Synthese (2021) 199:11609–11642
https://doi.org/10.1007/s11229-021-03304-3

ORIGINAL RESEARCH

Predictive processing and anti-representationalism

Marco Facchin[1]

Received: 6 February 2021 / Accepted: 5 July 2021 / Published online: 14 July 2021

Abstract
Many philosophers claim that the neurocomputational framework of predictive processing entails a globally inferentialist and representationalist view of cognition. Here, I contend that this is not correct. I argue that, given the theoretical commitments these philosophers endorse, no structure within predictive processing systems can be rightfully identified as a representational vehicle. To do so, I first examine some of the theoretical commitments these philosophers share, and show that these commitments provide a set of necessary conditions the satisfaction of which allows us to identify representational vehicles. Having done so, I introduce a predictive processing system capable of active inference, in the form of a simple robotic "brain". I examine it thoroughly, and show that, given the necessary conditions highlighted above, none of its components qualifies as a representational vehicle. I then consider and allay some worries my claim could raise. I consider whether the anti-representationalist verdict thus obtained could be generalized, and provide some reasons favoring a positive answer. I further consider whether my arguments here could be blocked by allowing the same representational vehicle to possess multiple contents, and whether my arguments entail some extreme form of revisionism, answering in the negative in both cases. A quick conclusion follows.

Keywords Anti-representationalism · Predictive processing · Structural representations · Mental content · Sensorimotor contingencies

1 Introduction

Many philosophers argue that the neurocomputational framework of predictive processing (PP) entails a form of global representationalism and inferentialism about cognition. Their reasoning seems the following: PP casts perception as a top-down process in which brains try to actively predict the incoming sensory inputs. Since

✉ 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

this process approximates Bayesian inferences, PP is an inferentialist theory of perception (e.g. Kiefer, 2017). But inferences requires representations; and in fact, PP extensively quantifies over generative models, which, *being models*, are structural representations: vehicles representing their targets by mirroring their inner relational structure (e.g. Gładziejewski, 2016; Kiefer & Hohwy, 2018, 2019). Since the *same kind* of top-down processing appears to explain cognitive processes in general (Friston, Hohwy, 2015; Spratling, 2016), then cognitive processes *in general* are inferential processes involving representations. As a consequence, inferentialism and representationalism hold about cognition in general.[1]

Here, I scrutinize these interpretations. I argue that, given the theoretical commitments they endorse, the physical structures instantiating generative models do not seem to qualify as representational vehicles. Rather, they appear as *non-representational* structures instantiating an agent's mastery of sensorimotor contingencies (i.e. the ways in which bodily movements systematically alter sensory states). So, if my arguments are on the right track, and PP really has the explanatory breadth most of its supporters believe,[2] then PP seems to naturally lead towards global anti-representationalism; that is, anti-representationalism about cognition in general.

To substantiate my claim, I examine a minimal PP system: a simple robotic "brain" able to predict the incoming input and to act out certain predictions through active inference. I argue that, given the relevant theoretical commitments endorsed by inferentialist and representationalist readings of PP, nothing in that "brain" appears to qualify as a representational vehicle. I also argue that the same conclusion likely generalizes to other PP systems. In this way, the physical structures instantiating generative models will more naturally appear as *non-representational* structures instantiating an agent's sensorimotor mastery.

Importantly, my argument is not based on Ramsey's (2007) job-description challenge.[3] Thus, my argument differs from other popular arguments claiming that PP is not a representationalist theory (Downey, 2018; Orlandi, 2014, 2016, 2018). These arguments consider different seemingly representational PP posits (e.g. priors, predictions, etc.) and argue that, on their own, these posits function either as detectors or as mere biases. Since detectors and biases fail the job-description challenge (i.e. they do not perform any *representational* function), these arguments conclude that PP is not really a representationalist theory, because its posits are not really *representational* posits. Proponents of the inferentialist and representationalist readings

[1] This reconstruction surely downplays the differences between "radical" and "conservative" interpretations of PP (see Clark, 2015a; Gładziejewski, 2017). Yet, since both interpretations are committed to representationalism and inferentialism, I here clump them together.

[2] Importantly, the claim that PP explains cognition in general is highly speculative, and some cognitive processes might lie beyond the explanatory reach of PP (see Williams, 2020). Here, I *assume for the sake of argument* that PP has the explanatory breadth the proponents of its inferentialist-representationalist reading believe. Given this assumption, if my arguments are correct, then *global* anti-representationalism seemingly follows. Yet, if that assumption is not correct, then my arguments *only support a form of limited anti-representationalism*: representations might still be involved in the cognitive processes PP *does not* account for. Many thanks to an anonymous reviewer for having pressed me to clarify this point.

[3] Many thanks to the reviewers for having suggested expanding upon this point, and to Nina Poth who advised me to make this point explicit from the onset.

of PP, however, contend that these arguments simply miss the mark, because, on the representationalist reading they favor, the relevant representational posit is the *entire generative model*, of which priors, predictions and the like are just parts (e.g. Kiefer, 2017: pp. 11–12; Kiefer & Hohwy, 2018: pp. 2394–2395).[4] Here, I will directly confront this representationalist reading on its own terms.

I structure the essay as follows. Section 2 briefly[5] introduces PP, showing how generative models and sensorimotor contingencies are related. Section 3 identifies some theoretical commitments of the representationalist and inferentialist view of PP, showing that they form a set of necessary conditions the satisfaction of which allows us to identify representational vehicles. Section 4 introduces a simple PP system in the form of a robotic "brain", and argues that none of its structures appears to satisfy all the conditions previously identified. Section 5 argues that the anti-representationalist verdict thus obtained likely generalizes to more complex PP systems, and responds to two worries raised by the reviewers. Lastly, Sect. 6 succinctly concludes the essay.

2 Predictive processing: a short introduction

As a theory of perception, PP starts by assuming that sensory states are under-informative in respect to their worldly causes. Each retinal image, for instance, could *in principle* have been caused by indefinitely many environmental layouts (e.g. Palmer, 1999: p. 25). So, to perceive, brains must *estimate* the causes of their sensory inputs, by combining under-informative signals and some prior knowledge on how these signals have likely been produced. PP suggests such an estimate is found by inverting a generative model operating according to a predictive coding algorithm (Friston, 2005).

Generative models are data structures capturing how sensory states might have been produced. Sampling from these models can *generate* sensory states (e.g. Hinton, 2007a), which are predicted (or expected) under the model. This procedure can be intuitively understood as realizing a mapping from external causes (e.g. carrots) to predictable sensory states, given these causes (e.g. orange retinal images).[6]

According to the predictive coding algorithm (Huang & Rao, 2011; Rao & Ballard, 1999), these predictions are then contrasted with the actual sensory states, typically, but not exclusively, by subtraction (Spratling, 2017). Their comparison yields a signal known as prediction error, which is used to revise predictions, so

[4] See (Sims & Pezzulo, 2021) for a nice rational reconstruction of this debate.

[5] As PP is now fairly well-known among philosophers, I will only cover the most essential aspects of it. For more introductory material, see (Clark, 2013, 2016; Hohwy, 2013; Tani, 2016; Wiese & Metzinger, 2017).

[6] To model rich bodies of data such as our sensory states, generative models must be hierarchically organized, so as to capture the hierarchical nesting of worldly causes. However, this only means that each hierarchical layer learns to predict only the layer directly below, predicting the patterns of activation it displays (e.g. Hinton, 2007b).

as to minimize the incoming error.[7] In this way, the system searches for a global minimum of error which, when reached, *inverts* the generative model, as it maps the incoming input (e.g. orange retinal images) onto its most likely cause (e.g. carrots). Importantly, as the cause thus selected approximates an exact Bayesian posterior, PP seems to cast perception as an inferential process performed by means of prediction error minimization.

A similar description holds for action, or active inference.[8] The basic idea behind it is that brains are skewed towards a set of (multimodal) sensory expectations. The error relative to the proprioceptive facets of these predictions is then used to trigger spinal reflexes (Adams et al., 2013; Friston, 2011), so as to bring about the predicted interoceptive sensory states, and eventually encounter the whole multimodal prediction. Active inference is thus a process of error minimization in which the predicted sensory states are *brought about* through movement (e.g. Namikawa et al., 2011); thereby bringing about the evidence in favor of one's generative model (Hohwy, 2016, 2020).

Importantly, the predictions triggering active inference are always multimodal, and non-proprioceptive predictions can sometimes (more or less directly) drive active inference too (see Pio-Lopez et al., 2016). In fact, if, as PP suggests, the only function of the brain is that of keeping prediction error at a minimum over time (see Friston, 2009, 2010), it is hard to see how these predictions can be *but* multimodal. A brain unable to predict the visual consequences of a saccadic eye movement, for instance, would be unable to effectively minimize prediction error, as each saccade would bring about unpredicted (i.e. error inducing) visual input. This immediately connects generative models to sensorimotor contingencies.

Sensorimotor contingencies are law-like relations capturing how the sensory states of a system evolve, given a system's movements and the relevant features of a system's sensorium and environment (Brette, 2016; O'Regan, 2011; O'Regan & Noë, 2001). Approaching an object, for instance, will make its retinal image expand; whereas backing away from it will make the retinal image contract.[9] Although theorists introducing sensorimotor contingencies never specified what sort of structures could realize a system's knowledge of sensorimotor contingencies, generative models appear to be ideal candidates (Pezzulo et al., 2017; Seth, 2014; Vásquez, 2019; see also Hemion, 2016; Laflaquiere, 2017).

To briefly see why, consider the role of forward models in motor control. Forward models are *special purpose* generative models, tasked with converting motor commands into the predictable sensory consequences of movement. Clearly, to function properly, a forward model must encode the relevant sensorimotor contingencies, as its role is precisely that of predicting how bodily movements alter sensory states (see Maye & Engel, 2013: p. 425; Pezzulo, 2011).

[7] Prediction error is often weighted according to the expected signal-to-noise ratio of the data. Roughly, this is how PP accounts for attention (see Feldman & Friston, 2010).

[8] This sacrifices precision to ease of exposition: active inference is also responsible for changes of bodily states that are not actions (see Seth & Friston, 2016).

[9] Provided, of course, that the moving system has eyes, that there is light in the environment, and that the object is still. These are examples of the "relevant features" of the system's sensorium and the environment.

In traditional theories of motor control, forward models work in tandem with *inverse* models, converting goal states into motor commands. A copy of the motor command thus computed is sent to the forward model, to estimate the expected sensory consequences of movement. The estimate is needed for a variety of reasons. For instance, it allows the agent to control and correct actions on-line, in spite of the fact that the reafferent signal is noisy and delayed. It also allows the agent to proactively adjust in regard to the foreseeable consequences of its own actions (Franklin & Wolpert, 2011). Forward models can also act as filters, allowing the agent to ignore the predictable, and thus uninformative, "bits" of the reafferent signal (Blackmore et al., 1999).

According to PP, however, there is *only* the forward model. More precisely, there is only one integrated generative model busy predicting the motor-dependent sensory states the agent "desires to encounter"[10]; the motor plant itself will then bring them about through movement (Friston, 2011; Pickering & Clark, 2014). Generative models able to perform active inference, thus, appear as ideal candidates to implement an agent's sensorimotor mastery, as they must encode parsimonious descriptions of sensorimotor loops (Baltieri and Buckely, 2019; Tschantz et al., 2020).[11]

Inferentialist and representationalist interpretations of PP conceive generative models as structural representations: vehicles representing their targets in virtue of the structural similarity holding between them. If they are on the right track, and my presentation of PP is correct, it thus follows that the structures instantiating our sensorimotor mastery are representational vehicles. But what does it take to be one?

