Minds and Machines (2021) 31:277–303
https://doi.org/10.1007/s11023-021-09559-6

Are Generative Models Structural Representations?

Marco Facchin[1]

Received: 1 August 2020 / Accepted: 24 March 2021 / Published online: 30 March 2021

Abstract
Philosophers interested in the theoretical consequences of predictive processing often assume that predictive processing is an inferentialist and representationalist theory of cognition. More specifically, they assume that predictive processing revolves around approximated Bayesian inferences drawn by inverting a generative model. Generative models, in turn, are said to be structural representations: representational vehicles that represent their targets by being structurally similar to them. Here, I challenge this assumption, claiming that, at present, it lacks an adequate justification. I examine the only argument offered to establish that generative models are structural representations, and argue that it does not substantiate the desired conclusion. Having so done, I consider a number of alternative arguments aimed at showing that the relevant structural similarity obtains, and argue that all these arguments are unconvincing for a variety of reasons. I then conclude the paper by briefly highlighting three themes that might be relevant for further investigation on the matter.

Keywords Structural representations · Antirepresentationalism · Predictive processing · Representation wars · Generative models

1 Introduction

Predictive processing is a neurocomputational framework surrounded by a number of philosophical disputes. Some of these disputes concern foundational matters: is realism about the theoretical posits of predictive processing warranted (e.g. Colombo et al. 2018)? And, if yes, are these posits representations (e.g. Kirchhoff and Roberston 2018)? Other controversies concern what predictive processing (henceforth PP) entails: does PP support an internalist or externalist view of the mind (Clark, 2017; Hohwy, 2016)? Does it provide a complete account of cognition, or is it unable to

✉ Marco Facchin
marco.facchin@iusspavia.it

[1] Department of Human and Life Sciences, Istituto Universitario Di Studi Superiori IUSS Pavia, Palazzo del Broletto, Piazza della Vittoria n. 15, 27100 Pavia, Italy

account for the systematic nature of human thought and the curiosity so many intelligent animals blatantly manifest (Sims, 2017; Williams, 2018a)? Is consciousness really just the brain's best guess (Dolega & Dewhurst, 2020; Hohwy, 2013)?

In this latter kind of disputes it is often assumed that PP is an inferentialist and representationalist theory of cognition. Rendered technically, the assumption is that PP revolves around approximated Bayesian inferences drawn by inverting a generative model operating under a predictive coding message-passing scheme (Clark, 2013; Hohwy, 2013). More mundanely, it is assumed that PP revolves around statistical inferences performed leveraging probabilistic models of the world. Models, in turn, are understood as *structural representations*: vehicles that represent their targets by mirroring the targets' relational structure (e.g. Williams, 2017, 2018b). The tie between predictive processing and models runs so deep that some have suggested that predictive processing would not be *humanly intelligible* without them (Clark, 2015).

Here, I challenge this assumption. I claim that the *only*[1] argument offered to identify generative models with structural representations (Gładziejewski, 2016) is flawed, and that it cannot be easily ameliorated. By doing so, I hope to bring a small contribution to the disputes surrounding the philosophical foundations of PP.

The essay is structured as follows. In Sect. 2, I introduce the theoretical apparatus of PP. In Sect. 3, I introduce structural representations, and summarize the argument Gładziejewski offers to identify generative models with them. In Sect. 4, I turn from exposition to criticism, showing a flaw in Gładziejewski's argument, and claiming that it cannot be easily adjusted. A brief concluding paragraph follows.

2 A Brief Introduction to Predictive Processing

Here, I provide a short introduction to PP. Since PP is now largely known, I will sketch only its most fundamental aspects.[2]

To successfully orchestrate behavior, an agent's brain must first determine in which environmental situation the agent is embedded; that is, what are the environmental causes of the energies impacting the agent's transducers. PP assumes that this task is burdened with uncertainty, as sensory states are under-determined in respect to their causes (Friston, 2005).[3] Depending on the context, different causes might generate similar sensory states, just as a single object can generate an unruly manifold of different inputs.

To cope with this uncertainty, PP suggests that the brain resorts to a form of Bayesian inference, as it yields an optimal way to determine the *most likely* cause

[1] A reader might contest this, noting that numerous accounts of generative models as structural representations have been offered (e.g. Kiefer and Hohwy 2018, 2019; Wiese 2018). I am aware of the existence of such accounts. However, to me they all seem to presuppose the success of Gładziejewski's (2016) original argument, to then improve on it in various ways.

[2] See (Clark 2013; 2016; Hohwy 2013; Tani 2016) for more introductory material.

[3] Notice that this is a theoretical assumption, that can be theoretically contested (e.g. Orlandi, 2016).

of a sensory state, given the incoming input and some prior knowledge of how environmental cause generate sensory states (Yuille and Kernsten 2006; Hohwy, 2013, pp. 13–40). Importantly, since exact Bayesian inferences are often computationally intractable, PP suggests that the brain *approximates* their results by inverting a generative model operating under a predictive coding processing regime.[4]

Generative models are data structures capturing the relations holding between some observable data (here, sensory states) and their hidden causes (here, worldly objects). These models are said to be generative, as the knowledge they embody can be leveraged "from the top-down" to *generate* expected instances of data (Hinton, 2007a; Danks, 2014, p. 44). Since real sensory states are generated by the nested interaction of multiple causes operating at different spatiotemporal scales, generative models need to be *hierarchically organized* to capture these nested causal relations. Importantly, this only requires that each hierarchical level l_N treats the hierarchically lower level l_{N-1} as a data source, capturing the regularities it displays (see Hinton, 2007b). Lastly, these models must be *probabilistic*, embodying their causal knowledge formatted in terms of the probability density functions that are required to approximate Bayesian inferences (e.g. Knill & Pouget, 2004).

Predictive coding is a message passing scheme which deploys generative models as follows (Huang & Rao, 2011; Rao & Ballard, 1999). With the exception of the bottommost level, each level l_N of the model generates a prediction signal: an "expected" pattern of activity of l_{N-1}, which is then conveyed to l_{N-1} through a set of descending connections. Hence, collectively, the levels of the model will generate a "downstream" flow of progressively spatiotemporally refined predictions about the incoming sensory inputs, ideally flowing from "higher" associative areas to primary sensory cortices (see Mesulam, 2008). As this signal is received, each level contrasts it with its own actual activity (or the incoming input in the case of the bottommost level), and computes the mismatch between the two. The magnitude of the mismatch, known as prediction error, is then conveyed "upwards", from l_N to l_{N+1}, courtesy of a second set of *ascending* connections. As prediction error is received, each level changes the prediction signal conveyed downwards so as to minimize the incoming prediction error. This process is then iterated until the entire hierarchy reaches a global minimum of prediction error. Since the states of the generative model that best minimize the error correspond to the *most likely* causes of the incoming sensory signal (given the body of knowledge the generative model encodes), minimizing prediction error *inverts* the generative model, mapping the signal onto its most likely causes, implicitly realizing a form of Bayesian inference (Clark, 2013, 2016; Hohwy, 2013, 2019; Kiefer & Hohwy, 2019).

This crude sketch calls for significant amendments. Firstly, it is silent upon *lateral* connections, which allow different, explicitly coded, hypotheses to compete in

[4] Importantly, model inversion is not *essentially* an approximated process. So, by saying that a generative model is inverted one has not yet shown *how* the intractability problem is solved. Since the technical details are fairly complex (see Bogacz, 2017) and will not matter for my argument, I will not sketch them here. An anonymous referee has my gratitude for having noticed this issue.

the interpretation of incoming data (Friston, 2005).[5] It is also silent on the *expected precision* of the incoming signals, which constantly modulates the message passing, determining the "impact" of error signals (Hohwy, 2013, pp. 59–74; Clark, 2016, pp. 53–82). Yet, the most significant amendment this sketch needs is the following: this mechanism is not *just* a mechanism of perception. It is also the engine of action (Hohwy, 2013, pp. 75–96; Friston, 2013a; Clark, 2016, pp. 111–137). To see how, consider the following two points.

First: prediction error minimization can occur under *two* directions of fit (Shea, 2013). One can change the predictions to make them fit the input, as sketched above. But one can also "keep the predictions still", and force *the input* to fit them. Secondly, given that the agent's body is, just like the external world, a source of sensory signals (and given that these two sources interact: moving towards an object will change the stimuli the object generates), the generative model must also model the agent's body (Hohwy, 2015). Mashing these two observations together yields the gist of how PP accounts for action: agents act by predicting specific bodily signals, to then cancel out the error relative to these predictions through movement (Adams, 2013). Actions appear thus to be generated by self-fulfilling predictions.

PP thus casts action and perception as two complementary sides of the same computational process of prediction error minimization. Given that processes of error minimization are inferential processes, as demonstrated by the brief analysis of the account of perception PP offers, this means that action is an inferential process too.[6] This is why, in the PP literature, action is referred to as *active inference*. More specifically, an agent engaged in active inference *tests* its model of the world, seeking sensory evidence to confirm the predictions licensed by that model (Hohwy, 2015, 2016, 2017, 2018).

Importantly, however,*active inference* has a broader scope than action as usually understood (i.e. bodily movements fulfilling one's intentions). Nothing *obliges* the self-confirming expectations involved in active inference to be *proprioceptive* expectations, to be confirmed by bodily movements. They might be visual (i.e. exteroceptive) predictions, and elicit saccades (Friston et al. 2010). Or they might be interoceptive predictions, servicing an agent's homeostatic control (Seth, 2015) and emotional regulation (Seth & Friston, 2016). From this perspective, prediction error minimization is not *primarily* a tool for accurate perception and goal-directed action. Rather, it appears as a mean to the more fundamental end of maintaining an agent within its physiological bounds of viability. This line of reasoning connects PP to an ambitious framework in theoretical biology, namely the free energy principle (see Friston, 2013b, 2019; Allen & Friston, 2018). But the free energy principle will not be considered here,[7] and the sketch of PP just proposed seems a sufficient introduction, given the task at hand.

