfitting, because it seems a straightforward case of imputing causes in the world where none exist. It is similar to a case in which ink is spilled on a map used to navigate an environment, introducing spurious represented features that don't correspond to anything in the modeled structure.

6 Concluding remarks

In this paper, we have defended a proposal for understanding the representational contents of mental states implemented by statistical models that is functionalist in spirit, but that relies also on the notion of structural representation and recasts traditional functional-role approaches in precise terms derived from the theory of variational Bayesian inference. As we have indicated throughout, this proposal is one way of developing existing work on structural representations within the PEM framework.

To recap, we have first argued on conceptual grounds for a marriage of structural and functional-role approaches to content. We then argued that well-known anti-representationalist criticisms have missed this sort of proposal, and further motivated the approach by arguing that it is consistent with approaches to content implicit in theories of unsupervised learning in neural networks. In the remainder of the paper, we described the prediction error minimization (PEM) framework for approximate Bayesian inference in some detail, and discussed how representations arise within Bayesian hierarchies. After explicating the notion of variational inference, we appealed to the KL divergence between posterior generative and approximate recognition densities in a hierarchical Bayesian network to define a subjectively accessible measure of misrepresentation, and showed how this measure relates to objective misrepresentation in terms of correspondence with the facts.

Acknowledgements We wish to thank Michael Kirchhoff and two anonymous reviewers for comments. JH is supported by Australian Research Council Grants FT100100322 and DP160102770, and by the Research School Bochum and the Center for Mind, Brain and Cognitive Evolution, Ruhr-University Bochum.

References

Adams, R. A., Huys, Q. J. M., & Roiser, J. P. (2015). Computational psychiatry: Towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery & Psychiatry*, *87*(1), 53–63.