3 Some necessary features of representational vehicles

Here, I expose some of the theoretical commitments endorsed by inferentialist and representationalist interpretations of PP. Each commitment spells out a condition that, according to these views, an item *must* satisfy in order to qualify as a representational vehicle. Hence, they jointly provide a minimal set of criteria to determine whether the physical structures instantiating generative models qualify as representational vehicles.

[10] Strictly speaking, these are the sensory states the agent predicts to encounter, given its priors; see (Friston et al., 2012a, b, c).

[11] Here, I mainly focus on perception and action, ignoring PP explanations of other cognitive processes (see Friston 2009, 2010; Spratling 2016). One, however, might fear that an agent's sensorimotor mastery *alone* will not support cognitive processing beyond simple sensorimotor coordinations. Computationally speaking, there are *some* reasons to believe that sensorimotor mastery can support more "thought-like" cognitive processes (e.g. Hay et al., 2018; Le Hir et al., 2018), but that evidence is not conclusive. However, within the PP literature, "thought like" cognitive processes are typically supposed to be supported by the offline functioning of the machinery responsible for perception and action (e.g. Pezzulo, 2017; Tani, 2016). Thus, it seems to me, the image of PP I just painted is not *significantly* removed from the official one.

3.1 Vehicles can be assigned distal and determinate content

Representations are type-identified by their contents, which are both distal and determinate (e.g. Egan, 2012: p. 256). Representations "are about" *well specified worldly targets*, rather than the proximal conditions by means of which these targets are causally encountered. Hence, representational vehicles can always be assigned a determinate and distal content, given a theory of content.

Here, the relevant senses of "distality" and "determinacy" are the ones at play in the horizontal disjunction/stopping problem (Dretske, 1986; Godfrey-Smith, 1989; Neander, 2017). A correct theory of content must allow us to say that a vehicle V represents one, and only one, target T, rather than the disjunction of two or more targets (T or T*). This is determinacy. Moreover, a vehicle must represent a target appropriately "out there". Cognitive agents represent objects and states of affairs of the *distal world*, rather than the more proximal states of affairs causally mediating one's encounter with the distal world, such as the states of one's transducers. Two distinct reasons support this assumption.

The first is empirical adequacy: cognitive scientists, by and large, *do ascribe* determinate and distal content to representations. A neuropsychologist, for instance, might claim that a given pattern of activation of the fusiform face area represents faces. I know of no neuropsychologist claiming that such a pattern of activation represents (faces or face-like conformations); or that it represents face-shaped retinal images. Hence, to be consistent with the normal conduct of cognitive science, philosophical theories of content need to deliver determinate and distal contents.[12]

The second reason is conceptual. Representations must be able to *misrepresent*. But disjunctive (i.e. non determinate) and/or proximal contents do not allow for misrepresentation to occur. To see why, consider Fodor's (1987: pp. 99–102) crude causal theory, according to which a vehicle V represents whichever target causes its tokening. If dogs cause the tokening of V, then V represents dogs. Suppose now a sheep causes a "wild" tokening of V. We would like to say that V *misrepresents* the sheep as a dog. Yet, the crude causal theory prevents us from saying so. If V represents whatever causes its tokening, and its tokening is caused by dogs *or* by sheep, then V represents (dogs or sheep), and tokens of V caused by sheep are not misrepresentations. Further, it could be argued that the tokening of V is not really caused by dogs (or sheep), but by some more proximal conditions, such as quadruped-shaped retinal images. Again, in this case, it seems that "wild" tokenings of V do not misrepresent dogs as sheep. Rather, they correctly represent some more proximal condition, which happens to be disjunctively caused by both dogs and sheep.

Notice that although the horizontal disjunction problem ties them together, distality and determinacy are two *logically independent* requirements, which can *independently* fail to obtain (see Artiga & Sebastián, 2018; Roche & Sober, 2019). A theory

[12] Importantly, this passage and the following should not be taken to imply that all currently existing theories of content do not have the conceptual resources needed to assign determinate and distal contents. Indeed, at least some theories of content seem to have the resources to do so (e.g. Neander, 2017: Ch. 7 and 9).

of content can be in trouble both because it does not provide appropriately determinate content (as in the "vertical" disjunction problem, see Fodor, 1990) or because it provides determinate, but only proximal, content.

Thus presented, distality and determinacy seem two requirements that *a theory of content* must satisfy; and, traditionally, they have been articulated in that way. Their traditional articulation is roughly as follows: representational vehicles have determinate and distal contents. If a given theory of content C does not assign them determinate and distal contents; then C is wrong and ought to be rejected. Notice the argument *assumes* representationalism, and *assesses* theories of content based on their ability to satisfy distality and determinacy.

Yet, the issues concerning distality and determinacy allow to formulate an argument working the other way around; namely, by *assuming* that a given theory of content is correct, one can *assess* whether a candidate vehicle really qualifies as a vehicle, by checking whether it is assigned an appropriately determinate and distal content by the theory. In fact, a correct theory of content supposedly assigns determinate and distal contents to *all and only* representational vehicles. Therefore, if given such a theory a candidate vehicle is not assigned an appropriately distal and determinate content, then the candidate vehicle really is no vehicle. If it were, it would have been assigned a determinate and distal content.

I take this to be the first necessary feature of vehicles of content: vehicles of content must be assigned determinate and distal contents, given a correct theory of content. Clearly, this procedure presupposes a theory of content, whose correctness has to be assumed. In the following, I grant representationalists and inferentialist reading of PP their theory of content of choice. I examine it in the next subsection, showing that it imposes further constraints on the properties of candidate vehicles.

3.2 Exploitable structural similarity

Inferentialist and representationalist accounts of PP argue that generative models are structural representations: vehicles representing their targets *in virtue of* their exploitable structural similarity (Gładziejewski, 2016; Kiefer & Hohwy, 2018, 2019; Wiese, 2018; Williams, 2018). The relevant theory of content they endorse is thus based on two ingredients: (1) structural similarity and (2) exploitability. I unpack them in turn.

Structural similarity is often unpacked as homomorphism (Kiefer & Hohwy, 2018; Wiese, 2018) or "second order structural resemblance", which is a partial homomorphism (Gładziejewski, 2016). Here, I adhere to the second reading. This is because second order structural resemblances are easier to obtain than homomorphisms, and so sticking to it allows me to provide a more charitable reconstruction of inferentialist and representationalist readings of PP.

On this reading, a system V bears a structural similarity to a system T if, and only if, there is a *one-to-one* mapping from at least some constituents v_x of V onto at least some constituents t_x of T such that an identical abstract pattern of relations among

constituents is preserved on both sides of the mapping.[13] Several clarifications seem needed.

First, structural representations are defined in terms of vehicles representing a target in virtue of the (exploitable) structural similarity that ties them together. Structural similarity is thus a relational property of *vehicles*. Hence, candidate vehicles must be structurally similar to their targets. Notice that the relevant structural similarity partially constitutes the relevant content of the vehicle. Hence, to secure distality and determinacy, the relevant structural similarity must hold between a candidate vehicle and some appropriately *distal and well determined* target.

Secondly, both the vehicles and the targets of structural representations must have some internal degree of complexity: they are made up by *constituents* among which certain *relations* hold. I will denote constituents using uncapitalized letters with a subscript (e.g. v_x is an arbitrary constituent of a vehicle V, t_a is a specific constituent of a target T).

Lastly, notice that what is mirrored on both sides of the mapping is a *pattern* of relations, not a relation. This means that the relations holding among the constituents of V and T may differ. Only their *patterns* need to be identical. Suppose, for instance, that V is constituted, among other, by three constituents ordered in the triplet (v_a, v_b, v_c) by their relative magnitude; whereas T is constituted, among other, by three constituents ordered in the tripled (t_a, t_b, t_c) by their relative frequency. If constituents with identical subscripts map one to one onto each other, V and T are structurally similar, in spite of the fact that no common relation holds among their constituents.[14]

Structural similarity alone is clearly insufficient to determine content. Structural similarities do not have the logical properties of representations (Goodman, 1969) and are extremely cheap to come by; so cheap that any two arbitrary systems can be said to be structurally similar in some regard (McLendon, 1955: pp. 89–90; Shea, 2018: p. 112). This is why the relevant vehicle-target structural similarity must be *exploitable*.

Exploitability is canonically defined as the conjunction of two requirements (Shea, 2018: p. 120). First, the relevant relations holding among vehicle constituents must have some systematic downstream effect on the computational operations of the system in which the vehicle is tokened. Secondly, both the constituents of T and their relations must be "of significance" to the system. Here, significance should be unpacked in terms of the system's task functions. Roughly put, task functions are the outputs that a system produces in response to a range of inputs in a range of different conditions and that the system is *supposed to* produce, in virtue of the system's history of selection, individual learning, or explicit (human) design (see Shea, 2018: Ch. 3).

Exploitable structural similarity naturally combines with distality and determinacy, yielding a further requirement on representational vehicles. A candidate

[13] Here, I'm trading precision for ease of exposition. See (O'Brien & Opie, 2004: p. 11) for the canonical formal definition of second order structural resemblance.

[14] Notice also, for the sake of clarity, that V and T need not have any property in common.

vehicle really is a vehicle only if it is assigned a determinate and distal content *in virtue of the exploitable structural similarity* it bears to a determinate and distal target.

3.3 Mathematical contents constrain representational contents

A further constraint must be taken into account. According to representationalist and inferentialist accounts of PP, the representational (distal and determinate) content of a vehicle must at least cohere with its *mathematical content* (Wiese, 2016, 2018). Thus, mathematical content constrains representational content.

Mathematical content is a kind of narrow content which is ascribed to vehicles in virtue of the relevant computational description that the system satisfies; that is, in virtue of the mathematical functions the system computes (Egan, 2014, 2018). Consider for instance how prediction error is computed. Saying that prediction error is computed by subtracting the values of expected and actually received sensory signals means that there is a robust mapping between vehicles and numerical values, such that, anytime the vehicle of the prediction signal maps onto value x and the vehicle of the incoming signal maps onto value y, the prediction error signal produced will map onto value $(x–y)$. In this example, the numerical values of x, y and $(x–y)$ are the *mathematical contents* carried by the vehicles.

The idea that mathematical contents must constrain representational content is attractive because we can explain the functioning of PP systems in two ways. One explanation leverages the mathematical tools of computational theory. Explanations of this kind are provided, for instance, when we say that prediction error encodes the difference between predicted and actual signal, computed by subtraction. The other explanation relies instead on the familiar lexicon of representational contents. According to this kind of explanation, for instance, prediction error represents what was missing from the original prediction; that is, the unexpected features displayed by a perceptual take (e.g. Clark, 2015b: pp. 5–6). Given that both accounts are *literal* explanations detailing how the PP machinery works, they must be at least coherent with each other. For this reason, the assignment of mathematical contents can place some constraints on the assignment of representational contents.[15]

But which constraints does it place? The literature is not explicit on this matter. Wiese (2018: p. 209) only explicitly states that mathematical contents pose "strong constraints" on representational contents, which get stronger as computational theories grow (Wiese, 2016: pp. 724–725).[16] It seems, however, that the constraints mathematical contents place on representational contents are strong enough to *at least partially determine* representational contents. I list some examples below.