[5] Many thanks to the anonymous reviewer who noticed that the original formulation of this point was too strong.

[6] And in fact, according to PP, action too requires the inversion of a generative model (see Friston 2011).

[7] This is not *entirely* correct: some aspects of the free energy principle, namely the ones most related to neuroscience, will be considered here. But since these aspects tend to boil down to PP (see Friston, 2009; 2010), I do not think I need to explicitly discuss the free energy principle here.

In the next section, I briefly introduce structural representations and then summarize Gładziejewski's (2016) argument to identify generative models with them.

3 Structural Representations and Predictive Processing

3.1 Structural Representations

As hinted at in the introduction, structural representations are representations whose *vehicles* represent their targets *by mimicking the inner relational structure* of the targets. Consider, for instance, a cartographic map. It might depict a gulf being north of an isle by placing a certain element, corresponding to the gulf, above a second element, corresponding to the isle. Importantly, the PP literature on structural representations (e.g. Dolega, 2017; Hohwy, 2020; Kiefer & Hohwy, 2018, 2019; Ramstead et al. 2019; Wiese, 2017, 2018; Williams, 2017) points to a single formalized account of structural representations; namely Gładziejewski's (2015, 2016) account. According to Gładziejewski:

A state R of a system S is a structural representation of a target T *only if*:

(a) R is structurally similar to T; &
(b) R guides S's action aimed at T; &
(c) R can satisfy (b) when decoupled from T; &
(d) S can detect the representational error of R

Each point calls for clarification.

Point (a) clarifies that structural representations are iconic: their representational properties are (at least partially) grounded in the *similarity* holding between them and their targets. Yet notice that the relevant kind of similarity mentioned in (a) is *structural* similarity. The relevant[8] definition of structural similarity is provided in (O'Brien and Opie 2004, p. 11):

> Suppose $S_V=(V,\mathfrak{R}_V)$ is a system comprising of a set V of objects, and a set $\mathfrak{R}_V$ of relations defined on the members of V. The objects in V may be conceptual or concrete; the relations in $\mathfrak{R}_V$ may be spatial, causal, structural, inferential, and so on. […] We will say that there is a *second-order resemblance* between two systems $S_V=(V,\mathfrak{R}_V)$ and $S_O=(O,\mathfrak{R}_O)$ if, for at least *some* objects in V and *some* relations in $\mathfrak{R}_V$, there is a one-to-one mapping from V to O and a one-to-

[8] Here, "relevant" means "the one adopted by Gładziejewski". Other definitions of structural similarity are surely possible (e.g. Shea, 2018, p. 117). However, since my focus here is Gładziejewski's argument, I will stick to the definition Gładziejewski favors.

> one mapping from $\mathfrak{R}_V$ to $\mathfrak{R}_O$, such that when a relation $\mathfrak{R}_V$ holds of objects in V, the corresponding relation $\mathfrak{R}_O$ holds of the corresponding objects in O.

There are several important things to highlight about this definition. One is that it can be straightforwardly applied to point (a) assuming that R is S_V and T is S_O. Another is that structural resemblance does not require *first order* resemblance to obtain. In fact, nothing in the definition requires S_V and S_O to have any common property. All they need to share is *a common pattern of relations* among their elements. Thirdly, the definition of structural similarity is tripartite. For S_V to be structurally similar to $S_{O,}$ it must be the case that: (i) at least *some* of the *objects* of which S_V and S_O are constituted map one-to-one onto each other; and (ii) at least *some* of the *relations* holding among these objects map one-to-one onto each other; and (iii) *corresponding* objects stand in *corresponding* relations in both S_V and S_O. Notice that (i) to (iii) need to obtain in conjunction. Notice, lastly, that (i) to (iii) obtaining in conjunction entails that S_V is *semantically unambiguous* in respect to S_O. By this I mean that once the mapping rule is known, it is *always in principle possible* to determine, for all elements of S_V mapping onto $S_{O,}$ to which element of S_O each element of S_V corresponds.

According to this definition, a structural similarity might hold among *any* two systems. However, the relevant structural similarity exhibited to satisfy point (a) must hold between a *representational vehicle* and the represented target (Kiefer & Hohwy, 2018; O'Brien, 2015). This is entailed by the *definition* of a structural representation: a representational *vehicle* that represents a target by being structurally similar to it.[9] It is thus immediately clear that the relevant structural similarity holds between *the vehicle* (the concrete particular doing the representing) and the represented target. Therefore, R must be a representational vehicle: a concrete particular encoding content.

Point (b) establishes that structural representations are *causally responsible* for S's behavior (Gładziejewski & Miłkowski, 2017). Gładziejewski unpacks point (b) in terms of *exploitable* structural similarity (Gładziejewski, 2015, 2016; Gładziejewski & Miłkowski, 2017). Hence, the relevant structural similarity in (a) must be exploitable[10] (Shea, 2014, 2018, p. 120). Put simply:

[9] Alternatively, structural representations can be defined as: "A collection of representations in which a relation on representational vehicles represents a relation on the entities they represent" (Shea, 2018, p. 118). This definition stresses the important fact that each element of the structural representation is also a representational vehicle, whose content is determined by the relevant structural similarity in which it participates. For instance, each object on a map stands for (i.e. represents) an environmental landmark, and spatial relations among objects on a map represent spatial relations holding among the corresponding landmarks. Notice that such a nesting of representational vehicles is entirely unproblematic: after all, both a sentence and the words forming it are representational vehicles in an entirely intelligible sense. Notice further that according to both Shea's and Gładziejewski's definition, the relevant structural representation is the *entire structure* of related objects, rather than any single part of that structure. That is, the elements (V and $\mathfrak{R}_V$) of a structural representation need not be, on their own, structural representations.

[10] This might or might not require a representational consumer. Gładziejewski asserts that a consumer is necessary in his (2015); but his (2016) does not mention consumers. Shea's definition of exploitable structural similarity (to which Gładziejewski adheres) does not require consumers, so I will skip them here. Notice that I adapted the notation in Shea's definition for the sake of orthographic consistency.

R's structural similarity with T is exploitable by S *only if*:

(iv) $\mathfrak{R}_V$ is a set of relations S's downstream computational processing is systematically sensitive to; &
(v) $\mathfrak{R}_O$ and O are of significance to S

Condition (iv) requires S to be sensitive, in its downstream processing, to the relevant relations $\mathfrak{R}_V$ in virtue of which R structurally resembles T. Given that computational processing is a mechanical affair, this requires the relations in $\mathfrak{R}_V$, or at least the objects in V upon which $\mathfrak{R}_V$ is defined, to causally impact the processing of S in some systematic way. Condition (v) requires R to structurally resemble a target that matters to S's purposes; that is, a target which matters to S's computational functioning.

Importantly, *exploitable* structural similarity is not a reflexive and symmetric relation (see Shea, 2014; Williams & Collings, 2017). For this reason, structural representations are immune to the objections that were fatal to iconic representations defined in terms of first order resemblance (see Goodman, 1969, p. 3–4). Notice further that insofar as *exploitable* structural similarity determines representational content, R's content is not just causally efficacious (Gładziejewski & Miłkowski, 2017): it is also *intrinsic* to R's material constitution as a representational vehicle (O''Brien & Opie, 2001; O'Brien, 2015; Lee, 2018), as it is literally inscribed in the physical form of R.

Condition (c) captures the idea that genuine representations function as *stand-ins* for their targets, enabling S to perform processes aimed at these targets even when they are absent (Grush, 1997; Pezzulo, 2008; Webb, 2006). Gładziejewski unpacks this idea in terms of decouplability, defining it as follows (Gładziejewski, 2015): R is *weakly decoupled* from T only if R and T are in no causal contact; R is instead *strongly decoupled* from T only if S and T are in no causal contact.[11]

Lastly, (d) captures the idea that representations can be semantically evaluated by the systems leveraging them (see Bickhard, 1999). Importantly, representations are often semantically evaluated only *indirectly*, by assessing how successful they are in guiding action (Gładziejewski, 2015 pp. 78–79; 2016, p. 569). This indirect route of evaluation seems a natural outcome of exploitability. If, as (iv) entails, R's content determines downstream processing in S (and eventually S's behavior), then *successful* behaviors directly depend on R's content being *correct* (accurate and/or truthful). Pragmatic successes and failure thus appear as reliable indicators of the semantic status of a system's representational resources.

Before moving forward, it is important to clarify the scope of Gładziejewski's account of structural representations. Following Chemero (2009: pp. 67–68), it is possible to distinguish between an *epistemic* representationalist claim and a *metaphysical* representationalist claim. Bluntly put, the epistemic representationalist

[11] In its original formulation, the definition of decouplability also mentions representational consumers (see Gładziejewski, 2015). Here, I omit them for the reasons given in the previous footnote.

claim is the claim that our *best explanations* of cognition need to posit representations. The metaphysical representationalist claim is instead the claim that cognitive systems contain components that *really* are representations. The two claims can in principle come apart. A fictionalist about representations, for instance, endorses the epistemic claim while denying the metaphysical one (Sprevak, 2013; Ramsey, 2020; see also Downey, 2018 for a fictionalist interpretation of PP). Gładziejewski's account of structural representations aims at vindicating both claims (Gładziejewski, 2015: 70).[12] Thus, his account of structural representations succeeds just in case the relevant representational posits of PP (i.e. generative models) satisfy features (a) to (d) *and* these are the relevant sort of structures identified as representation by our best explanatory practices.

3.2 Generative Models as Structural Representations.

Gładziejewski (2016) holds that his account of structural representations straightforwardly applies to generative models. The general outlook of his argument is as follows:

(P1) Items satisfying conditions (a) to (d) in conjunction are structural representations
(P2) Generative models satisfy conditions (a) to (d) in conjunction; therefore
(C) Generative models are structural representations.

The argument needs little clarification. (P1) follows directly from the definition of structural representations. Thus, (P2) carries alone the whole weight of the argument. Here, I sketch the reasoning Gładziejewski provides to substantiate (P2). I will focus in particular on point (a), as it will be central in the following discussion.