Allen, M., & Friston, K.J. (2016). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*. doi:10.1007/s11229-016-1288-5.
Anderson, M., & Chemero, A. (2013). The problem with brain GUTs: Conflation of different senses of 'prediction' threatens metaphysical disaster. *Behavioral & Brain Sciences*, *36*, 204–205.
Apps, M. A. J., & Tsakiris, M. (2014). The free-energy self: A predictive coding account of self-recognition. *Neuroscience & Biobehavioral Reviews*, *41*, 85–97.
Barrett, L. F. (2016). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*, 1.
Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711.
Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, *1*(2), 1–127.
Bengio, Y., Courville, A. C., & Vincent, P. (2012). Unsupervised feature learning and deep learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828.
Bishop, C. M. (2007). *Pattern recognition and machine learning*. Cordrecht: Springer.
Block, N. (1994). Advertisement for a semantics for psychology. In S. P. Stich & T. Warfield (Eds.), *Mental representation: A reader*. Oxford: Blackwell.
Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, *76*(Part B), 198–211.
Brandom, R. (1994). *Making it explicit*. Cambridge: Harvard University Press.
Bruineberg, J. (2016). The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese*. doi:10.1007/s11229-016-1239-1.
Burr, C., & Jones, M. (2016). The body as laboratory: Prediction-error minimization, embodiment, and representation. *Philosophical Psychology*, *29*(4), 586–600.
Carreira-Perpiñán, M. A., & Hinton, G. E. (2005). On contrastive divergence learning. In *Proceedings of the tenth international workshop on artificial intelligence and statistics*.
Clark, A. (2012). Dreaming the whole cat: Generative models, predictive processing, and the enactivist conception of perceptual experience. *Mind*, *121*(483), 753–771.
Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral & Brain Sciences*, *36*(3), 181–204.
Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. New York: Oxford University Press.
Colombo, M., & Seriés, P. (2012). Bayes in the brain–On Bayesian modelling in neuroscience. *The British Journal for the Philosophy of Science*, *63*, 697–723.
Colombo, M., & Wright, C. (2017). Explanatory pluralism: An unrewarding prediction error for free energy theorists. *Brain and Cognition*, *112*, 3–12.
Corlett, P. R., & Fletcher, P. C. (2012). The neurobiology of schizotypy: Fronto-striatal prediction error signal correlates with delusion-like beliefs in healthy people. *Neuropsychologia*, *50*(14), 3612–3620.
Cummins, R. (1994). Interpretational semantics. In S. Stich & T. Warfield (Eds.), *Mental representation: A reader*. Oxford: Blackwell.
Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience*. Cambridge, Mass: MIT Press.
Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation*, *7*(5), 889–904.
Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, MA: MIT Press.
Eliasmith, C. (2000). How neurons mean: A neurocomputational theory of representational content. Ph.D., Washington University in St.Louis.
Farennikova, A. (2014). Perception of absence and penetration from expectation. *Review of Philosophy and Psychology*, *6*, 1–20.
Field, H. (1977). Logic, meaning and conceptual role. *Journal of Philosophy*, *74*(69), 379–409.
Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, *10*(1), 48–58.
Fodor, J. A. (1975). *The language of thought*. Cambridge: Harvard University Press.
Fodor, J. A. (1990). *A theory of content and other essays*. Cambridge, Mass: MIT Press.
Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions: Biological Sciences*, *369*(1456), 815–836.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.
Friston, K. (2013). Life as we know it. *Journal of The Royal Society Interface*, *10*(86), 1–12.
Friston, K., Levin, M., Sengupta, B., & Pezzulo, G. (2015). Knowing one's place: A free-energy approach to pattern regulation. *Journal of The Royal Society Interface*, *12*(105), 20141383.
Gallagher, B. (2006). Matching structure and semantics: A survey on graph-based pattern matching. In *AAAI fall symposium on capturing and using patterns for evidence detection* (pp. 45–53). American Association for Artificial Intelligence.
Gallagher, S., & Allen, M. (2016). Active inference, enactivism and the hermeneutics of social cognition. *Synthese*. doi:10.1007/s11229-016-1269-8.
Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, *193*(2), 559–582.
Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: Causally relevant and different from detectors. *Biology and Philosophy*. doi:10.1007/s10539-017-9562-6.
Godfrey-Smith, P. (1996). *Complexity and the function of mind in nature*. Cambridge: Cambridge University Press.
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 27, pp. 2672–2680).
Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, *27*, 377–442.
Harman, G. (1973). *Thought*. Princeton: Princeton University Press.
Harman, G. (1999). *Reasoning, meaning and mind*. Oxford: Oxford University Press.
Harrison, L. M., Stephan, K. E., Rees, G., & Friston, K. J. (2007). Extra-classical receptive field effects measured in striate cortex with fMRI. *NeuroImage*, *34*(3), 1199–1208.
Hinton, G. E. (2005). What kind of graphical model is the brain? In *International joint conference on artificial intelligence 2005, Edinburgh*.
Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, *11*(10), 428–434.
Hinton, G. E., & Sejnowski, T. J. (1983). Optimal perceptual inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
Hinton, G. E., & Sejnowski, T. J. (1999). Unsupervised learning: Foundations of neural computation. In G. E. Hinton & T. J. Sejnowski (Eds.), *Unsupervised learning: Foundations of neural computation*. Cambridge, MA: MIT Press.
Hohwy, J. (2004). Top–down and bottom–up in delusion formation. *Philosophy, Psychiatry and Psychology*, *11*(1), 65–70.
Hohwy, J. (2011). Phenomenal variability and introspective reliability. *Mind & Language*, *26*(3), 261–286.
Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.
Hohwy, J. (2015a). The neural organ explains the mind. In T. Metzinger & J. M. Windt (eds.) *Open MIND* (pp. 1–23). Frankfurt am Main: MIND Group.
Hohwy, J. (2015b). Prediction error minimization, mental and developmental disorder, and statistical theories of consciousness. In R. Gennaro (Ed.), *Disturbed consciousness: New essays on psychopathology and theories of consciousness* (pp. 293–324). Cambridge, MA: MIT Press.
Hohwy, J. (2016a). Prediction, agency, and body ownership. In A. Engel, K. Friston, & D. Kragic (Eds.), *Where is the action? The pragmatic turn in cognitive science*. Cambridge, MA: MIT Press.
Hohwy, J. (2016b). The self-evidencing brain. *Noûs*, *50*(2), 259–285.
Hohwy, J. (2017). Priors in perception: Top-down modulation, Bayesian perceptual learning rate, and prediction error minimization. *Consciousness and Cognition*, *47*, 75–85.
Hohwy, J., & Michael, J. (2017). Why would any body have a self. In F. Vignemont & A. Alsmith (Eds.), *The body and the self*. Cambridge, MA: MIT Press.
Hosoya, T., Baccus, S. A., & Meister, M. (2005). Dynamic predictive coding by the retina. *Nature*, *436*(7047), 71.
Hutto, D. (2017). Getting into the great guessing game: Bootstrap heaven or hell? *Synthese*. doi:10.1007/s11229-017-1385-0.
Hutto, D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Cambridge, MA: MIT Press.