[15] This is a significant departure from Egan's (2014, 2018) account of mathematical content. On her account, mathematical contents do not constrain ascription of representational content. Moreover, Egan claims that representational content cannot be naturalized. Conversely, inferentialist and representationalist accounts of PP endorse naturalism about representational content.

[16] A somewhat similar idea seems to be endorsed by (Ramstead et al., 2020b), even if I doubt that Ramstead and colleagues' notion of "representation" is the same notion of representation used in this essay.

Wiese (2016: p. 733) claims that computational models of active inference determine at least some representational contents, as they interpret the signals reaching the motor plant as conveying predicted sensory states rather than motor commands. But predicted sensory states and motor commands are not numerical values, hence they are not mathematical contents. Rather, they are representational contents, which, on Wiese's view, "fall off" directly from the computational rendering of the theory. In a further publication, Wiese (2018: pp. 215–218) suggests that the representational content of a generative model includes everything that can be described by the same set of equations which describe the model computational behavior. Again, it seems that here too Wiese is suggesting that mathematical contents *at least partially* determine representational contents. And it seems to me that Wiese is not alone in endorsing this view.

Gładziejewski (2016: p. 573) argues that the relevant structural similarity holding between generative models and their targets should be construed in terms of the *prior probabilities* of certain events and the *likelihoods* of sensory states, given external events. Prior probabilities and likelihoods are mathematical contents—they are numerical values upon which (some) PP systems compute. Yet, in Gładziejewski's view, they also partially determine the relevant structural similarity; and so the representational content of a generative model.

Clark (2015c: p. 2) and Williams (2018: pp. 162–163) claim that, in PP models, the naturalization of content falls within the scope of computational neuroscience. In their view, what needs to be done to naturalize content just is detailing the *computational functioning* of generative models, showing how such mechanisms "get a grip" on the world (see Hutto & Myin, 2020: pp. 93–97 for further discussion). (Kiefer & Hohwy, 2018, 2019) go as far as proposing a mathematical measure of misrepresentation.

Due to space limitations, I cannot examine any of these proposals in detail. Nevertheless, they are here worth mentioning, to show that representationalist and inferentialist accounts of PP really are committed to the claim that there is a significant interplay between mathematical and representational content; so significant, indeed, that in many cases representational contents seem to derive, more or less immediately, from mathematical ones. Notice, importantly, that this is entirely compatible with the claim that representational contents are determined by exploitable structural similarities. In fact, the relevant structural similarity itself might be visible only under some quite specific mathematical description (e.g. Gładziejewski, 2016; Wiese, 2018: pp. 215–217).

Taking stock: according to inferentialist and representationalist accounts of PP, representational vehicles have a determinate and distal content, which they acquire in virtue of an exploitable structural similarity with an appropriate (i.e. determinate and distal) target. Moreover, the representational (distal and determinate) content of these vehicles is at least coherent with (if not more or less directly determined by) their mathematical content: the numerical values they must represent to allow the computational operations defined over them to take place.

In the next two sections, I will argue that no component of a generative model seems to satisfy that description. Hence, given these theoretical commitments,

generative models will naturally appear as *non-representational structures* instantiating a system's sensorimotor knowledge.

4 The structures instantiating generative models do not appear to be representational vehicles

I split this section into two sub-sections. The first introduces a minimal generative model able to perform active inference. The second examines it, arguing that none of its components qualifies as a representational vehicle, given the requirements highlighted above.

4.1 A minimal generative model capable of active inference

According to PP, generative models are physically instantiated by patterns of neural activation and axonal connections (Friston, 2005: pp. 819–820; Buckley et al., 2017: p. 57). So, patterns of activation and connections are the candidate vehicles of generative models. Hence, connectionist systems are ideally suited to examine the representational commitments of PP (Dołega, 2017; Kiefer & Hohwy, 2018, 2019).

Consider the network Bovet (2007) engineered as a control system for robotic agents, enabling them to display a variety of behaviors involving simple sensorimotor coordinations, such as returning to a "nest" after having explored the environment (Bovet, 2006), smoothly moving using different gaits (Iida & Bovet, 2009) or successfully navigating simple T-mazes (Bovet & Pfeiffer, 2005a, b).

The network is a series of homogeneously connected artificial neural networks, one for each sensory modality of the robotic agent ("motor" modality included). Each net consists of the following three input populations (ending in "S") and two output populations (ending in "C"):

(CS) or *current state* population, receiving input from the sensor or effector of one modality.
(DS) or *delayed state* population, receiving the same input of (CS) after a small delay.
(VS) or *virtual state* population, receiving input from all other nets.
(SC) or *state change* population, receiving input from (CS) and (DS).
(VC) or *virtual change* population, receiving input to (CS) and (VS), and sending output to all other (VS)s (Fig. 1).

The number of neurons of each population varies *across* modalities, but remains constant *within* each modality. This allows the various populations of a single modality to be "copies" of each other. In particular, (DS)s and (VS)s can be "copies" of (CS)s; whereas (VC)s can "mimic" (SC)s. Within each net, the connections running from input to output populations are not trained, and have opposite weights. Moreover, these connections are *neuronwise*: the *n*th neuron of each input population projects only to the *n*th neuron of the relevant output population. Thus, the

patterns of activation of the output populations are defined as the neuron-to-neuron subtraction of activity patterns of the corresponding input populations. Conversely, connections *between* nets are trained, and involve all neurons of the (VC) population of a modality and all the neurons of the (VS)s of all other modalities.

To understand how the network works, consider first (CS)s: they encode, in each modality, the state of the relevant sensor. In the visual modality, for instance, (CS) will reflect the image captured by a camera. (DS)s do the same, but after a small delay: in the visual modality, (DS)'s activity reflects the image captured by the camera *one timestep ago*. (CS)s and (DS)s jointly determine the activation pattern of (SC)s, which thus reflect how the sensory state has changed in a timestep.[17] Continuing with the previous example, (SC) in the visual modality captures how the camera image changed during the delay; for instance, whether it expanded or contracted.

Consider now any two arbitrary modalities *a* and *b*: there will be patterns of coactivation between the neurons in (SC) of modality *a* and those in (CS) of modality *b*. For instance, when visual (SC) encodes the expansion of the camera image, the motor (CS) is typically encoding the fact that the motors are pushing forward. These patterns of coactivation are then used to train, in a purely Hebbian fashion, the connections running from (VC) of modality *a* to (VS) of modality *b*. If the *n*th neuron in (SC) of modality *a* and the *m*th neuron in (CS) of modality *b* fire together, the *n*th neuron in (VC) of modality *a* and the *m*th neuron in (VS) of modality *b* wire together.

This allows the information flowing from (VC)s to (VS)s to be transformed in a way so as to induce, in (VS)s, a pattern of activation that corresponds to the sensory state that modality typically occupies as the other modalities change in a given way; that is, the sensory state expected, given the activity in all (SC)s.[18] Thus, the activity of (VS) estimates (or predicts) a sensory state, given the motor-dependent changes of sensory states in all other modalities. And, in fact, the connections from all (VC) s to all (VS)s constitute a simple generative model, which predicts the sensory states expected, given the robot's activity. In this way, they constitute a simple generative model instantiating an agent's knowledge of its relevant sensorimotor contingencies: they allow the network to predict the incoming stimulation, given the robot's movements.

Recall now that the connections running from (CS)s and (VS)s to (VC)s are not trained, and have opposite weights. This means that the pattern of activity in each (VC) will reflect the difference between current and predicted sensory states, which is just prediction error, computed in the simplest possible way. Prediction error is then forwarded to all (VS)s, enabling them to update their estimate just as PP requires.[19]

[17] Notice that each change in sensory state is always due to the behavior of the robot or, during the learning period, the fact that an experimenter "moved" the robot's body around.

[18] To be clear, (SC)s do not project on (VS)s. Only (VC)s do. But since within each modality each population has the same number of neurons, the (VC) of each modality can mimic the (SC) of that modality.

[19] Notice that albeit here all nets are homogeneously connected (and so there is no hierarchy) PP allows for *horizontal* (i.e. within level) message passing of error, see (Friston, 2008: p. 16). Intriguingly, such an horizontal message-passing is rarely implemented in robotic models inspired by PP, see (Ciria et al., 2021).

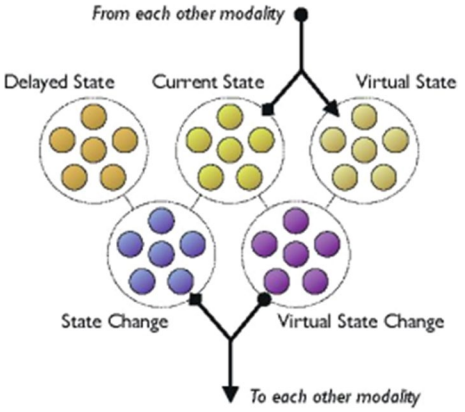

Fig. 1 Implementation of the model at the neural level (one modality). See text for details. © IEEE. Reprinted, with permission, from (Bovet & Pfeiffer, 2005b)

Notice further that in the motor modality, (VS) directly controls the motors. In this way, the robot will move so as to bring about the sensory states the network expects. The robot's behavior is thus driven directly by the network's motor predictions, and indirectly by the ensemble of expected sensory states. This is because the input to the motor (VS) just is prediction error from all other modalities. Thus, the robots will act if, and only if, the network needs to minimize prediction error in *some* modalities, and the robot will act so as to bring about the sensory stimulation the network expects, thereby minimizing prediction error in all modalities.[20] In this way, Bovet's networks qualify as minimal PP systems, able to "actively infer" the sensory states expected in all modalities.

Before I move forward, let me stress that it is *essential* not to confuse networks and robots. Only networks *literally* are PP systems, generating and minimizing prediction errors. And only networks host connections and units exhibiting activation patterns. So, *only networks* are candidate vehicles of generative models. This is important because Bovet describes networks and robots differently. Robots are described distally, in terms of interactions with the environment (e.g. navigating a T-Maze). But networks are described only *proximally*, without any reference to environmental states of affairs. For instance: "The essence of this neural architecture […] is the following. (1) All *signals of the sensors and motors* the robot is equipped with are represented through the activity of artificial neurons." (Bovet, 2007: p. 12, emphasis added). The point motivates Bovet's research: he aims at showing that coherent and intelligent behaviors can be enabled by networks that only learn correlations between the states of the robots' sensors and motors (*ibidem*).[21]

[20] This is because the (VC) in each modality effectively "mimics" the (SC) of that modality. Thus, the activity of (VC)s elicit in motor (VS) a pattern of activity corresponding to the motor state expected, given that change in sensory states. In this way, the robot will act so as to minimize *that* error.

[21] Many thanks to an anonymous reviewer for having advised me to be clearer on this point from the onset.

Now, to see this simple generative model in action, consider the following experiment in which the network enabled a form of "phonotaxis"[22] comparable with that of female crickets (Bovet, 2007: pp. 79–105). When a female cricket hears the song of a conspecific, she turns in the direction of the sound source and approaches the male to mate. The turning behavior of the cricket, however, generates optic flow in the opposite direction[23]; and optic flows tend to trigger the cricket's optomotor response: a simple reflex that tries to correct for the visual flow, re-orienting the cricket in her original position. Clearly, in order for the cricket to reach her mate, her optomotor response needs to be inhibited. Empirical studies suggest that the inhibition is carried out through reafference cancellation: a simple forward model predicts the visual flow caused by the cricket reorientation, and that prediction is used to suppress the optomotor reflex (e.g. Payne et al., 2010; Webb, 2019).