To claim that generative models satisfy condition (a), Gładziejewski (2016, pp. 571–573) reasons as follows. Generative models can formally be treated as Bayesian nets. Bayesian nets are directed acyclic graphs, and thus are *graphs*: sets of nodes connected by edges. And, in general, graphs are *always* structurally similar to their targets, because: (i) each node stands for one, and only one, environmental variable, (ii) the edges connecting the nodes map onto some relation holding among environmental variables; and (iii) two nodes are connected by an edge if, and only if, the corresponding variables are in the relation of interest (see Danks, 2014, pp. 39–41). Notice that graphs are not *just* structurally similar to their targets: they are *homomorphic* to them; and homomorphism is a special, *stronger*, case of structural similarity (see O''Brien and Opie, 2004, p. 12).

Gładziejewski (2016, pp. 571–573) also suggests a specific way in which generative models structurally resemble their targets: the nodes of each hierarchical layer l_N in the model map onto the likelihoods of the corresponding sensory states. Interlevel connections between nodes mimic the temporal evolution of sensory states. Lastly, each

[12] This commitment seems shared by the majority of accounts of generative models as structural representations (e.g. Kiefer and Hohwy, 2018; 2019; Wiese, 2018; Williams, 2017).

node maps onto the prior probability of the corresponding worldly cause. The validity of the mapping on offer might be contested (e.g. Wiese, 2017), but the general point Gładziejewski is trying to make is clear: generative models are graphs and, *as such*, they are structurally similar to their targets. Therefore, they easily satisfy condition (a).

Point (b) is satisfied because generative models can engage in active inference. Recall: during active inference, the model "purposefully" generates a false prediction, to then cancel the prediction error it generates through movement. Thus, generative models are the engines of action, and guide behavior exactly as (b) requires.

To show that generative models satisfy condition (c), Gładziejewski exhibits the following evidence. To begin with, prediction error minimization seems responsible for a variety of paradigmatically *offline* cognitive phenomena, such as memory and imagination (e.g. Clark, 2016, pp. 84–107). Furthermore, early sensory cortices are involved in acts of imagination (e.g. Albers, 2013). Given that these cortices are commonly taken to be part of the neural machinery implementing our generative model, these studies suggest that generative models can in fact function when decoupled from their targets. Moreover, generative models must be *counterfactually deep* (Seth, 2014): to function effectively, they need to be able to predict how the sensory signals *would* change, *were* the agent to move in such-and-such a way. Given that counterfactual scenarios have no causal powers, our generative models will always encode some information about targets from which we are strongly decoupled.

Lastly, generative models clearly satisfy condition (d). When a hypothesis about the incoming sensory flow is selected and tested in active inference, the *prediction error* that the hypothesis generated is a measure of its inaccuracy. For instance, when I drink coffee, I (subpersonally) expect a flow of gustatory sensations which I try to bring about through the ingestion of coffee. But if the ingestion of the liquid does not bring about *these* sensations, the ensuing error will inform me that I'm in fact *not* drinking a coffee, prompting my generative model to revise its expectations. In this way, the error ensuing from failed active inferences indicates the representational inaccuracy of the tested hypothesis.

This concludes the exposition of Gładziejewski's argument. In the next section, I will attack it, claiming that it does not succeed in equating generative models to structural representations.

4 Generative Models and Structural Representations: An Unjustified Identification

I divide this section in two sub-sections. The first section examines Gładziejewski's argument for (a), and argues that the argument Gładziejewski offers fails in vindicating (a). The second section examines some alternative arguments for (a).

4.1 A Problem in Gładziejewski's Argument for Point (a)

Recall Gładziejewski's (2016) argument for (a). The argument is as follows: generative models can be thought of as Bayesian graphs, Bayesian graphs (and graphs

in general) are structurally similar to their targets; therefore, generative models are structurally similar to their targets. *Therefore, (a) obtains*.

Yet, I believe that this line of reasoning cannot vindicate (a). This is because the relevant structural similarity exhibited to vindicate (a) must hold between a *representational vehicle* and a represented target. But graphical models *are not* representational vehicles. So the argument Gładziejewski provides does not substantiate (a).[13] Let me unpack.

If one is a realist about representations (as Gładziejewski surely is, see Gładziejewski, 2015, 2016; Gładziejewski & Miłkowski, 2017), then one is committed to the claim that representations are *concrete particulars* encoding content (e.g. Shea, 2018, pp. 25–43). Accordingly, structural representations are defined as *concrete particulars* (i.e. representational vehicles) which carry content in virtue of a relevant exploitable structural similarity that holds between them and their representational targets. They are *representational vehicles* that do the representing by structurally resembling. Hence, the relevant structural similarity must hold between a representational vehicle (a concrete particular) and a represented target. This is why the content of structural representations is supposed to be *intrinsic* to their material constitution (e.g. O'Brien & Opie, 2001; O'Brien, 2015; Williams & Colling, 2017; Lee, 2018). Their content is intrinsic because it is inscribed in the physical form of the representational vehicle (the concrete particular that does the representing).

But Bayesian nets, and graphs in general, are *mathematical objects* (e.g. Danks, 2014, p. 40; Leitgeb, 2020). They are defined as a finite *set* of *nodes* connected by a finite *set* of *edges* (Koski & Noble, 2009, p. 41) and sets, nodes and edges are mathematical objects. Mathematical objects might or might not be particulars (it is irrelevant for the purpose of the argument), but definitely are not *concrete*. So they cannot be representational vehicles, given that representational vehicles are *concrete* particulars. Hence, the structural similarity Gładziejewski exhibits cannot be used to vindicate point (a). It just isn't what point (a) requires.

The same issue can also be appreciated from another point of view. Thus, consider condition (b). Recall that, in order to satisfy (b), the relevant structural similarity *that satisfies (a)* must be exploitable. Exploitability is partially defined by condition (iv): $\mathfrak{R}_V$ is a relation S's downstream computational processing is systematically sensitive to. But, as clarified above, for a computation to be sensitive to $\mathfrak{R}_V$, either $\mathfrak{R}_V$ or the objects V upon which $\mathfrak{R}_V$ is defined, must possess the relevant causal powers needed to systematically orient S's processing. Computational processing is, at the end of day, a *physical* affair, which is ultimately responsive only to the *physical* properties of the computational states (e.g. Williams & Collings, 2017, p. 1945). However, the only structural similarity Gładziejewski provides to vindicate (a) is defined over mathematical objects. But mathematical objects do not seem to have the causal powers needed to systematically influence S's downstream

[13] Notice that I'm not claiming that graphical models are not structurally similar to their targets. They are. As clarified above, a structural similarity might hold among any pair of entities. Yet, the relevant class of structural similarities that can be used to vindicate (a) is the class of structural similarities holding between representational vehicles and their targets; and graphs are not representational vehicles.

processing. Hence, the structural similarity that Gładziejewski exhibits to satisfy (a) cannot be exploitable.

This places Gładziejewski's line of argument in a dire situation: either the structural similarity he shows does not vindicate (a), or it does. But if it does, then (b) fails to obtain. Either way, the argument seems to fail, leaving the claim that generative models are structural representations unjustified.

As I understand it, the overall flaw in Gładziejewski's argument is the following. Structural representations are defined in terms of their vehicle properties. Hence they should be identified at the level of the *physical machinery* doing the computation (what Marr would call the implementation level). Structural representations are bits of an information processing system *literally* resembling their target according to an appropriate mapping rule. But graphs, wherever they sit in the explanatory hierarchy of cognitive science, surely do not sit at the level of the *physical machinery* doing the computing (Danks, 2014, pp. 13–37; 218–221). Hence, Gładziejewski's argument seems to be pitched at the wrong level of the relevant explanatory hierarchy. For this reason, I conclude that Gładziejewski's argument does *not* justify the claim that generative models are structural representations.

But there might be other means to justify that claim.

4.2 Alternative Arguments for (a)

4.2.1 Alternative Argument 1

The claim that graph theoretic notions cannot sit at the level of the physical machinery might not be warranted. For instance, graph theoretic notions are used to map the connections between different neural regions (e.g. Sporns, 2010). Thus, graph theoretic notions *can* sit at the level of the physical machinery. This suggests an alternative way to vindicate (a): if graphs are structurally similar to their targets, and these graphs can be transparently mapped onto cortical structures and/or neural activity patterns, then, since structural similarity is a *transitive* relation, the relevant cortical structures/activity patterns are structurally similar to the target of the graphs. Some (e.g. Kiefer, 2017) defend structural representations along precisely these lines.[14]

This alternative argument for (a) is fairly attractive, as it is maximally conservative over the structure of Gładziejewski's original argument. It also nicely integrates with the existing *scientific* literature on PP, at least insofar some generative models, rendered as Bayesian nets, have been mapped onto cortical structures (e.g. Bastos,

[14] Or, at least, so it *seems*. To be honest, I believe that Kiefer is no longer committed to the claim that generative models are structural representations. Rather, it seems to me that Kiefer is committed to some form of functional role semantics. To be precise, Kiefer (2017, p. 12) seems to endorse the claim that generative models are structural representations. However, he seems to have quickly changed his mind about this point, as, in numerous later publications (Kiefer and Hohwy, 2018, p. 2393; 2019, p. 401–403; Kiefer, 2020, footnote 19) he takes the content of generative models to be determined by internal functional roles rather than by the structural similarity holding between a generative model and its target. I will more directly confront this issue in the main text, when dealing with the fourth alternative argument for (a). Many thanks to an anonymous referee for having pressed me on this issue.

2012; Friston, 2017a). Isn't this *sufficient* to show that at least these generative models are structurally similar to their targets?