Kirchhoff, M. (2016). Autopoiesis, free energy, and the life–mind continuity thesis. *Synthese*. doi:10.1007/s11229-016-1100-6.
Klein, C. (2016). What do predictive coders want? *Synthese*. doi:10.1007/s11229-016-1250-6.
Le, Q. V., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G. S., et al. (2012). Building high-level features using large scale unsupervised learning. In *Proceedings of the 29th international conference on machine learning, Edinburgh*.
Limanowski, J., & Blankenburg, F. (2013). Minimal self-models and the free energy principle. *Frontiers in Human Neuroscience*, *7*, 1–12.
Loughlin, V. (2016). Jakob hohwy: The predictive mind. *Phenomenology and the Cognitive Sciences*. doi:10.1007/s11097-016-9479-6.
Lupyan, G. (2015). Cognitive penetrability of perception in the age of prediction: Predictive systems are penetrable systems. *Review of Philosophy and Psychology*, *6*(4), 547–569.
Macpherson, F. (2017). The relationship between cognitive penetration and predictive coding. *Consciousness and Cognition*, *47*, 6–16.
Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., et al. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, *8*, 825.
Metzinger, T., & Wiese, W. (Eds.). (2017). *Philosophy and predictive processing*. MIND Group: Frankfurt am Main.
Millikan, R. (1984). *Language, thought, and other biological categories*. Cambridge, MA: MIT Press.
Millikan, R. (1989). Biosemantics. *The Journal of Philosophy*, *86*(6), 281–291.
Neal, R. M., & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. *NATO ASI Series D Behavioural and Social Sciences*, *89*, 355–370.
O'Brien, G., & Opie, J. (2004). Notes toward a structuralist theory of mental representation. In H. Clapin, P. Staines, & P. Slezak (Eds.), *Represenation in mind: New approaches to mental representation*. Oxford: Clarendon Press.
Orlandi, N. (2014). *The innocent eye: Why vision is not a cognitive process*. Oxford: Oxford University Press.
Orlandi, N. (2016). Bayesian perception as ecological perception. *Philosophical Topics*, *44*(2), 327–351.
Palmer, C. J., Lawson, R. P., & Hohwy, J. (2017). Bayesian approaches to autism: Towards volatility, action, and behavior. *Psychological Bulletin*, *143*(5), 521–542.
Papineau, D. (1984). Representation and explanation. *Philosophy of Science*, *51*(4), 550–572.
Rao, R., & Ballard, D. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*, 79–87.
Sellars, W. (1953). Inference and meaning. *Mind*, *62*(247), 313–338.
Sellars, W. (2007). *In the space of reasons*. Cambridge: Harvard University Press.
Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, *17*(11), 565–573.
Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience*, *5*(2), 97–118.
Seth, A. K., Suzuki, K., & Critchley, H. D. (2012). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology*, *2*, 1–16.
Shea, N. (2014). Exploitable isomorphism and structural representation. *Proceedings of the Aristotelian Society Supplementary*, *114*(2), 123–144.
Sims, A. (2016). A problem of scope for the free energy principle as a theory of cognition. *Philosophical Psychology*, *29*(7), 967–980.
Trappenberg, T. (2010). *Fundamentals of computational neuroscience*. Oxford: Oxford University Press.
Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., Lee de-Wit, L., et al. (2014). Precise minds in uncertain worlds: Predictive coding in autism. *Psychological Review*, *121*(4), 649–675.
Vetter, P., & Newen, A. (2014). Varieties of cognitive penetration in visual perception. *Consciousness and Cognition*, *27*, 62–75.