Bovet's experiment was simple. First, he created a network mounted on a "cricket robot", possessing four modalities: an "auditory" modality, a visual modality, a motor modality and a battery level modality, which equipped the robot with a minimal form of viscerception. The network was then trained (by making the robot interact with its environment) so that it could learn the relevant sensorimotor contingencies. Crucially, each time the robot reached the "auditory source", the battery level was increased.

After training, the experimental session began. The network's visceroceptive (VS) was increased; and the mismatch between visceroceptive (CS) and (VS) propagated prediction error. Since increases of battery level highly correlated with certain patterns of activation of the "auditory modality" (recall, the battery level *increased* anytime the robot was in proximity of the "auditory source"), the "auditory" (VS) instantiated those patterns. The mismatch between "auditory" (CS) and (VS) was then propagated to all other modalities. Hence, the network "expected" the patterns of stimulation generated by movements towards the "auditory source": a certain kind of motor activation, and the corresponding optic flow. The error relative to these expectations was then minimized through active inference; that is, by making the robot reach for the "auditory source".

Then, the (VS) of the motor modality was injected with some noise, and the robot's "phonotactic" behavior was tested under two conditions. In the first, the synaptic coupling between motor and visual modality was removed; whereas in the second it was left untouched. In the first condition, the robot was often unable to display the "phonotactic" behavior. This is because the noisy activity in motor (VS) forced the robot to take sudden curves, and, given that the visual and motor modalities were disconnected, the visual modality was unable to predict the corresponding optic flow. This generated visual prediction error, which was propagated in the network, triggering the optomotor reflex, thereby hindering "phonotaxis". The

[22] Due to the robotic hardware employed, "phonotaxis" really was phototaxis (i.e. the sound source really is a light source). This is why "phonotaxis", "auditory modality" and "sound source" will appear under scare quotes in the text.

[23] That is, when the cricket turns left, the optic flow optic flow moves to the right. This is a simple sensorimotor contingence.

competition between "phonotactic" and optomotor behaviors can be seen in (Bovet, 2007: p. 90, figs. 5-7): the robot's trajectories exhibit the zig-zag typical of two competing orienting reflexes. Yet, when the synaptic coupling between motor and visual modalities was re-established, the visual modality was able to predict the incoming optic flow. Thus, no optomotor reflex ensued, and the robot swiftly reached for the "sound source".[24] Hence, the synaptic coupling between visual and motor modality constituted a simple forward model[25]; and, more generally, the connection between various modality constituted a simple generative model, enabling the *network* to predict the incoming input and to make some of those predictions come true through active inference. Notice further that the network qualifies as a genuine forward model, rather than merely as a system exhibiting a simple compensatory bias. In fact, its predictions are targeted to enhance or suppress behaviorally relevant stimulation, are modulated so as to match the incoming feedback and are able to adapt in an experience-dependent manner (see Webb, 2004).[26]

4.2 The network hosts no representational vehicle

It is now possible to check whether the connections or the activity patterns of the network qualify as representational vehicles given the theoretical commitments endorsed by inferentialist and representationalist readings of PP.

Consider first patterns of activity. In the connectionist literature it is standardly assumed that patterns of activity of the *hidden layers* are representational vehicles (e.g. Goodfellow et al., 2016: Ch. 15). But the network has no hidden layers. It is thus doubtful whether we should consider its activity patterns as candidate representational vehicles.[27]

Suppose we should. Are patterns of activity structurally similar to relevant environmental targets? As far as I can see, the answer is in principle positive: structural similarities are cheap to come by and can even be arbitrarily defined (Shea, 2018: pp. 112–113). Hence, it is extremely likely that the patterns of activation of the

[24] Strikingly, a similar synaptic coupling enabling optic flow predictions has been observed in mammalian brains, and it nicely fits a number of theoretical predictions coming from PP, see (Leinweber et al., 2017).

[25] Notice, importantly, that I'm here using the term "forward model" just to denote the fact that such a synaptic coupling allowed the network to predict the sensory consequences of the movements of the robot. I'm not implying that the synaptic coupling estimated the sensory consequences of behavior from motor commands. In fact, there are *no* motor commands in such an architecture, and the robot's behavior is directly controlled by the network's sensory predictions, just as active inference prescribes.

[26] On experience-dependent adaptability, see (Bovet & Pfeiffer, 2005a, b).

[27] As a reviewer noticed, it is intuitive to define hidden layers as layers which are neither input nor output layers. Given this definition, (VS)s seem to be hidden layers: they do not receive inputs from the sensors (so they are not input layers) nor forward output to effectors (so they are not output layers). So, why am I claiming the networks have no hidden layer? Mainly, because this is how Bovet characterizes them: "The network does not contain any so-called 'hidden' layer of inter-neurons" (Bovet 2007: 29). Perhaps it could be argued that *both* the reviewer and Bovet are right: if we focus on *single* modalities, then (VS)s naturally appear as input layers. Yet, when focusing on the *entire* network, (VS)s are more naturally considered as hidden layers. However, as far as I can see, granting (VS)s the status of hidden layers does not impact my argument.

network will turn out to be structurally similar to at least some environmental target. The relevant point is thus whether these structural similarities will be *exploitable*.

Recall: exploitability is the conjunction of two requirements (Shea, 2018: p. 120). First, the system must be systematically sensitive to the relations holding among vehicle constituents. Secondly, the relevant target must be of significance to the system; that is, it must be relevant to the system's task functions: the outputs the system has been stabilized or purposefully designed to produce. As Bovet's networks are artificially designed, the designer dictates their task functions, thereby (partially) determining which structural similarities are exploitable.

However, Bovet defines the function of his networks in squarely *proximal* terms. For instance, he states that (CS)s are, by design, "supposed to" produce a pattern of activity that corresponds to the state of one sensor or motor. As he writes: "In the visual modality for instance, the activity of each neuron corresponds to the *brightness of a pixel in the camera image*" (Bovet, 2006: p. 528, italics added). Similarly, he states (SC)s have been designed to reflect how the *sensory inputs* have changed in a timestep. Equally proximal descriptions are in fact given for each neural population.

It thus seems that, *by design*, the network's task functions target only proximal states, and therefore only proximal states will be of significance to it. But exploitable structural similarities can hold only between candidate vehicles and targets that are of significance to the system. Thus, if exploitable structural similarities are used to determine the content of the candidate vehicles under scrutiny (i.e. patterns of activation), their content can only be proximal. But then the candidate vehicles fail to satisfy distality. Conversely, if we assign candidate vehicles distal targets, they will fail to satisfy exploitability. It thus seems that here candidate vehicles cannot satisfy distality and exploitable structural similarity in conjunction. The same holds if instead of single patterns of activations we focus on the entire activation space (e.g. Churchland, 2012), as focusing on the entire activation space will not change the task functions of the networks. Thus, the entire activation space can bear an *exploitable* structural similarity only to proximal stimuli (or, perhaps more appropriately, the space of possible proximal stimuli). As a result, it fails to satisfy either distality or exploitability just as single activation patterns.

What if, as a reviewer asks, we focus on *the robots*' task functions instead? Since the robots' behavior is distally characterized, it seems legitimate to expect *the robots'* task functions to be distally characterized (i.e. "long-armed") too. That would solve the problem of distality just raised. However, albeit Bovet *describes* the robots' behaviors in distal terms, he *never* assigns distal functions to them. In fact, he explicitly states that his robots *have no purpose* (Bovet, 2007: pp. 4–9). His aim is that of studying: "artificial systems endowed with a self-developing dynamics, yet *without any particular task or motivation*" (*ibidem*: 8, emphasis added). Given that robots are artificial systems, and so their functions are determined by their designer, it seems correct to conclude that Bovet's robots have just no task function, long-armed or otherwise.

Couldn't perhaps the patterns of activation have acquired some distally characterized function through the network's individual learning history? A negative answer seems warranted for two distinct reasons. First, albeit some philosophers do allow

individual learning histories to dictate functions, the scope of the claim is restricted to *supervised* forms of learning involving some sort of feedback (e.g. Dretske 1998; Shea, 2018: pp. 59–62). But Bovet's networks learn in a purely unsupervised manner, and no feedback is involved. Moreover, functions are typically understood as the upshot of processes of selection, in which certain features or traits are *selected over* competing features or traits in virtue of their effects. Hebbian learning, however, is not a process of selection. Hence, it cannot confer functions (Garson, 2012).[28] *Mutatis mutandis*, the same reasoning seems to apply to entire robotic agents.

Maybe we should assign content to single activation patterns in a different way. Wiese (2018: pp. 219–223) has in fact recently suggested a different procedure to do so. In his view, the (generative) model as a whole represents the causal structure of the world in virtue of the exploitable structural similarity holding between the two. However, he adds that the contents of individual patterns of activation should be determined by looking at the statistical dependencies holding between them and their worldly causes. Relying on Eliasmith's theory of content, Wiese suggests that the target of a neuronal response is the set of causally related events upon which the neural response statistically depends the most under all stimulus conditions (see Eliasmith, 2000: p. 34). That is, a neuronal response represents the events that, on average, make its tokening most likely. Does this suggestion allow the candidate vehicles under scrutiny to meet distality and exploitable structural similarity? The answer seems to me negative for two reasons.

First, resorting to Eliasmith's theory of content seems redundant. Wiese (2018: pp. 219–222) intends to use it to assign contents to individual neuronal responses, which he takes to be "proper parts" (i.e. vehicle constituents) of the generative model. He also maintains that the generative model is, as a whole, structurally similar to the causal structure of the world. However, in structural representations, the way in which each vehicle constituent participates to the structural similarity is already *sufficient* to determine its content (Cummins, 1996: p. 96; Shea, 2018: p. 125; Kiefer & Hohwy, 2018: p. 2391). Consider, for instance, a map. As a whole, the map (V) is structurally similar to a target territory (T). This is because V's constituents ($v_a \ldots v_n$) map one to one onto T's constituents ($t_a \ldots t_n$) in a way such that the same *pattern* of spatial relations holds among both ($v_a \ldots v_n$) and ($t_a \ldots t_n$). But if this is the case, then it is entirely correct to say that v_a represents t_a and v_b represents t_b and so on. Since individual vehicle constituents acquire content in virtue of the role they play in the overall structural similarity, there seems to be no need of resorting to Eliasmith's theory of content.

Secondly, suppose that content is assigned to vehicle constituents as Eliasmith's theory of content suggests. Will the contents thus assigned be consistent with the ones assigned by the relevant structural similarity? If yes, then resorting to Eliamith's theory of content adds nothing to what structural similarity already provides. But if not, then there are at least some cases in which a vehicle constituent v_x

[28] Notice also that PP only requires Hebbian forms of learning, see (Bogacz, 2017). Thus, given that Hebbian learning is not a selectionist process, it could be argued that *no* PP system can acquire functions through individual learning.

represents both t_x by structural similarity and t_y by Eliasmith's theory. But then v_x fails determinacy, because its content is disjunctive. In fact, given that v_x represents t_x, its conditions of satisfaction obtain whenever t_x is the case. And, given it *also* represents t_y, its conditions of satisfaction obtain whenever t_y is the case. Hence, v_x will misrepresent if, and only if, both t_x and t_y are not the case. But these are the conditions of satisfaction of a vehicle representing (t_x *or* t_y).