I believe that a negative answer is warranted. This is because the graphs presented in (Bastos et al. 2012; Friston et al. 2017a) and a number of similar publications in the PP literature do not model any worldly target. There is thus no specific worldly target that they represent. So, even if the cortical machinery is in some relevant sense structurally similar to these graphs, there is no *third* element to which the cortical machinery can be structurally similar to by being structurally similar to these graphs. For this reason, it seems to me correct to conclude that the alternative argument for (a) provided above fails.

But what, then, is the purpose of the graphs in (Bastos et al. 2012; Friston et al. 2017a)? The answer, if I understand the literature correctly, is the following: these graphs are, in a sense, didactic tools, aimed at showing, with a fair degree of approximation, that the cortical machinery is arranged in a way such that it can easily perform the inferential processes PP revolves around (see Bastos et al. 2012, p. 703; Friston et al 2017a, p. 393). In fact, it seems to me that, within the PP literature, graphical models are often deployed to capture the *message passing* within the brain; that is, how inference is performed (see, for instance, de Vries & Friston, 2017; Friston, 2017b; Friston, 2017c; Donnarumma, 2017; Matsumoto & Tani, 2020; see also Hinton & Sejnowski, 1983 for an earlier model).[15] I believe that this is an important point to notice for two distinct reasons.

First, if these graphical models are intended to be models of the relevant *message passing*, it seems more natural to suppose they will map onto the cortical machinery *performing* the inferences, rather than on the representational vehicles manipulated in inferential processes.[16] Secondly, and relatedly, if those graphical models are accurately characterized as portraying the inferential message passing in the brain, it seems to me that they *presuppose* the presence of some relevant representational vehicle, as inferences are defined over representations.[17] These representations might (but, as far as I can see, need not) be structural representations. However, as these graphical models seem to *presuppose* the presence of representations, it seems to me that they cannot be invoked to *justify* one's representationalist claim, on pain of circularity.

Importantly, I do not believe that the considerations offered above rule out in any way the *possibility* of using graphical models to justify (a). As far as I can see, one might still resort to a graphical model to argue that at least some representational vehicles in the brain are structurally similar to their targets using the argument by transitivity sketched above. However, to do so, one would need a graphical model

[15] Notice that the scope of my claim is restricted to PP and the usage of graphical models in the PP literature. I make no claim on how graphical models are used in the rest of cognitive neuroscience (and related disciplines). Many thanks to the reviewer who advised me to be more cautious on this point.

[16] Importantly, this seems exactly how Kiefer interpreted these models, see (Kiefer 2017, pp. 12–16).

[17] The same two points seems to apply whether these models are intended to capture computational processes more generally, given that computational processes are often defined in terms of representations (see Fodor, 1981; Shagrir, 2001; Ramsey, 2007, pp. 68–77; Sprevak, 2010; Rescorla, 2012). This latter point, however, is not entirely uncontested (e.g. Piccinini, 2008).

depicting some specific worldly target. And, to the best of my knowledge of the PP literature, no such graphical model has yet been proposed.

4.2.2 Alternative Argument 2

Artificial neural networks might provide a different way to leverage graph theoretic notions to vindicate (a). As formal objects, artificial neural networks *are* graphs. But they are also somewhat plausible sketches of the *physical machinery* implementing or realizing some given computational process of interest (see Haykin, 2009, pp. 1–18; Rogers & McClelland, 2014). Moreover, at least some artificial neural networks encoding generative models (such as Helmholtz machines) are Bayesian graphs (e.g. Dayan & Hinton, 1996). Therefore, even if these artificial neural networks cannot *prove* that generative models in the brain are structural representations, they can show that generative models can be structural representations, thereby providing circumstantial evidence in favor of (a) obtaining. If our plausible sketches of the physical machinery encoding generative models are graphs (or at least graph-like), then we have a solid reason to believe the *real* physical machinery encoding generative models is graph-like. And, given that graphs are structurally similar to their targets, we have a solid reason to believe that (a) obtains. However, I think such a belief would be misplaced. Indeed, it seems to me that a closer consideration of artificial neural networks provides a reason to believe that (a) *does not* obtain.

To see why, consider first that artificial neural networks are often said to encode generative models in their weighted connections (e.g. Dayan & Hinton, 1996; Hinton, 2014; Spratling, 2016, p. 3).[18] But weighted connections (or, more precisely, weight matrices) are typically considered to be *superposed* representations. And the vehicles of superposed representations are *not* structurally similar to their targets. As a consequence, if considering artificial neural networks provides evidence regarding the status of (a), the evidence they provide is *not* in favor of (a) obtaining.

To elaborate a little, consider first the notion of a *superposed* representation. A representation R is said to be a superposed representation of two targets T and T' when R encodes information about T and T' using the *same* set of physical resources. When applied to weight matrices, the idea is that weight matrices superpositionally represent their targets when each individual weight is assigned a value such that the network can exhibit the functionality needed to operate on all its targets (Clark, 1993: pp. 17–19, see Van Gelder, 1991, 1992 for further discussion). For instance, a single net can be first trained to recognize (or generate) instances of T. If the network is then trained so as to recognize (or generate) both instances of T and T', then the weights of the net will encode information about *both* representational targets, and the weight matrix will represent both in a superposed fashion.

However, in weight matrices: "Each memory trace is distributed over many different connections, and each connection participates in many different memory traces"

[18] A similar claim is sometimes made in the neuroscientific literature on PP (ad es. Friston, 2005, p. 820; Shipp, 2016, p. 3). The claim, however, might not be entirely correct, as I will soon clarify in the main text.

(McClelland & Rumelhart, 1986, p. 176). So, it seems each individual weight maps onto *many* different representational targets (or aspects thereof). But if this is the case, then either condition (i) or (ii)[19] of structural similarity are blatantly violated, since they require a one-to-one mapping. As a further proof of their violation, recall that the obtaining of (i) to (iii) in conjunction *entails* semantic unambiguity (see Sect. 3.1). That is, if (i) to (iii) jointly obtain, it is always *in principle possible* to tell, for each "bit" of the representational vehicle, which "bit" of the represented target it corresponds to. However, in superposed representations: "It is impossible to point to a particular place where the memory of a particular item is stored" (Rumelhart & McClelland, 1986, p. 70). Superposed representations are thus not semantically unambiguous. Therefore, *at least one* condition among (i) and (iii) is not met. As a consequence, superposed representations do not support the claim that (a) obtains.[20]

The argument outlined above can be challenged in two ways.[21] First, generative models are not encoded in connections alone; they are *jointly* encoded by connections *and activity vectors* (e.g. Buckley, 2017, p. 57). Secondly, the definition of structural similarity relevant to the obtaining of (a) quantifies only over *some*. Thus, noticing that connections do not participate in any one-to-one mapping does not provide an argument to the effect that (a) does not obtain: connections might simply be excluded from the objects V or relations $\mathfrak{R}_V$ of R participating in the structural similarity. I address these challenges in turn.

Are generative models really *jointly* encoded by activity vectors and weighted connections as the first challenge suggests? As far as I can see, the answer is positive; and focusing *only* on weighted connections (as I did above) is a mistake. But, to my excuse, it is a mistake the that the PP literature encourages:

> We allowed the network to learn a hierarchical internal model of its natural image inputs by maximizing the posterior probability of generating the observed data. *The internal model is encoded in a distributed manner within the synapses* of the model at each level. (Rao & Ballard, 1999, p. 80, emphasis added)
>
> The representation at any given level attempts to predict the representation at the level below; at the lowest level this amounts to a prediction of the raw

[19] Or both. The formulation in terms of "either (i) or (ii)" is due to the fact that it seems to me that one might interpret weighted connections either as parts of a structural representation (i.e. as members of V) or as *relations* among parts (i.e. as relations in $\mathfrak{R}_V$).

[20] Notice that I'm not denying that weight matrices encode the invariant relations that hold among the elements of the domain upon which the network has been trained to operate. I am only denying that there is a mapping from weight matrices (that is, from individual weights or sets of weights) to relations such that the mapping satisfies (i) or (ii). In simpler terms, I'm not denying that weight matrices represent invariant relations, I'm only denying that weight matrices represent invariant relations by being structurally similar to the target domain (or by participating in some relevant structural similarity with the target domain). Notice, importantly, that not all invariant relations need to be encoded in a vehicle that is structurally similar to its target. We might, for instance, stipulate that the sign "§" represents the fact that my father is *n* years older than me. If we do so, then "§" encodes an invariant relation holding between me and my father, and yet there just seems to be no structural similarity holding between "§" and the target it represents. Many thanks to an anonymous reviewer for having pressed me on this point.

[21] Many thanks to an anonymous referee for having raised these objections.

> sensory input. *It is the backward connections, therefore, that instantiate the generative model*. (Shipp, 2016, p. 3, emphasis added)
>
> The prediction error minimization (PEM) framework in cognitive science is an approach to cognition and perception centered on a simple idea: organisms represent the world by constantly predicting their own internal states. [...] Cascades of predictions are matched against the incoming sensory signals, which act as negative feedback *to correct a generative model encoded in the top-down and lateral connections*. (Kiefer & Hohwy, 2019, p. 384, emphasis added)
>
> The generative model, which in theories such as hierarchical predictive coding is hypothesized to be *implemented in top-down cortical connections*, specifies the *Umwelt* of the organism, the kinds of things and situations it believes in independently of the current sensory data [...] (Kiefer, 2020, p. 2, emphasis added)

Sadly, this excusation does not address the first challenge. Importantly, however, it seems to me that the first challenge is really no challenge at all. Allowing (so to speak) activity patterns to participate in the relevant structural similarity alongside weighted connections does not change the fact that, at least *prima facie*, weighted connections do not appear to map *one-to-one* onto any target. Thus, simply counting activity patterns as elements of V does not vindicate (a). This is because weighted connections are still considered elements of either V or $\mathfrak{R}_V$, and, as a result, at least some elements of the vehicle do not map one-to-one onto elements of the target as required by (a). Counting activity vectors in, on its own, is not enough: one must also be able to exclude that weighted connections participate in the relevant structural similarity.

This brings me to the second challenge. Is it possible to define some relevant vehicle target structural similarity without involving weighted connections? I think that the correct answer is negative.

To start, it is surely *possible* to define some relevant network-target structural similarity without appealing to weighted connections. There is nothing particularly new in this claim: Paul Churchland's state-space semantics is the most obvious example of a network-target structural similarity that does not involve connections (see Churchland, 1986, 2012; see also O'Brien and Opie, 2004). In his view, the entire *activation space* of the hidden layers of a network structurally resembles the target domain upon which the network has been trained to operate. And I'm fairly confident that a similar structural similarity can be found by considering artificial neural networks encoding generative models.[22]

Isn't this *just conceding* that (a) obtains? I do not think so. For activation spaces (the first relevant *relatum* of the structural similarity) are not vehicles, because they are not concrete particulars. They are abstract mathematical spaces that are used to account for the systematic behavior of artificial neural networks. So, they fail to vindicate (a) for the same reasons Gładziejewski's argument fails to vindicate (a).

[22] And even if my confidence were misplaced, I would concede the point for the sake of discussion.

Notice that I'm *not* claiming that the activation space-target domain structural similarity cannot determine the content of each individual activity vector. The relevant structural similarity holding between the activation space and the target *might* be enough to determine the content of each individual vector (i.e. of each individual element of V, using O'Brien and Opie's notation). However, the fact that each individual vector acquires content in virtue of the relevant structural similarity holding between the entire activation space and the target domain does not entail that each individual vector is a structural representation. This is because the elements of the vehicle (i.e. the objects of V and relations of $\mathfrak{R}_V$) involved in a structural similarity need not be structurally similar to elements of the target (i.e. the members of O and $\mathfrak{R}_O$) they correspond to. Given that, to my knowledge, only individual vectors are tokened in connectionist systems, the relevant structural similarity holding between the state space of a network and the network's target domain is insufficient to substantiate the claim that structural representations are tokened within the system.

Moreover, I honestly doubt that it is possible to *rightfully* exclude weighted connections from the relevant structural similarity. To see why, consider the following: if a vehicle represents in virtue of the structural similarity it bears to a target, then the more the vehicle and the target are structurally similar, the more the representation will be accurate. The accuracy of a structural representation non-accidentally increases when (and, at least *prima facie*, only when) the elements of the vehicle are rearranged in a way that increases the extent to which the vehicle is structurally similar to the target.

If this is correct, then there seems to be a solid reason to deny that we can exclude connections from the relevant network-target structural similarity, for modifications of weighted connections made in accordance to the learning algorithm *do improve* the representational accuracy of connectionist systems. Thus, if these systems represent by means of structural similarity, it seems that weighted connections *must* be counted among the elements participating in the similarity. Surely, the relevant definition of structural similarity provided when unpacking condition (a) quantifies only over some, but that "some" seems to include weighted connections. However, if my argument based on superpositionality is correct, then weighted connections do not map one-to-one on their targets as (a) requires. In short: if artificial neural networks deploys structural representations, connections must be involved. Yet, their involvement seems to prevent the obtaining of (a).[23]

4.2.3 Alternative argument 3

Alternative argument 3: maybe one does not need to look at artificial neural networks to vindicate the claim that generative models are structurally similar to their

[23] At this point, it might be tempting to wonder whether the relevant definition of structural similarity could be relaxed, so as to allow connections to be elements in the structural similarity in spite of the lack of any intelligible *one-to-one* mapping holding between them and the elements of the target domains. As an anonymous reviewer aptly noticed, O'Brien and Opie's (2004) definition of structural similarity is not the only one on the market, and at least some alternative formulations do not require a one-to-one mapping (e.g. Kiefer and Hohwy, 2019, p. 400; Shea, 2018, p. 117). As far as I can see, the mapping can be relaxed so as to allow *many* elements of the vehicle to map onto *one* element of the target. However, I believe the mapping *cannot* be relaxed so as to allow *one* element of the vehicle to

targets. In fact, PP theorists often point to a relevant structural similarity one might leverage to vindicate (a). Friston (2013a, p. 133) provides one clear example, worth quoting at length:

> [...] every aspect of our brain can be predicted from our environment. [...] A nice example is the anatomical division into *what* and *where* pathways in the visual cortex. Could this have been predicted from the free-energy principle? Yes – if the anatomical structure of the brain recapitulates the causal structure in the environment, then one would expect independent causes to be encoded in functionally segregated neuronal structures.

Since points (i) and (ii) in the definition of structural similarity quantify only over *some*, this quote by Friston provides us a structural similarity *sufficient* to vindicate (a): if Friston is right, there is a structure-preserving mapping form *some* cerebral regions onto *some* environmental targets. Furthermore, examples like the one highlighted in the quote seem fairly easy to multiply. It might be pointed out, for instance, that the anatomical segregation of visual and auditory cortices reflects the fact that visual and sensory input can have different worldly causes. So there *is*, I submit, a relevant brain-world structural similarity. Therefore, if *the whole brain* is the generative model (a claim that is not uncommon in the PP literature,[24] e.g. Bastos et al. 2012, p. 702), then condition (a) is met.

I must confess that, to me, this way of vindicating (a) seems to lead to a Pyrrhic victory at best. To begin with, claiming that the whole brain is a structural representation seems to prevent Gładziejewski's account from vindicating the epistemic

Footnote 23 (continued)
map onto *many* elements of the target. To see why this is the case, consider a minimal structural representation constituted by two objects $a^*$and b^* in a relation R^*. Suppose that R^* corresponds to a relation R, that a^* corresponds to an element a and that b^* maps onto two elements b and c. Now, given this mapping, the representation is accurate when aRb is the case. It is also accurate when aRc is the case. Hence, misrepresentation occurs only when *both* aRb and aRc are not the case. But, if this is correct, then the representation represents (aRb or aRc), and its content is disjunctive and thus indeterminate. Yet, it is widely assumed that a successful theory of content *must* deliver us determinate content. So, it seems to me that, in order for a structural-resemblance based theory of content to be successful, it must exclude *one-to-many* mappings. Now, the issue with weights in connectionist systems is that they seem to map one-to-many: each weight encodes information about many targets (see Clark, 1993, pp. 13–17; Van Gelder, 1991, pp. 42–47; Ramsey, Stich and Garon 1991, pp. 215–217 for early renditions of this point). Hence, it seems that each weight is bound to map onto *many* targets, generating the problem with content determinacy. Notice, importantly, that the same line of reasoning holds even when the *relations* map onto many. To see why, consider a modified version of the minimal structural representation considered above, in which a^* maps onto a, b^* maps onto b and R^* maps onto two relations R and F. Again, given this mapping, misrepresentation occurs only when *both* aRb and aFb are not the case, and so the representation represents (aRb or aFb). In both cases, the disjunction problem is brought about by the claim that one-to-many mappings might constitute structural similarities, so as to circumvent the problems raised by superspositionality. Hence, we should *not* allow one-to-many to constitute structural similarities. Thanks to an anonymous referee for having pressed me on this point.

[24] More precisely, it is common in the PP literature most heavily influenced by Friston's free energy principle. Many thanks to an anonymous referee for having noticed this imprecision.

representationalist claim. This is because, in our best explanatory practices, "representation" typically denotes states of information processing systems (e.g. Kandel et al. 2012 p. 372), rather than entire information processing systems.[25] The structural similarity presented above seems to enable us to vindicate *only* metaphysical representationalism about the whole brain (i.e. the claim that the whole brain *really* is a "big" representation). Given that Gładziejewski's account of structural representations aims at vindicating both metaphysical and epistemic representationalism, this way of vindicating (a) seems to lead to a partial failure of his account.[26]

Secondly, a complaint about content. What would such a "whole brain" structural representation represent? If I understand Friston correctly, the brain is supposed to recapitulate the causal structure of the world. Thus, the relevant structural similarity holds between the anatomical structure of the brain and the causal structure of the world. But a structural representation represents the target whose structure is mirrored in the structure of the vehicle, and here such a target is *the world* (see Wiese, 2018, p. 219; Williams, 2018a, 2018b, p. 154–155). This is not the kind of content naturalistic theories of content are supposed to deliver, for *the world* is not the kind of content invoked in the scientific explanations of our cognitive capacity, nor the kind of content relevant to our personal-level mental states. This isn't a knockdown objection against alternative argument 3. But it surely shows that the argument has some very undesirable consequences.

Lastly, and, I believe, most importantly, this way of vindicating (a) seems to prevent (c) from obtaining. If the entire brain is a single gigantic representation representing the world, it is very hard to see how decouplability might obtain. There is always *some* sort of causal contact between brains and worlds. Since point (c) spells out decouplability in terms of causal contact, this way of vindicating (a) seems to prevent the obtaining of (c).[27]

[25] Notice that the former usage of "representation" is consistent with the PP literature (e.g. Friston, 2005, p. 819; Kiefer and Hohwy, 2018, p. 2396).

[26] One might contend this verdict is premature. For the elements (i.e. objects of V and relations of $\mathfrak{R}_V$) of structural representations are representational vehicles in their own right (e.g. Shea, 2018, p.118; Ramsey, 2007, p. 79, footnote 3). Thus, claiming that the brain as a whole is a structural representation might in principle justify the claim that the relevant elements of the structural similarity (i.e. patterns of activation) are representations too, leading to a vindication of epistemic representationalism. I believe that the problem with this line of reasoning is the following: the brain-world structural similarity Friston envisages is *not* defined over patterns of activation in the brain. Rather, it is defined over the anatomical structure of the brain. The relevant elements in the structural similarity are not patterns of activation. Hence, this way of vindicating (a) fails to properly vindicate the epistemic representationalist claim.

[27] Here, one might be tempted to simply reject condition (c) and accept that entire brains are structural representations of the environment. As far as I can see, this is a legitimate move. However, it seems quite an ad hoc move. There are good independent reasons to hold that representations are necessarily decouplable from their targets (see Grush, 1997; Webb, 2006; Pezzulo, 2008: Orlandi, 2014, pp. 120–134). Moreover, abandoning (c) would likely make Gładziejewski's account of structural representations far too liberal, as Gładziejewski himself acknowledges (Gładziejewski, 2016, p. 571).