To restore determinacy, one needs to deny either that v_x represents t_x or that it represents t_y. Denying that v_x represents t_y rules out the contribution provided by Eliasmith's theory, which again is left with no role to play. But one cannot rule out that v_x represents t_x either, as that would deny that V, of which v_x is a constituent, is a structural representation. In fact, the statement "*if* V is a structural representation of T, *then* each constituent v_x of V represents the constituent t_x of T onto which it maps" is correct. So, by saying that v_x is not a representation of t_x one denies the consequent of a true statement. But if the consequent of a true statement is false, then the antecedent must be false too. Therefore, if v_x does not represent t_x, then V is not a structural representation of T.[29]

Summarizing: patterns of activation do not seem to bear any exploitable structural similarity to distal targets. Hence, if their content is determined by exploitable structural similarity, then distality does not obtain. Conversely, if their content is not proximal, then their content is not determined by an *exploitable* structural similarity. Appealing to a different content determination procedure appears to deepen the problem. I thus conclude that patterns of activation are not representational vehicles.

Now, what about the connections? As distality has thus far been particularly pressing, it offers a natural starting point: do connections have distal content? The answer seems negative.

To begin with, what should their content be? Connections encode all a network learns (e.g. Rogers and McClelland 2004). But all Bovet's networks learn is to predict the states of the sensors and motors of the robots they control. This seems definitely proximal content. Computationally speaking, connections are also trained in a simple Hebbian fashion. At each time step, the way in which the weight of a connection is modified is provided by a function that takes as arguments patterns of co-activation between the neurons in (CS) and (SC) and the learning rate (see e.g. Bovet, 2007: pp. 26–29). The mathematical content of these connections (i.e. their weight value) is thus exclusively determined by factors lying *inside* the system. If ascriptions of mathematical contents constrain ascriptions of representational contents, it seems that, in these cases, the mathematical contents constrain our ascriptions of representational contents in favor of proximal contents.

These arguments are not conclusive. So, I concede we *might* be able to assign distal contents to connections. But will it be assigned in virtue of an exploitable structural similarity? I believe the answer is again negative. This is because if connections are representations, they are superposed representations. And, given the

[29] An anonymous reviewer raised a challenge to the line of argument developed here. I discuss it in Sect. 5.2 to avoid having to place a long digression here. But the reader can read it now, if they so wish.

standard notion of superpositionality (see Clark, 1993: pp. 17–19; Van Gelder, 1991, 1992), superposed representations cannot be structurally similar to their targets.

Consider the standard definition of superpositionality. The definition is based on a further technical concept, that of a vehicle being conservative over a target (Van Gelder, 1991: p. 43). Bluntly put, a vehicle V is conservative over a target T just in case the minimal set of resources a system needs to leverage in order to represent T equals V. For instance, given the representational resources of natural languages, "John" is conservative over John. To represent John I need, minimally, to token "John". Moreover, "John" has no "representational space" left to represent something over and above John. On the other hand, "John loves Mary" is not conservative over John. To represent John I need not token the entirety of "John loves Mary", and "John loves Mary" has some representational space left to represent something other than John. Superpositionality can then be defined in terms of conservativeness as follows: a vehicle V is a superposed representation of a series of targets $T_a \ldots T_n$ just in case V is conservative over each member of $T_a \ldots T_n$. Notice the plural: superposed representations are always, by definition, conservative over more than one target (Clark, 1993: pp. 17–19; Van Gelder, 1992).

Structural representations, however, can be conservative over one target at most. If V is the vehicle of a structural representation, then there is at least one target T with which V is exploitably structurally similar. This entails that each relevant (i.e. similarity constituting) constituent of V $v_a \ldots v_n$ maps (in an exploitable way) onto *one, and only one*, constituent t_x of T. Now, if this mapping determines the content of each constituent, it seems that each constituent of V entirely "spends its representational credit" to represent one and only one constituent of T. Hence, each constituent of V will be conservative over one, and only one, constituent of T. By the same token, V will be conservative over one, and only one, target T.

Why can't a constituent v_x be conservative over two (or more) constituents t_x and t_y, making V conservative over T and T* (of which t_y is a constituent)? Because it would have to map onto *many*. But (exploitable) structural similarities are defined in terms of *one to one* mappings (see O'Brien & Opie, 2004: p. 11). Thus, it seems correct to say that if a vehicle represents by means of (exploitable) structural similarity, then it is conservative over one, and only one, target. Hence, if a vehicle is not conservative over one, and only one, target, then the vehicle does not represent by means of exploitable structural similarity. But superposed representations are not conservative over one and only one target. Hence, their vehicles fail to satisfy exploitable structural similarity.

Couldn't perhaps the relevant definition of structural similarity be relaxed, so as to allow superposed representations to count as structural representations? Allowing structural similarities to be defined in terms of one-to-many mappings would easily defuse my argument. However, allowing one-to-many mappings makes the content of structural representations disjunctive. In fact, if V is a structural representation of T and v_x maps onto many (e.g. onto both t_x and t_y), it follows that v_x misrepresents only when both t_x and t_y are not the case; and thus that v_x represents (t_x or t_y). Notice that, formally, this is the same problem faced by Wiese's (2018) suggestion on how to assign content to vehicle constituents.

Summarizing: it seems correct to say that connections fail to satisfy distality. And, were that verdict wrong, they would still fail to satisfy exploitable structural similarity. Hence, it seems correct to conclude that, in the networks under scrutiny, connections do not qualify as representational vehicles, given the theoretical commitments of inferentialist and representationalist accounts of PP.

Perhaps my analysis thus far has been unfair. Perhaps it is the network *as a whole* that instantiates the relevant generative model, rather than one of its parts (see e.g. Kiefer & Hohwy, 2018: pp. 2394–2395; Wiese, 2018: p. 219). Albeit I think this is a fair point, I fail to see how it might challenge my conclusion. After all, it seems to me still correct to say that the only things "of significance" to the network, given the task function it has by design, are proximal sensory states. Thus, it seems to me that even conceding, for the sake of discussion, that the network as a whole is, in some sense, exploitably structurally similar to its targets, it would still fail to meet distality.

In this section, I presented the simplest PP system able to perform active inference I know of, and checked whether the candidate vehicles of the relevant generative model (i.e. patterns of activations and connections) actually qualify as vehicles, providing a negative answer. Thus, albeit the network instantiates a simple generative model "knowing" the robot's sensorimotor contingencies, the structures instantiating that model do not qualify as representational vehicles. They are *non-representational structures* instantiating the robot's knowledge of its own sensorimotor contingencies.

Notice that my verdict does not hinge on "weird" metaphysical premises on what counts as a representation. Indeed, the criteria by which I assessed the metaphysical status of generative models are derived from inferentialist and representationalist accounts of PP, which surely provide the mainstream interpretation of the theory.

In the next section, I examine some concerns regarding the verdict here provided.

5 Some worries considered

Here, I examine three distinct concerns regarding the argument I have provided. The first regards the scope of my conclusion: can it be generalized to other PP systems? The second and third worries have been raised by two anonymous reviewers. They concern, respectively, the possibility of multiple assignments of contents and the revisionist implications of my argument. I examine these three worries in turn.

5.1 Will it generalize?

The most obvious objection to the analysis provided above is that its conclusion will not generalize to other PP systems. This is a genuine concern, which I cannot *fully* exorcise here. I will, however, propose a number of arguments and considerations to the effect that my conclusion *is likely to generalize*. To do so, I mainly consider the lines of reasoning that could block the generalization, arguing that none, at present, seems sufficient to block it. Of course, this is not *a proof* that my conclusion

generalizes. To *prove* it, I would probably have to examine *all possible* PP systems, which is clearly unfeasible.

One reason as to why my verdict will not generalize is that Bovet's networks do not resemble standard PP networks, such as Rao and Ballard's (1999) network. The conclusions drawn by looking at Bovet's networks might simply not apply to different PP networks.

Yet, albeit it is surely correct to say that Bovet's networks do not resemble other PP systems, it is worth noting that there is no *standard* PP network. They are all different. For instance, some PP networks do not have distinct error and prediction units (e.g. O'Reilly et al., 2014), and others do not embody distinct sets of ascending and descending connections (e.g. Matsumoto & Tani, 2020). And, when it comes to robotic implementations, there just is no standard PP model, connectionist or otherwise (Ciria et al., 2021). So, if the relevant conclusions drawn from these various models are supposed to generalize (as their modelers surely suppose), why shouldn't the conclusions drawn from *Bovet's* model generalize too?

Perhaps, then, the problem is that Bovet's networks lack an ingredient which, when considered, would force me to revise my verdict. But what could that ingredient be?

Hierarchy is an obvious candidate: Bovet's networks are non hierarchical, whereas the majority (but by no means all, see Tani, 2014; Lanillos & Cheng, 2018) of PP systems are. However, I simply fail to see how hierarchy would force me to revise my verdict. Adding hierarchy means adding hidden layers and connections to (and from) these layers. But these connections would be superposed representations just as the connections of Bovet's network. Thus, if the argument provided above is correct, they would surely fail to satisfy exploitable structural similarity.[30]

Moreover, it is doubtful that the patterns of activity in hierarchically higher layers could be assigned an appropriately distal content (O'Regan & Degenaar, 2014; Dołega, 2017: pp. 12–13). Strictly speaking, all a hierarchically higher level has to predict is what is going on in the layer directly below it, of which it can thus be rightfully said to be a model. Computationally speaking, hierarchically higher levels are said to "produce abstract statistical summaries of the *original visual input*" (Bulow et al., 2016: pp. 5–6; emphasis added; see also Hinton, 2007b; Foster, 2019). Thus, if these computational descriptions are correct, and the mathematical content assigned by these descriptions strongly constrains the ascription of representational content, there are good reasons to expect that the content of hierarchically higher layers to be only proximal.

Notice that sometimes this point seems to be acknowledged even by defenders of the representationalist and inferentialist view of PP. As Orlandi and Lee (2019: pp. 215–217) aptly noticed, it is not uncommon, in the PP literature, to characterize higher levels as models *of the level directly below them*, and to say that they predict

[30] A reviewer noticed I must here make a concession: I cannot exclude that, in a possible connectionist implementation of PP, weighted connections will be structural, rather than superposed, representations. I surely cannot rule out this *possibility*. However, at present, I do not see any *positive* reason to believe that such an exceptional connectionist system will be produced.

the behavior of the level directly below them (e.g. Clark, 2016: pp. 14–24). If these characterizations are correct, they definitely suggest that the content of hierarchically higher layers is only proximal.

But what about the standard account of representations in hidden (i.e. hierarchically higher) layers of artificial neural networks?[31] Careful mathematical analyses conducted on the pattern of activation of those layers often reveal a *structure-preserving* mapping holding between patterns of activation and features of the *distal* domain the network has been trained to operate upon (e.g. Churchland, 2012; Elman, 1991; Shagrir, 2012). This surely seems a hefty consideration in support of a structural-representationalist reading of these patterns. So, were similar patterns found in at least some PP systems, the representationalist-inferentialist reading of PP would be bolstered.

I cannot *in principle* exclude that some form of mathematical analysis on PP models could unravel similar patterns of activation. Yet, I do not believe that it would provide the desired support to the inferentialist and representationalist reading of PP. There are two broad reasons as to why that seems to me the case.

One is that the relevant structure-preserving mapping often holds among *many* patterns of activation (if not the entire activation space, as in Churchland, 2012) and their respective targets. Yet, it seems correct to say that different patterns of activation are *different vehicles*. Hence, the structural similarity would not hold between *a single vehicle* and its target (as it is in the case of structural representations), but rather between an ensemble of vehicles and the ensemble of their respective targets.