4.2.4 Alternative argument 4

Alternative argument 4: perhaps there is a way to make "whole brain" representations work. Thus, consider Kiefer and Hohwy's (2018, 2019) defense of generative models as structural representations.[28]

On the view Kiefer and Hohwy favor, we should conceive the brain as a complex causal network. If I understand them correctly, we should interpret each node in such a network as a definite pattern of neuronal activity, and the arrows connecting the nodes as causal relations between patterns (i.e. if node *a* is connected to node *b*, then pattern *a* causes pattern *b*). This network of causal relations, on the account Kiefer and Hohwy propose, structurally resembles the causal structures of the world as captured by "material inferences"; that is, inferences such as that from "It's raining" one infers "The street is wet" (see Kiefer & Hohwy, 2018, pp. 2392–2393). In this way, the entire brain (which instantiates the causal network), comes to reflect, and hence to represent, the causal structure of the world.

Kiefer and Hohwy's account of "whole brain" structural representations seems to me a significant improvement from the previously scrutinized one. For one thing, given that in this view the relevant elements of the structural representation are patterns of activation, and given that the elements of a structural representations can be counted as representations in their own right, Kiefer and Hohwy's proposal would allow one to substantiate the epistemic representationalist claim. Moreover, it can also assign each individual pattern of activation a determinate content, depending on its causal embedding within the network. However, it seems to me that relying on Kiefer and Hohwy's proposal to vindicate (a) has serious drawbacks.

To start, the problem with (c) is not solved by Kiefer and Hohwy's account.[29] If the brain is a complex causal network mirroring the causal structure of the world, it is correct to say that the relevant structural representation (i.e. the brain) represents the world. And I simply do not see how one could sever the constant brain-world causal contact so as to vindicate (c).[30]

[28] To be clear, Kiefer and Hohwy do not *explicitly* set out to defend "whole brain" representations. However, it seems to me that their account entails that the whole brain is a structural representation, at least insofar they take the entire causal network *instantiated by the brain* to be the relevant structural representation. A reviewer noticed that this characterization of Kiefer and Hohwy's position might be too ungenerous, since, strictly speaking, Kiefer and Howhy speak only of connections among *cortical* regions. Hence, their position is best described as a form of "whole cortex", rather than "whole brain" representationalism. However (and the reviewer seems to agree) noticing this does not substantially alter the dialectical situation. So, I will continue to speak of Kiefer and Hohwy as endorsing a form of "whole brain" representationalism, mainly for the sake of simplicity.

[29] Notice, importantly, that Kiefer and Hohwy *seem* to consider decouplability a necessary feature of representations, see (Kiefer and Hohwy 2019, p. 400).

[30] Of course, individual patterns of activation can be decoupled from the individual target they represent in virtue of the overall brain-world structural similarity. However, to be satisfied, point (c) requires that the *entire vehicle* of structural representation (in this case, the whole brain) is decoupled from its target (in this case, the world). Thus, noticing that in some cases (e.g. during dreaming) certain patterns of activation are tokened in a way that is functionally independent from the incoming sensory stimulation is not sufficient to vindicate point (c). This is because individual patterns of activations are not the *entire* vehicle of the structural representations, but rather elements of that vehicle. Thanks to an anonymous referee for having pressed me to clarify this point.

Secondly, Kiefer and Hohwy's account raises a puzzle about the inferential status of brain processes. If causal relations holding among patterns of activation are members of $\mathfrak{R}_V$, it seems to me that it logically follows that they *are part of the vehicle*. After all, according to O'Brien and Opie's (2004) definition of structural similarity, both the objects of V and the relations in $\mathfrak{R}_V$ are parts of S_V. But in order for the relevant structural similarity to satisfy (a), S_V must be a vehicle. Hence, Kiefer and Hohwy's proposal seems to imply that causal relations among patterns of activations are part of the vehicle. But if this is the case, then it seems to me that these causal relations cannot be inferential processes, for inferential processes seem to be distinct from the representational vehicles upon which they operate. So, it seems that if Kiefer and Hohwy's (2018, 2019) account of structural similarity is accepted, causal interactions among neural activity patterns cannot be rightfully called inferences. And this seems a problem, given that the inferentialist reading of PP tends to go hand in hand with the claim that generative models are structural representations (e.g. Gładziejewski, 2017; Hohwy, 2018; Kiefer, 2017).

Lastly, a wholesale acceptance of Kiefer an Hohwy's (2018, 2019) account might, paradoxically, force one to abandon the claim that generative models are structural representations. The point is subtle but important. According to Kiefer and Hohwy:

> 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. (Kiefer & Hohwy, 2018, p. 2393; see also Kiefer & Hohwy, 2019, p. 402; Kiefer, 2020, endnote 17)

But this is not how the parts (i.e. objects and relations) of a structural representation acquire their contents. The content of a structural representation is determined by the relevant structural similarity it bears to a target; and the content of the *parts* (i.e. the elements of V and $\mathfrak{R}_V$) of a structural representation is determined by the way in which they participate in the relevant structural similarity; that is, by the way in which they map onto a corresponding element of the target. The relevant relation determining the contents of the elements of a structural representation is the structural similarity holding between the vehicle and the target; not the relations $\mathfrak{R}_V$ holding among the various members of V. Surely, since structural similarity is *structural*, it must, in some relevant sense, be *sensitive to* the relevant members of $\mathfrak{R}_V$. But this does not entail that the content of the elements of a structural representation is determined by their relations $\mathfrak{R}_V$.

Another, perhaps more perspicuous, way to put the same point is this: were the content of the elements of V determined by the relations $\mathfrak{R}_V$ holding between them, then the elements of V would have content *whether the entire vehicle is structurally similar to something or not*. Moreover, even in cases in which the whole vehicle S_V is structurally similar to some target S_O, there is, as far as I can see, no *prior* guarantee that the content assigned to each element in V in virtue of the relations in $\mathfrak{R}_V$ *would* match the content each element of V would bear, were their contents determined by the relevant mapping from S_V to S_O constituting the structural similarity. To put it bluntly, what I'm trying to point out is this: Kiefer and Hohwy espouse a form of functional role semantics. But functional role semantics and structural similarity have no essential connections, *pace* Kiefer and Hohwy. It thus seems to me

that a wholesale adoption of Kiefer and Hohwy's proposal ends up undermining the broader structural-representationalist claim. Kiefer and Hohwy might provide a way to vindicate (a); but a wholesale acceptance of their proposal seems to make such a vindication redundant. If one adheres to functional role semantics,one does not *need* a structural similarity.

This is not to deny that Kiefer and Hohwy (2018, p. 2393; 2019, p. 402) stress that the relevant (i.e. content constituting) functional relations in $\mathfrak{R}_V$ mirror, in the relevant sense, the causal structure of the world (i.e. $\mathfrak{R}_O$): in their view, functional role semantics *entails* a relevant structural similarity. But this surely does not allow us to count Kiefer and Hohwy as defenders of structural representations. For, as many have stressed, causal/informational theories of content entail a relevant vehicle-target structural similarity too (Morgan, 2014; Nirshberg & Shapiro, 2020; Facchin 2021). But surely causal/informational theories of content are not structural resemblance-based theories of content, for at least three reasons. First, the claim that content is determined by causal/informational factors is logically distinct from the claim that content is determined by a relevant vehicle-target structural similarity. Secondly, it is not the case that the obtaining of all vehicle-target structural similarities hinges upon some appropriate causal/informational relation holding the vehicle and the target (see Shea, 2018, p. 139–140). Lastly, in the case of genuine structural representations, the relevant structural similarity is what determines the relevant content. Hence, it should not be a "side effect" of some other *content-determining* factor (see Gładziejewski & Miłkowski, 2017 for further discussion).[31] It seems to me that one needs only to substitute "causal/informational relations" with "functional relations" to make the same remarks about Kiefer and Hohwy's proposal.

4.2.5 Alternative argument 5

Alternative argument 5: one might further try to vindicate (a) by claiming that, since generative models can be rendered as Bayesian nets, and that Bayesian nets are computationally useful because they are structurally similar to their target (Danks, 2014 p. 39), *whatever piece of machinery* is instantiating the relevant generative models must also be structurally similar to the target to be computationally useful. This way of vindicating (a), however, seems flawed. Generative models can be run by everyday personal computers: Von Neumann architectures computing over arbitrary *symbols*. And symbols surely aren't structural representations: in fact, the two are typically contrasted (O'Brien & Opie, 2001; Williams & Collings, 2017).

There might be other ways to vindicate (a). But I'm not claiming that the proposition "generative models are structural representations" is *false*. I'm only claiming that it is presently *unjustified*. And the discussion above seems sufficient to substantiate that claim. If the arguments provided above are correct, Gładziejewski's

[31] One might object that Kiefer and Hohwy (2018, 2019) should be counted as defending structural representations because they stress that the relevant structural similarity is relevant for the system's success. As I understand it, the problem with this line of argument is that the same holds true also for causal theories of content (see Nirshberg and Shapiro, 2020, pp. 6–7; Facchin 2021, pp. 9–12).

original argument fails to substantiate the claim that generative models are structural representations; and, as far as I can see, there is no "rough and ready" way to vindicate that claim on offer in the current philosophical market.

5 Concluding Remarks

In this essay, I have tried to argue that the identification of generative models with structural representations is, at present, unjustified. Here, I wish to single out some recurrent themes that emerged in the discussion above, as they are likely to be important to understand the theoretical commitments of PP.

One issue that repeatedly emerged from the discussion above concerns the vehicles of generative models. The PP literature, I believe, is fairly confusing on this point. The word "model" is in fact applied to a variety of quite distinct things, including: (i) the whole brain (e.g. Bastos et al., 2012, p. 702), (ii) axonal connections (e.g. Shipps, 2016, p. 3), (iii) functionally specialized networks of neural areas (e.g. the mirror system as a model of bodily kinematics, see Kilner et al., 2007), (iv) neuronal responses and connections (Buckley et al. 2017, p. 57), (v) the spinal cord[32] (Friston, 2011, p. 491), (vi) single hierarchical levels[33] (e.g. Kiefer & Hohwy, 2019, p. 387) and I would not be surprised if this list is not complete. This liberal, almost casual, usage of "model" makes it very difficult to understand what piece of neural machinery should be taken as the vehicle of the relevant representation. It also makes unclear what sort of structural similarity would be appropriate to vindicate a structural representationalist claim. I believe that clarifying this point should be a priority for philosophers (and empirical scientists) interested in defending the claim that generative models are structural representations. This is because structural representations are defined in terms of a (relational) vehicle property; namely, structural similarity. Hence, determining what counts as the relevant vehicle is essential in order to vindicate the claim that the relevant vehicle is the vehicle of a structural representation.

A related problem is that it is unclear whether the candidate vehicle (that is, the candidate piece of neural circuitry) is supposed to *be* (or embody) a model or merely to *encode* a model.[34] As far as I can see, "being/embodying a model" and "encoding a model" are used roughly as synonyms in the PP literature. Yet, there seems to be an obvious difference between the two claims: the hard drive of my PC surely encodes numerous early drafts of this very essay, but my hard drive is *not* an early draft of this essay. If the core claim PP makes is that the relevant candidate vehicle

[32] To be precise, Friston suggests that the *dorsal horn* of the spinal cord embodies an *inverse* model. But an inverse model still seems to me to count as a model.

[33] Presumably, single, well identified, regions of the cortical hierarchy.

[34] This issue seems to me importantly related to the "having VS. being" a model in the literature on the free energy principle (see van Es, 2020; Baltieri et al., 2020; see also Bruineberg et al., 2020). I must confess, however, that I'm unsure about how to properly articulate such a relation.

is a model, and models really are structural representations, then *some* structural similarity *must* be found; otherwise, PP would be in trouble. But if the core claim PP makes is just that the relevant candidate vehicle simply *encodes* a model, then the absence of any relevant vehicle-target structural similarity might be entirely unproblematic (supposing that not each and every form of encoding entails a structural similarity).

This latter point also suggests that, where the core claim of PP that the brain only *encodes* a generative model, the reading of generative models as structural representations would not be mandatory. For this reason, I believe that philosophers willing to defend a representationalist account of PP need not *necessarily* commit themselves to a "structuralist" account of representations. Representations in cognitive science need not necessarily be structural representations; and it would be interesting to explore whether a representationalist account of PP not based on structural representations is viable.[35]

Lastly, a point about the rhetoric of the philosophical discussion surrounding PP. Many philosophers defending a representational reading of PP based on structural representations seem to hold that the "representation wars" are over, since PP has conclusively resolved the issue in favor of (structural) representationalism (e.g. Clark, 2015; Gładziejewski, 2016; Williams, 2017). Now, I find it sincerely hard to deny that structural representations are increasingly popular in cognitive neuroscience (e.g. Poldrack, 2020; Williams & Colling, 2017).[36] However, if the argument I have provided here is correct, the newfound popularity of structural representations might have very little to do with predictive processing. Identifying the factors that actually contribute to the popularity of structural representations might prove useful to fully understand the strength and merits of both representationalism and antirepresentationalism in cognitive science.

Acknowledgments The author wishes to thank the participants to the IUSS WIP seminars for useful feedback on the essay. Thanks also to Niccolò Negro and Giacomo Zanotti for their useful comments on some previous version of this essay. A special thanks goes to Eleonora, for her moral support.

Author Contribution MF is the sole author of the paper.

Funding This work has been funded by the PRIN Project "The Mark of Mental" (MOM), 2017P9E9N, active from 9.12.2019 to 28.12.2022, financed by the Italian Ministry of University and Research.

Declarations

[35] Arguably, Kiefer and Hohwy's (2018, 2019) account is one such account, given Kiefer and Hohwy's commitment to functional role semantics. However, given that they seem to take (wrongly, in my opinion) functional role semantics as a kind of structural resemblance, it is very hard to evaluate their proposal as an *alternative* to structural representations-based accounts of PP.

[36] A reviewer noticed that structural representations are less popular in the philosophy of mind, where teleosemantic theories of content still appear to dominate. It might be worth noticing, at this point, that teleosemanticists are increasingly willing to incorporate *some* forms of structural similarity in their accounts (e.g. Millikan, 2020; Neander, 2017). Moreover, the standard notion of *exploitable* structural similarity has been elaborated within a roughly teleosemantic framework (Shea, 2018). Yet, nothing, in my argument, hinges over this.

Conflict of interests The author declares no conflict of interests.

References

Adams, R., et al. (2013). Predictions, not commands: Active inference in the motor system. *Brain Structure and Function, 218*(3), 611–643.

Albers, A. M., et al. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology, 23*(15), 1427–1431.

Allen, M., & Friston, K. (2018). From cognitivism to autopoiesis: Towards a computational framework for the embodied mind. *Synthese, 195*(6), 2459–2482.

Baltieri, M., et al. (2020). Predictions in the eye of the beholder: An active inference account of Watt governors. *Artificial Life Conferences*. https://doi.org/10.1162/isal_a_00288.

Bastos, A. M., et al. (2012). Canonical microcircuits for predictive coding. *Neuron, 76*(4), 695–711.

Bickhard, M. H. (1999). Interaction and representation. *Theory and Psychology, 9*, 435–458.

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

Bruineberg, J., et al. (2020). The emperor's new Markov Blankets [preprint]. Accessed at http://philsci-archive.pitt.edu/18467/, Accessed 15 Dec 2020

Buckley, C. L., et al. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology, 81*, 55–79.

Chemero, A. (2009). *Radical embodied cognitive science*. . The MIT Press.

Churchland, P. M. (1986). Some reductive strategies in cognitive neurobiology. *Mind, 95*(379), 279–309.

Curchland, P. M. (2012). *Plato's Camera. How the physical brain captures a landscape of abstract universals*. . The MIT Press.

Clark, A. (1993). *Associative engines*. . The MIT Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204.

Clark, A. (2015). Predicting peace: the end of the representation wars. In T. Metzinger, J. M. Windt (Eds.). *Open MIND*: 7, Frankfurt am Main: The MIND Group. https://doi.org/https://doi.org/10.15502/9783958570979.

Clark, A. (2016). *Surfing uncertainty*. . Oxford University Press.

Clark, A. (2017). Busting out: predictive brains, embodied minds, and the puzzle of the evidentiary veil. *Noûs, 51*(4), 727–753.

Colombo, M., Elkin, L., & Hartmann, S. (2018). Being realist about Bayes, and the predictive processing theory of the mind. *The British Journal of Philosophy of Science*. https://doi.org/10.1093/bjps/axy059.

Danks, D. (2014). *Unifying the mind*. . The MIT Press.

Dayan, P., & Hinton, G. (1996). Varieties of Helmholtz machine. *Neural Networks, 9*(8), 1385–1403.

De Vries, B., & Friston, K. (2017). A factor graph description of deep temporal active inference. *Frontiers in Computational Neuroscience, 11*, 95.

Dolega, K. (2017). Moderate predictive processing. In T. Metzinger, W. Wiese (Eds.), *Philosophy and predictive processing*, 10, Frankfurt am Main: The MIND Group, https://doi.org/https://doi.org/10.15502/9783958573116.

Dolega, K., & Dewhurst, J. E. (2020). Fame in the predictive brain: a deflationatory approach to explaining consciousness in the prediction error minimization framework. *Synthese*. https://doi.org/10.1007/s11229-020-02548-9.

Donnarumma, F., et al. (2017). Action perception has hypothesis testing. *Cortex, 89*, 45–60.

Downey, A. (2018). Predictive processing and the representation wars: a victory for the eliminativists (via fictionalism). *Synthese, 195*(12), 5115–5139.

Facchin, M. (2021). Structural representations do not meet the job description challenge. *Synthese*. https://doi.org/10.1007/s11229-021-03032-8.

Fodor, J. (1981). The mind body problem. In J. Heil (Ed.), *(2004), Philosophy of mind: A guide and anthology*. (pp. 162–182). Oxford University Press.