The other is that it seems correct to say that, if V is structural representation of T, then changes to V which make it more structurally similar to T increase its representational accuracy. Now, when it comes to artificial neural networks, the changes that increase their representational accuracy surely include changes in the weighted connections. Hence, it seems that weighted connections must be vehicle constituents (or relations thereof) participating in the relevant structural similarity. Yet, in the case at hand, *only* patterns of activation are considered. It thus seems doubtful that a structural similarity defined *purely* in terms of patterns of activation will substantiate the representationalist reading of PP as desired. However, simply "adding connections to the mix" raises the problems with superpositionality described above, marring the relevant structural similarity.

I take this to be only circumstantial evidence in favor of the claim that hierarchically higher layers do not qualify as representations. So my argument is not conclusive: it could be argued that hierarchically higher levels are, as a matter of fact, exploitably similar to some distal target. And that might be done without violating the constraints mathematical contents place upon representational contents. Yet, as far as I know, an argument to that effect has still to be made. As things stand, I only see circumstantial evidence favoring the claim that higher layers do not qualify as representational vehicles. It thus seems that the available evidence favors my anti-representationalist verdict over the representationalist one.

[31] Thanks to an anonymous reviewer for having raised this objection.

A second missing ingredient from Bovet's network is precision. This might be worrisome, as PP suggests that precision plays a key role in enabling active inference (see Brown et al., 2013).

However, I believe that considering precision will not change my verdict. On the one hand, precision is only supposed to modify, in various ways (see Friston, 2012) the relevant patterns of activation to which it is applied. But if, as I argued, these patterns of activation are not representational vehicles in the first place, then any mechanism operating upon them should not be considered a representational mechanism. Moreover, from the computational point of view, precision is typically equated with the inverse variance *of the predicted signal* (Buckley et al., 2017). If, as I've argued, predictions only have proximal content, and the mathematical content of precision signals (i.e. inverse variance) constrains our ascription of representational contents, it then seems we can only ascribe *proximal* contents to precision signals too.

Perhaps the verdict I have provided here will not generalize because I've considered an *artificial* neural network whose task functions have been proximally defined by a human designer, whereas "natural" neural networks implementing PP have long-armed task functions. I think there are reasons to suspect this will not be the case.

To see why, it is important to notice that functions are normative: they are outcomes that a system is *supposed to* produce, in virtue of its design (natural or artificial) or learning history. Task functions (and, more generally, functions) dictate the standards against which to test the performance of a system (e.g. Neander, 2017, Ch. 3). A system can perform *optimally* or *abnormally* only given the standards determined by its functions.

This seems to speak against PP systems having long-armed functions. Consider, for instance, the fact that, on the account PP offers, perceptual illusions are *optimal* percepts (Brown & Friston, 2012). Now, if perceptual illusions are *optimal* percepts, it follows that the machinery producing them (i.e. the PP system) is not *malfunctioning* when a perceptual illusion is produced. But, if this is correct, then it seems that perceptual PP systems do not have long-armed functions. That is, their functions do not appear to be defined in terms of distal states of affairs (e.g. tracking the distal environment, recognizing the external causes of the sensory inputs, etc.). For the output produced by the system here does not match distal states of affairs; hence, were the system's function defined in terms of the latter, the system would have been *malfunctioning*. As a consequence, perceptual illusions would not have been *optimal* percepts.

Moreover, PP systems are often described as *just* in the task of minimizing prediction error (e.g. Friston, 2010; Hohwy, 2015).[32] In fact, the discussion about what PP systems are supposed to do is typically couched in *proximal* terms, such as

[32] Here, I trust neurocomputational modellers (e.g. Spratling, 2017; Tani, 2014) and consider free-energy minimization as a PP algorithm, bracketing the complex relation between the free-energy principle and PP "proper" (see Friston, 2019; Hohwy, 2020).

avoiding sensory states with high surprisal[33] or encountering the sensory states predicted by the model (see Hohwy, 2020). Notice that the purely proximal rendering of what PP systems are supposed to do is no accident: it is actually needed to account for how these systems function in practice. Since PP systems have by assumption[34] access only to *proximal* states, the relevant tasks they are "supposed to" perform must be defined in terms of these states.

As further evidence of the proximal character of what, according to PP, generative models are supposed to do, consider the so-called "dark room" problem (see Sims, 2017 for discussion). The problem is roughly as follows: why, if PP systems are only trying to minimize prediction error, they do not lock themselves in environments delivering extremely predictable stimuli, such as a completely dark room? Notice that such a problem would be immediately dispelled if PP systems were assigned long-armed functions: if PP systems were supposed to, say, find mates to reproduce (rather than just minimize prediction error) it would be immediately clear why they do not end up in dark rooms: there just are no mates there. Notice further that the standard reply to the "dark room" problem is not to concede that PP systems are supposed to do more than minimizing prediction error. Rather, the reply is that "dark room" *sensory states* are prediction-error inducing, given the models possessed by PP systems (Friston et al., 2012c).

All this suggests that, according to PP, all PP systems have to do can be spelled out in *proximal* terms: they have to minimize the error relative to the expected sensory input. But if this is the case, there seems to be little reason to think that "natural" PP systems will be assigned long-armed functions. Thus, there seems to be little reason to think that "natural" PP systems will satisfy *both* distality and exploitable structural similarity in the desired way.[35]

[33] In extremely crude terms, surprisal is an information theoretic quantity (also known as self-information) which captures how improbable a sensory state is, given a model.

[34] This assumption is a corollary of the assumption that sensory states are under-informative in respect to their worldly causes (see Orlandi, 2016; Anderson, 2017 for discussion).

[35] A reviewer wonders whether considering PP systems in the context of the free-energy principle could deliver long-armed functions. The reviewer also points to Hohwy (2013: pp. 179–181; expanded in Hohwy, 2020) as providing some argument to that effect. Now, I cannot introduce the free-energy principle here, but I think I can make a few remarks motivating some skepticism about the free-energy principle providing long-armed functions. The first remark is the following: all the free-energy principle "commands" is to minimize free-energy, which is a quantity *internal* to systems. Indeed, it is precisely because free-energy is internal to systems that free-energy is useful in the first place (see Hohwy, 2020: pp. 5–8). It is thus very hard to see how abiding the free-energy principle would confer long-armed functions. Moreover, insofar PP is the "process theory" by means of which the free-energy principle is abided, to minimize free-energy *just is* to minimize prediction error on average (e.g. Friston, 2009: p. 295; Hohwy, 2013: p. 180). So, it seems that all the free-energy principle "commands" PP systems to do is to minimize prediction error, and this seems to be a proximally defined function. Lastly, it is not clear to me whether the normativity the free-energy principle brings to the table is the normativity of functions in the relevant sense (Hohwy, 2013: p. 181; 2020: pp. 18–20 seems to agree). The relevant normativity of the free-energy principle seems to be based *on the very existence* of free-energy minimizing systems (e.g. "Rather, the FEP's conceptual analysis allows us to see how existence (analysed as self-organization) is *at the same time both causal and normative*", Hohwy, 2020: p. 16, emphasis added). A system failing to abide by this kind of normativity, would simply "fail to exist" as a system. But, given the relevant notion of function under consideration, something can exist, and yet systematically *fail* to perform its function

The verdict I provided could also be challenged arguing that Bovet's networks enable the robotic agents to perform only very "low level" sensorimotor coordinations with the surrounding environment. Had I considered different (and, plausibly, more complex) networks, enabling "representation hungry" tasks requiring coordination with non-present targets, my verdict would have been different, as coordinating with non-present targets *requires* representations to be in place (Clark & Toribio, 1994). A related worry is that the robots guided by Bovet's network might be "merely reactive": they just *respond* to the incoming environmental inputs. Many hold that merely reactive behaviors do not require any internal model, whereas proactive, self-generated behaviors do (e.g. Pezzulo, 2008; Tani, 2007). Both objections share a common theme; namely, that the networks upon which my verdict is based is *too simple of a model* to allow my verdict to generalize to more complex PP systems. However, I believe these objections are misguided in at least two respects.

To begin with, it is, as a matter of fact, *false* that the robots guided by Bovet's network are merely reactive systems enabling only low level sensorimotor coordinations. As a matter of fact, in numerous experiments (e.g. Bovet, 2006; Bovet & Pfeiffer, 2005a, b) the robot *self-initiated* its own behavioral routines, because the network was expecting sensory inputs that the environment did not deliver, thereby triggering active inference. Moreover, the network architecture Bovet engineered is capable of delayed reward learning in the context of T-maze tasks (Bongard & Pfeiffer, 2005a; b). In such tasks, however, agent and target (reward[36]) are *not* immediately coupled, and so, from the point of view of the robotic agent, the target is absent at the start of each trial. Furthermore, delayed reward learning is supposed to require some form of *working memory*, which is needed to correctly associate cue, motor decision and outcome (e.g. Kim, 2004; see also Carvalho & Nolfi, 2016). Thus, it seems to me correct to say that delayed reward tasks in a T-maze setting are a *bona fide* instance of "representation hungry" cognition. Nevertheless, that non-representational network managed to solve the task with a high degree of accuracy, only by learning a set of relevant sensorimotor associations. More precisely (but see Bovet & Pfeiffer, 2005a, b and Bovet, 2007: pp. 123–153 for the full account), the network enabled the robot to solve the task only by learning to *predict shifts of visual flow* conditioned on the activity of tactile sensors stimulated by the cue. The mismatch between expected and actually received visual flow was then minimized through active inference, thus making the robot turn so as to bring about the

Footnote 35 (continued)
(Millikan, 1989). To use a well-known example: malformed hearts are *supposed to* pump blood, but they always fail to do so.

[36] A PP enthusiast might question my use of the word "reward" in this context, as active inference does not, strictly speaking, posit rewards (see Friston et al., 2012a, b). It is thus worth noting that Bovet himself acknowledges that "reward" and "punishment" are *arbitrary* tags, which he uses to simplify the discussion. The "reward" modality of the net really only tracks the state of the robot batteries, and the reward itself is a reduction of prediction error between the predicted and actually sensed state of the batteries (see Bovet and Pfeiffer 2005a, b). Notice further that, in Bovet's architecture, a "reward" *only* aligns expected and actually sensed battery states. Hence, "rewards" *just are* highly predictable sensory states, exactly as PP suggests.

expected visual flow. But by turning, the robot also entered in the correct arm of the T-maze, thus "stumbling upon" the reward.[37]

Secondly, I do not think that these objections can be rightfully formulated within the theoretical framework of PP, at least if really PP offers "a cognitive package deal" able to account with the same set of resources for cognition *in general* (Clark, 2016; Pezzulo, 2017; Spratling, 2016). If really PP can account for all cognitive phenomena using the same set of resources functioning in the same way, then it seems to me that representationalism or anti-representationalism should be valid across the board. If these resources are representational resources, then it seems that they will be representational even when they are enabling simple sensorimotor interactions with a present target. And, if, as I've argued, these resources do not qualify as representational, then they will not be representational even when the target they are enabling a system to proactively coordinate with is absent.