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

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences, 13*(7), 293–301.
Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138.
Friston, K. (2011). What is optimal about motor control? *Neuron, 72*(3), 488–498.
Friston, K. (2013a). Active inference and free-energy. *Behavioral and Brain Sciences, 36*(3), 132–133.
Friston, K. (2013b). Life as we know it. *Journal of The Royal Society Interface, 10*(86), 20130475.
Friston, K. (2019). Beyond the desert landscape. In M. Colombo, E. Irvine, & M. Stapleton (Eds.), *Andy clark and his critics.* (pp. 174–190). Oxford University Press.
Friston, K., et al. (2010). Action and behavior, a free-energy formulation. *Biological Cybernetics, 102*(3), 227–260.
Friston, K., et al. (2017a). The graphical brain: belief propagation and active inference. *Network Neuroscience, 1*(4), 381–414.
Friston, K., et al. (2017b). Active inference: A process theory. *Neural Computation, 29*(1), 1–49.
Friston, K., et al. (2017c). Active inference, curiosity and insight. *Neural Computation, 29*(10), 2633–2683.
Gładziejewski, P. (2015). Explaining cognitive phenomena with internal representations: A mechanistic perspective. *Studies in Logic, Grammar and Rhetoric, 40*(1), 63–90.
Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese, 193*(2), 559–582.
Gładziejewski, P. (2017). Just how conservative is conservative predictive processing? *Internetowy Magazyn Filozofinczny Hybris, 38*, 98–122.
Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: Causally relevant and different from detectors. *Biology and Philosophy, 32*(3), 337–355.
Goodman, N. (1969). *The languages of art*. . Oxford University Press.
Grush, R. (1997). The architecture of representation. *Philosophical Psychology, 10*(1), 5–23.
Haykin, S. (2009). *Neural networks and machine learning*. . Pearson.
Hinton, G. (2007a). To recognize shapes, first learn to generate images. *Progress in Brain Research, 165*, 535–547.
Hinton, G. (2007b). Learning multiple layers of representations. *Trends in Cognitive Sciences, 11*(10), 428–434.
Hinton, G. E. (2014). Where do features come from? *Cognitive Science, 38*(6), 1078–1101.
Hinton, G. E., & Sejnowski, T. E. (1983). Optimal perceptual inference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 448.
Hohwy, J. (2013). *The predictive mind*. . Oxford University Press.
Hohwy, J. (2015). Prediction, agency, and body ownership. In A. K. Engel, K. Friston, & D. Kragic (Eds.), *The pragmatic turn.* (pp. 109–138). The MIT Press.
Hohwy, J. (2016). The self-evidencing brain. *Noûs, 50*(2), 259–285.
Hohwy, J. (2017). How to entrain your evil demon. In T. Metzinger, W. Wiese (Eds.), *Philosophy and Predictive Processing*, 2, Frankfurt am Main: The MIND Group, https://doi.org/https://doi.org/10.15502/9783958573048.
Hohwy, J. (2018). The predictive processing hypothesis. In A. Newen, L. De Bruin, & S. Gallagher (Eds.), *The Oxford handbook of 4E cognition.* (pp. 129–146). Oxford University Press.
Hohwy, J. (2019). Prediction error minimization in the brain. In M. Sprevak & M. Colombo (Eds.), *The routledge handbook of the computational mind.* (pp. 159–172). New York: Routledge.
Hohwy, J. (2020). New direction in predictive processing. *Mind & Language*. https://doi.org/10.1111/mila.12281.
Huang, Y., & Rao, P. (2011). Predictive coding. *Wiley Interdisciplinary Reviews, 2*(5), 580–593.
Kandel, E. R., Schwartz, J. H., Jessel, T. M., Siegelbaum, S. A., & Hudspeth, A. J. (Eds.). (2012). *Principles of neural science*. (5th ed.). The MacGraw-Hill Companies.
Kiefer, A. (2017). Literal perceptual inference. In T. Metzinger, W. Wiese (Eds.), *Philosophy and Predictive Processing*, 17, Frankfurt am Main: The MIND Group, https://doi.org/https://doi.org/10.15502/9783958573185.
Kiefer, A. (2020). Psychophysical identity and free energy. *Journal of the Royal Society Interface*. https://doi.org/10.1098/rsif.2020.0370.
Kiefer, A., & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese, 195*(6), 2387–2415.

Kiefer, A., & Hohwy, J. (2019). Representation in the prediction error minimization framework. In S. Robins, J. Symons, & P. Calvo (Eds.), *The Routledge companion to philosophy of psychology.* (2nd ed., pp. 384–410). Routledge.
Kilner, J., Friston, K., & Frith, C. (2007). Predictive coding: An account of the mirror neuron system. *Cognitive Processing, 8*(3), 159–166.
Kirchhoff, M. D., & Robertson, I. (2018). Enactivism and predictive processing: a non-representational view. *Philosophical Explorations, 21*(2), 264–281.
Knill, D., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Cognitive Science, 27*(12), 712–719.
Koski, T., & Noble, J. (2009). *Bayesian networks: An introduction*. . Wiley.
Lee, J. (2018). Structural representations and the two problems of content. *Mind & Language, 34*(5), 606–626.
Leitgeb, H. (2020). On non-eliminative structuralism Unlabeled graphs as a case study, part A. *Philosophia Mathematica*. https://doi.org/10.1093/philmat/nkaa001.
Matsumoto, T., & Tani, J. (2020). Goal directed planning for habituated agents by active inference using a variational recurrent neural network. *Entropy, 22*(5), 564.
McClelland, J., & Rumelhart, D. (1986). *Parallel distributed processing*. (Vol. II). The MIT Press.
Mesulam, M. (2008). Representation, inference, and trascendent encoding in neurocognitive networks of the human brain. *Annals of Neurology, 64*(4), 367–378.
Millikan, R. G. (2020). Neuroscience and teleosemantics. *Synthese*. https://doi.org/10.1007/s11229-020-02893-9.
Morgan, A. (2014). Representations gone mental. *Synthese, 191*(2), 213–244.
Neander, K. (2017). *A mark of the mental*. . The MIT Press.
Nirshberg, G., & Shapiro, L. (2020). Structural and indicator representations: a difference in degree, not in kind. *Synthese*. https://doi.org/10.1007/s11229-020-02537-y.
O'Brien, G. (2015). How does the mind matter?. In T. Metzinger, J. M. Windt (Eds.), *Open MIND*: 28, Frankfurt am Main: The MIND Group https://doi.org/https://doi.org/10.15502/9783958570146.
O'Brien, G., & Opie, J. (2001). Connectionist vehicles, structural resemblance, and the phenomenal mind. *Communication and Cognition, 34*(1/2), 13–38.
O'Brien, G., & Opie, J. (2004). Notes towards a structuralist theory of mental representations. In H. Clapin, P. Staines, & P. Slezak (Eds.), *Representation in mind: New approaches to mental representaion.* (pp. 1–20). Elsevier.
Orlandi, N. (2014). *The innocent eye*. . Oxford University Press.
Orlandi, N. (2016). Bayesian perception is ecological perception. *Philosophical Topics, 44*(2), 327–352.
Pezzulo, G. (2008). Coordinating with the future: The anticipatory nature of representation. *Minds and Machines, 18*(2), 179–225.
Piccinini, G. (2008). Computation without representation. *Philosophical Studies, 137*(2), 205–241.
Poldrack, R. (2020). The physics of representation. *Synthese*. https://doi.org/10.1007/s11229-020-02793-y.
Ramsey, W. (2007). *Representation reconsidered*. . Cambridge University Press.
Ramsey, W. (2020). Defending representational realism. In J. Smortchkova, K. Dolega, & T. Schlich (Eds.), *What are mental representations?* (pp. 54–78). Oxford University Press.
Ramsey, W., Stich, S. P., & Garon, J. (1991). Connectionism, eliminativism and the future of folk psychology. In W. Ramsey, S. P. Stich, & D. E. Rumelhart (Eds.), *Philosophy and connectionist theory.* (pp. 199–228). Routledge.
Ramstead, M., Kirchooff, M. D., Friston, K. (2019). A tale of two densities: active inference is enactive inference. *Adaptive Behavior*, 1059712319862774.
Rao, R., & Ballard, D. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience, 2*(1), 79–87.
Rescorla, M. (2012). How to integrate representations is computational modeling, and why we should. *Journal of Cognitive Science, 13*(1), 1–38.
Rogers, T. T., & McClelland, J. L. (2014). Parallel Distributed Processing at 25: Further explorations in the microstructure of cognition. *Cognitive Science, 38*(6), 1024–1077.
Rumelhart, D., & McClelland, J. (1986). *Parallel distributed processing*. (Vol. I). The MIT Press.
Seth, A. (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. (2015). The cybernetic Bayesian brain. In T. Metzinger, J. M. Windt (Eds.), *Open MIND*: 35, Frankfurt am Main: The MIND Group https://doi.org/https://doi.org/10.15502/9783958570108.

Seth, A., & Friston, K. (2016). Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B, 371*(1708), 20160007.
Shagrir, O. (2001). Content, computation and externalism. *Mind, 110*(438), 369–400.
Shea, N. (2013). Perception versus action: the computations might be the same but the direction of fit differs. *Behavioral and Brain Sciences, 36*(3), 228–229.
Shea, N. (2014). VI: Exploitable isomorphism and structural representation. *Proceedings of the Aristotelian Society, 114*(22), 123–144.
Shea, N. (2018). *Representations in Cognitive Science*. . Oxford University Press.
Shipp, S. (2016). Neural elements for predictive coding. *Frontiers in Psychology, 7*, 1792.
Sims, A. (2017). The problems with prediction. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*: 23, Frankfurt am Main: The MIND Group https://doi.org/https://doi.org/10.15502/9783958573246
Sporns, O. (2010). *Networks in the brain*. . The MIT Press.
Spratling, M. W. (2016). Predictive coding as a model of cognition. *Cognitive Processing, 17*(3), 279–305.
Sprevak, M. (2010). Computation, individuation and the received view on representation. *Studies in History and Philosophy of Science Part A, 41*(3), 260–270.
Sprevak, M. (2013). Fictionalism about neural representations. *The Monist, 96*(4), 539–560.
Tani, J. (2016). *Exploring robotic minds*. . Oxford University Press.
van Es, T. (2020). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior*. https://doi.org/10.1177/1059712320918678.
Van Gelder, T. (1991). What is the "D" in "PDP"? A survey of the concept of distribution. In W. Ramsey, S. P. Stich, & D. E. Rumelhart (Eds.), *Philosophy and connectionist theory.* (pp. 33–61). Rutledge.
Van Gelder, T. (1992). Defining distributed representations. *Connection Science, 4*(3–4), 175–191.
Webb, B. (2006). Transformation, encoding and representation. *Current Biology, 16*(6), 184–185.
Wiese, W. (2017). What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences, 16*(4), 715–736.
Wiese, W. (2018). *Experienced wholeness: Integrating insights from gestalt theory, cognitive neuroscience and predictive processing*. . The MIT Press.
Williams, D. (2017). Predictive processing and the representation wars. *Minds And Machines, 28*(1), 141–172.
Williams, D. (2018a). Predictive coding and thought. *Synthese, 197*(4), 1749–1775.
Williams, D. (2018b). Predictive minds and small-scale models: Kenneth Craik's contribution to cognitive science. *Philosophical Explorations, 21*(2), 245–263.
Williams, D., & Colling, L. (2017). From symbols to icons: The return of resemblance in the cognitive science revolution. *Synthese, 195*(5), 1941–1967.
Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Science, 10*(7), 301–308.