5.2 Can the "two-level attribution" save representationalism?

My argument against Wiese's (2018) appeal to Eliasmith's theory of content *presupposes* that vehicles can be assigned contents in only one way. But what if, as a reviewer asks, vehicles could be assigned *multiple* contents according to *multiple* theories of content, based on one's explanatory focus? For instance, if one's focus is centered on the inner workings of Bovet's network, it might be appropriate to assign it only proximal contents via exploitable structural similarity. But if one's explanatory focus is how the entire robot interacts with the environment, it might be appropriate to assign it distal content resorting to Eliasmith's theory of content (or vice versa). Given that contents thus attributed sit at different explanatory levels, they *need not* be mutually exclusive. Such a "two-level attribution"[38] of content can thus allow us to follow Wiese's suggestion, without *thereby* inviting the problems I raised before. How can I respond?

To start, I wish to point out an ambiguity. Talking of "assigning content" is ambiguous between two readings. On a first reading, content assignments are *mere* ascriptions of content: given our explanatory aims, we speak of a vehicle as if it represents something, but as a matter of fact the vehicle *does not* represent that thing. This seems a form of content pragmatism (Mollo, 2020: p. 109). On a second reading, content assignments are not *mere* ascriptions: the vehicle *really has* multiple contents, perhaps in virtue of the fact that it satisfies multiple content-determining relations with multiple targets. Our explanatory interests only select, among the many contents a vehicle *really and objectively* bears, the one that best serves our explanatory needs.

Now, which is the intended reading of the "two-level attribution" the reviewer suggests? I think the second one. The reviewer is presumably trying to rescue

[37] In this way, it seems to me that Bovet's systems provide some empirical support to the enactivists' claim that complex non-representational structures instantiating sensorimotor knowledge are sufficient for "higher"/"representation hungry" cognition (Bruineberg et al., 2019; Kiverstein & Rietveld, 2018).

[38] The phrase has been coined by the anonymous reviewer.

Wiese's (2018) account, which espouses realism about content (as all the inferentialist and representationalist readings of PP do). Moreover, it could be argued that inferentialist and representationalists accounts of PP *already* ascribe multiple contents to vehicles: they do accept that a vehicle has both *mathematical* and *representational* contents. Isn't this a "two-level attribution" of the kind the reviewer suggests?[39]

Yet, I see a problem with the "two-level attribution" thus interpreted. It can be exposed by means of a simple example. Suppose V satisfies (at the same time) the conditions spelled out by two theories of content C and C*. According to C, V represents T; whereas it represents T* according to C*. Accept the "two-level attribution" as sketched above: V *really and objectively* represents T *as well as* T*. Thus V has *two* contents, and we are free to "pick one" based on our explanatory needs.

Now, V is a representational vehicle *objectively bearing* some content. So, there are some tokenings of V which *objectively are* misrepresentations—but which ones? I think there are only three possible cases:

(a) A tokening of V is a misrepresentation when T, and only T, is not the case (*mutatis mutandis* for T*)
(b) A tokening of V is a misrepresentation when at least one among T and T* is not the case
(c) A tokening of V is a misrepresentation when both T and T* are not the case

If (a) is accepted, then it seems that V represents only T (or only T*). It's accuracy conditions are sensitive only to Ts, just as those of a vehicle representing *only* Ts, and thus having *only one* content, determined only by C (or C*).

If (b) is accepted, then V appears to be representing (T *and* T*). In fact, a vehicle misrepresenting when T or T* are not the case *just is* a vehicle representing (T *and* T*). But then it seems that V has a *single* "conjunctive" content, determined by neither C nor C*.

If (c) is accepted, then V appears to represent (T or T*). A vehicle misrepresenting only when both T and T* are not the case *just is* a vehicle representing (T or T*). But then, again, V seems to have a *single* disjunctive content, determined by neither C nor C*.

So, it seems that, in all cases,[40] the "two-level attribution" view entails that V does not have *multiple* contents, but only a single (perhaps disjunctive or "conjunctive")

[39] Perhaps it is, but an important difference should be noticed. Mathematical and representational contents are different *kinds* of content (Egan, 2014: p. 118). One is narrow, the other is (typically) wide. One is determined by the computations a system performs, the other by some privileged naturalistic relation holding between vehicles and targets. But the "two level attribution" the reviewer proposes assigns different contents of the *same* kind (representational) to the same vehicle.

[40] A reader might wonder why I have not considered option (b) when considering Wiese's proposal. The answer (embarrassingly) is that I had not noticed its viability when the manuscript was first conceived. Noticing the presence of option (b), however, does not solve the problems with determinacy Wiese's proposal suffers from. Indeed, it seems to me that it makes them *harder* to solve. For now it is unclear whether following Wiese's suggestion delivers us vehicles representing (T *or* T*) or (T *and* T*).

content. Moreover, in two cases out of three, that content is *not* determined by *any* of the theories of content accepted (C and C*). This seems to put these theories under pressure, as it suggests that those theories inadequately capture the content that representational vehicles bear. A defender of the "two-level attribution" view might object that content is as a matter of fact determined in a way that it is *only partially* captured by C and C*, and that only by wielding them together we understand what vehicles really represent. But why then shouldn't we resort to a third theory C** "mashing up" C and C*? Indeed, if either option (b) or (c) is accepted, C** looks desirable: it would be the *single* theory of content capturing the *single* ("conjunctive" or disjunctive) content possessed by vehicles.

Now, the above is too quick of a discussion for me to declare that the "two-level attribution" view is untenable. Its defenders might convincingly reply to my quick argument. At present, however, the "two-level attribution" view does not really seem viable.[41]

5.3 Radical revisionism?

A different reviewer asks how the anti-representationalism advocated here squares with the representationalism of cognitive science. Am I committed to a strong form of revisionism? I clearly cannot reply in full here. Yet, I can quickly motivate a negative answer.

To start, notice that I (as any other anti-representationalist) am committed to *some* form of revisionism. Cognitive science really seems strongly committed to representationalism. Arguing that certain structures are not representations or that no such commitment is present (e.g. Ramsey, 2007) *is* a form of revisionism: at least one ontological commitment of cognitive science should be modified. *H*ow *radical* should the revision be? There are, I think, two reasons as for why the anti-representationalism defended here does not seem to have *radically* revisionist implications.

First, the anti-representationalist conclusion has been motivated using an artificial neural network; and artificial neural networks surely are central in the current empirical practice of cognitive science. The form of anti-representationalism I'm arguing for stems from cognitive science as it is currently practiced, rather than some *alternative* research program developing *alternative* empirical methods and research practices (e.g. Chemero, 2009). So, *at least prima facie*, the form of anti-representationalism argued for here does not invite a radical departure from the current epistemic routines of cognitive science.

Secondly, in the context of the free-energy principle, generative models have *already* been characterized as *non-representational* structures mediating agent-environment interactions (Bruineberg & Rietveld, 2014; Ramstead et al., 2020a). Indeed, it could be argued that this sense of "model" is the *core* sense of model in the free-energy framework, for, according to it, models, in the relevant sense, just are *controllers* (e.g. Seth, 2015: pp. 6–8). So, in a way, my argument only extends

[41] I'm also leaving the possibility of reading the "two level attribution" as a form of content pragmatism undiscussed. Owing to space limitations, I cannot address content pragmatism here.

an already existing conceptual characterization of generative models from the free-energy principle to PP proper. Surely this isn't a conceptual revolution.

But what about the term "model" itself (as well as other representational terms)? Should we police our language so as to systematically avoid them? I think the answer is negative. As hinted above, "model" has a *technical* meaning, which does not align with the philosophically loaded meaning of models *as structural representations*; or so, at least, I'm suggesting. But once the point has been made, I see no strong reason to systematically police our language so as to erase any occurrence of "model".

6 Concluding remarks

In this essay I have argued that, given the theoretical commitments of representationalist and inferentialist accounts of PP, the structures instantiating generative models do not appear to qualify as representational vehicles. The physical realizers of generative models seem to be just non-representational structures instantiating an agent's knowledge of sensorimotor contingencies. So, if the theoretical commitments of inferentialist-representationalist readings of PP are correct, then PP does not seem to qualify as a representationalist theory of cognition. And, if, as these views hold, PP really explains *all* aspects of our cognitive lives, then it seems that PP invites a form of global anti-representationalism about cognition.

Contra (Gładziejewski, 2016), PP might be as *anti-representationalist* as cognitive science can possibly get.

Acknowledgements This paper owes much to Elmarie Venter, Bartosz Radomski, Tobias Schlicht, Tobias Starzak, Nina Poth, François Kammerer, Adrian Downey and Krys Dołega, who extensively commented on a previous version of the essay. Thanks also to Bruno Cortesi, Arianna Beghetto, and Giacomo Zanotti for their nice comments on a very early draft. I also wish to thank the two anonymous referees, who provided extensive feedback which immensely improved the essay.

Authors' 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

Conflict of interest The author declares that have no conflict of interest.

References

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

Anderson, M. L. (2017). Of Bayes and bullets. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*. (Vol. 4). Frankfurt am Main: The MIND Group. https://doi.org/10.15502/9783958573055

Artiga, M., & Sebastián, M. A. (2018). Informational theories of content and mental representation. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-018-0408-1

Baltieri, M., & Buckley, C. (2019). Generative models as parsimonious descriptions of sensorimotor loops. *Behavioral and Brain Sciences*. https://doi.org/10.1017/s0140525x19001353

Blackmore, S., Frith, C., & Wolpert, D. (1999). Spatiotemporal prediction modulates the perception of self-produced stimuli. *Journal of Cognitive Neuroscience, 11*(5), 551–559.

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

Bovet, S., & Pfeiffer, R. (2005a). Emergence of coherent behaviors from homogeneous sensorimotor coupling. In *ICAR '05: Proceedings of the 12th international conference in advanced robotics*. Seattle, WA.: IEEE. https://doi.org/10.1109/ICAR.2005.1507431

Bovet, S., & Pfeiffer, R. (2005b). Emergence of delayed reward learning from sensorimotor coordination. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, https://doi.org/10.1109/IROS.2005.1545085.

Bovet, S. (2006). Emergence of insect navigation strategies from homogeneous sensorimotor coupling. In *Proceedings of the 9th international conference on intelligent autonomous systems (IAS 9)* (pp. 525–533). Tokyo.

Bovet, S. (2007). *Robots with Self-Developing Brains*. Ph.D. dissertation, University of Zurich (CH).

Brette, R. (2016). Subjective physics. In A. El Haldy (Ed.), *Closed loop neuroscience* (pp. 145–169). Elsevier.

Brown, H., et al. (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing, 14*(4), 411–427.

Brown, H., & Friston, K. (2012). Free-energy and illusions: The Cornsweet effect. *Frontiers in Psychology, 3*, 43.

Bruineberg, J., Chemero, A., & Rietveld, E. (2019). General ecological information supports engagement with affordances for "higher" cognition. *Synthese, 196*(12), 5231–5251.

Bruineberg, J., & Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience, 8*, 599.

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

Bulow, P., et al. (2016). Concepts as semantic pointers: A framework and computational model. *Cognitive Science, 40*(5), 1128–1162.

Carvalho, J. T., & Nolfi, S. (2016). Cognitive offloading does not prevent but rather promotes cognitive development. *PLoS ONE, 11*(8), e0160679.

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

Churchland, P. M. (2012). *Plato's camera*. The MIT Press.

Ciria, A., et al. (2021). Predictive processing in cognitive robotics: A review. *Neural Computation, 33*(6), 1402–1432.

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. (2015a). Radical predictive processing. *The Southern Journal of Philosophy, 53*, 3–27.

Clark, A. (2015b). Embodied prediction. In T. Metzinger & J. Windt (Eds.), *Open MIND: 7(T)*. Frankfurt am Main: The MIND Group. https://doi.org/10.15502/9783958570115

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

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

Clark, A., & Toribio, J. (1994). Doing without representing? *Synthese, 101*(3), 401–431.

Cummins, R. (1996). *Representations, targets and attitudes*. The MIT Press.

Dołega, 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/10.15502/9783958573116

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

Dretske, F. (1986). Misrepresentation. In R. Bodgan (Ed.), *Belief: Form, content and function* (pp. 17–36). Oxford University Press.

Dretske, F. (1998). *Explaining behavior*. The MIT Press.

Egan, F. (2012). Representationalism. In E. Margolis, S. Samuels, & P. Stich (Eds.), *The Oxford handbook of philosophy of cognitive science* (pp. 250–272). Oxford University Press.
Egan, F. (2014). How to think about mental content. *Philosophical Studies, 170*(1), 115–135.
Egan, F. (2018). The nature and function of content in computational models. In M. Sprevak & M. Colombo (Eds.), *The Routledge Handbook of the computational mind* (pp. 247–258). Routledge.
Eliasmith, C. (2000). *How neurons mean: A neurocomputational theory of representational content*. Ph.D. Dissertation: Washington University in St. Louis, MO.
Elman, J. (1991). Distributed representations, simple recurrent neural networks and grammatical structure. *Machine Learning, 7*, 195–225.
Feldman, H., & Friston, K. (2010). Attention, uncertainty and free-energy. *Frontiers in Human Neuroscience, 4*, 105.
Fodor, J. (1987). *Psychosemantics*. The MIT Press.
Fodor, J. (1990). *A theory of content and other essays*. The MIT Press.
Foster, D. (2019). *Generative deep learning*. Sebastopol, CA.: O'Reilly.
Franklin, D. W., & Wolpert, D. M. (2011). Computational mechanisms for sensorimotor control. *Neuron, 72*(3), 425–442.
Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences, 360*(1456), 815–836.
Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology, 4*(11), e1000211.
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 optimal motor control? *Neuron, 72*(3), 488–498.
Friston, K. (2012). Predictive coding, precision and synchrony. *Cognitive Neuroscience, 3*(3–4), 238–239.
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., Adams, R., & Montague, R. (2012a). What is value - accumulated reward or evidence? *Frontiers in Neurorobotics, 6*, 11.
Friston, K., Samothrakis, S., & Montague, R. (2012b). Active inference and agency: Optimal control without cost functions. *Biological Cybernetics, 106*(8–9), 523–541.
Friston, K., Thornton, C., & Clark, A. (2012c). Free energy minimization and the dark-room problem. *Frontiers in Psychology, 3*, 130.
Garson, J. (2012). Function, selection and construction in the brain. *Synthese, 189*(3), 451–481.
Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese, 193*(2), 559–582.
Gładziejewski, P. (2017). Just how conservative is conservative predictive processing? *Internetowy Magazyn Filozoficzny Hybris, 38*, 98–122.
Godfrey-Smith, P. (1989). Misinformation. *Canadian Journal of Philosophy, 19*(4), 533–550.
Goodfellow, I., et al. (2016). *Deep learning* (Vol. I & II). Cambridge, MA.: The MIT Press.
Goodman, N. (1969). *Languages of art*. Oxford University Press.
Hay, N. *et al*. (2018). Behavior is everything: Towards representing concepts with sensorimotor contingencies. In *32nd AAAI conference on artificial intelligence* (pp. 2–7). New Orleans, LA.
Hemion, N. J. (2016). Discovering latent states for model learning: applying sensorimotor contingencies theory and predictive processing to model context.
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.
Hohwy, J. (2013). *The predictive mind*. Oxford University Press.
Hohwy, J. (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.), *Open MIND, 19(T)*. Frankfurt am Main: The MIND Group. https://doi.org/10.15502/9783958570016
Hohwy, J. (2016). The self-evidencing brain. *Noûs, 50*(2), 259–285.
Hohwy, J. (2020). Self-supervision, normativity, and the free energy principle. *Synthese*. https://doi.org/10.1007/s11229-020-02622-2
Huang, Y., & Rao, R. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*(5), 580–593.

Hutto, D., & Myin, E. (2020). Deflating deflations about mental representations. In J. Smortchkova, K. Dolega, & T. Schlicht (Eds.), *What are mental representations?* (pp. 79–100). Oxford University Press.

Iida, F., & Bovet, S. (2009). Learning legged locomotion. In A. Adamatzky & M. Komosinski (Eds.), *Artificial life models in hardware* (pp. 21–33). Springer.

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/10.15502/9783958573185

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.

Kim, D. E. (2004). Evolving internal memory for T-maze tasks in noisy environments. *Connection Science, 16*(3), 183–210.

Kiverstein, J. D., & Rietveld, E. (2018). Reconceiving representation-hungry cognition: An ecological-enactive proposal. *Adaptive Behavior, 26*(4), 147–163.

Laflaquiere, A. (2017). Grounding the experience of a visual field through sensorimotor contingencies. *Neurocomputing, 268*, 142–152.

Lanillos, P., & Cheng, G. (2018). Adaptive robot body learning and estimation through predictive coding. In *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. https://doi.org/10.1109/IROS.2018.8593684

Le Hir, N., et al. (2018). Identification of invariant sensorimotor structures as a prerequisite for the discovery of objects. *Frontiers in Robotics and AI, 5*, 70.

Leinweber, M., et al. (2017). A sensorimotor circuit in the mouse cortex for visual flow prediction. *Neuron, 95*(6), 1420–1432.

Matsumoto, T., & Tani, J. (2020). Goal-directed planning for habituated agents by active inference using a variational recurrent neural network. *Entropy, 22*(5), 564.

Maye, A., & Engel, A. K. (2013). Extending sensorimotor contingency theory: Prediction, planning, and action generation. *Adaptive Behavior, 21*(6), 423–436.

McLendon, H. J. (1955). Uses of similarity of structure in contemporary philosophy. *Mind, 64*(253), 79–95.

Millikan, R. G. (1989). In defense of proper functions. *Philosophy of Science, 56*(2), 288–302.

Mollo, D. C. (2020). Content pragmatism defended. *Topoi, 39*(1), 103–113.

Namikawa, J., et al. (2011). A neurodynamic account of spontaneous behavior. *PLoS Computational Biology, 7*(10), e1002221.

Neander, K. (2017). *A mark of the mental*. The MIT Press.

O'Regan, J. K. (2011). *Why doesn't red sounds like a bell?* Oxford University Press.

O'Regan, J. K., & Degenaar, J. (2014). Predictive processing, perceptual presence, and sensorimotor theory. *Cognitive Neuroscience, 5*(2), 130–131.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*(5), 939–973.

O'Reilly, R. C., *et al.* (2014). Learning through time in the thalamocortical loops.

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 representation* (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.

Orlandi, N. (2018). Predictive perceptual systems. *Synthese, 195*(6), 2367–2386.

Orlandi, N., & Lee, G. (2019). How radical is predictive processing? In M. Colombo, E. Irvine, & M. Stapleton (Eds.), *Andy Clark and his critics* (pp. 206–221). Oxford University Press.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology* (Vol. I). The MIT Press.

Payne, M., Hedwig, H., & Webb, B. (2010). Multimodal predictive control in crickets. In S. Doncieux, B. Girard, A. Guillot, J. Hallam, J.-A. Meyer, & J.-B. Mouret (Eds.), *From animals to animats 11* (pp. 167–177). Berlin: Springer.

Pezzulo, G. (2008). Coordinating with the future: The anticipatory nature of representation. *Minds and Machines, 18*(2), 179–225.

Pezzulo, G. (2011). Grounding procedural and declarative knowledge in sensorimotor anticipation. *Mind and Language, 26*(1), 78–114.
Pezzulo, G., et al. (2017). Model-based approaches to active perception and control. *Entropy, 19*(6), 266.
Pezzulo, G. (2017). Tracing the roots of cognition in predictive processing. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing: 20.* Frankfurt am Main: The MIND Group. https://doi.org/10.15502/9783958573205
Pickering, M. J., & Clark, A. (2014). Getting ahead: Forward models and their place in cognitive architecture. *Trends in Cognitive Sciences, 18*(9), 451–456.
Pio-Lopez, L., et al. (2016). Active inference and robot control: A case study. *Journal of the Royal Society Interface, 13*(122), 20160616.
Ramsey, W. (2007). *Representation reconsidered*. Cambridge University Press.
Ramstead, M. J. D., Friston, K., & Hipòlito, I. (2020b). Is the free-energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy, 22*(8), 889.
Ramstead, M. J. D., Kirchhoff, M., & Friston, K. (2020a). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior, 28*(4), 225–239.
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.
Roche, W., & Sober, E. (2019). Disjunction and distality: The hard problem for purely probabilistic causal theories of mental content. *Synthese*. https://doi.org/10.1007/s11229-019-02516-y
Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition*. Cambridge, MA: The MIT Press.
Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience, 5*(2), 97–118.
Seth, A. K. (2015). The Cybernetic Bayesian Brain. In T. Metzinger & W. Wiese (Eds.), *Open MIND: 35T.* Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958570108
Seth, A. K., & Friston, K. (2016). Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1708), 2016007.
Shagrir, O. (2012). Structural representations and the brain. *The British Journal of Philosophy of Science, 63*(3), 519–545.
Shea, N. (2018). *Representation in Cognitive Science*. Oxford University Press.
Sims, A. (2017). The problems with prediction: the dark room problem and the scope dispute. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing: 23.* Frankfurt am Main: The MIND Group. https://doi.org/10.15502/9783958573246
Sims, M., & Pezzulo, G. (2021). Modeling ourselves: What the free-energy principle reveals about our implicit notions of representation. *Synthese*. https://doi.org/10.1007/s11229-021-03140-5
Spratling, M. W. (2016). Predictive coding as a model of cognition. *Cognitive Processing, 17*(3), 279–305.
Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition, 112*, 92–97.
Tani, J. (2007). On the interactions between top-down anticipation and bottom-up regression. *Frontiers in Neurorobotics, 1*, 2.
Tani, J. (2014). Self-organization and compositionality in cognitive brains: A neurorobotics study. *Proceedings of the IEEE, 102*(4), 586–605.
Tani, J. (2016). *Exploring robotic minds*. Oxford University Press.
Tschantz, , et al. (2020). Learning action oriented models through active inference. *PLoS Computational Biology, 16*(4), e1007805.
Van Gelder, T. (1991). What is the "D" in "PDP"? A survey of the concept of distribution. In W. Ramsey, S. Stich, & D. Rumelhart (Eds.), *Philosophy and Connectionist Theory* (pp. 33–60). Rutledge.
Van Gelder, T. (1992). Defining distributed representations. *Connection Science, 4*(3–4), 175–191.
Vásquez, M. J. C. (2019). A match made in heaven: Predictive approaches to (an unhortodox) sensorimotor enactivism. *Phenomenology and the Cognitive Sciences*. https://doi.org/10.1007/s11097-019-09647-0
Webb, B. (2004). Neural mechanism for prediction: Do insects have forward models? *Trends in Neurosciences, 27*(5), 278–282.
Webb, B. (2019). The minds of insects. In M. Colombo, E. Irvine, & M. Stapleton (Eds.), *Andy Clark and his critics* (pp. 254–265). Oxford University Press.
Wiese, W. (2016). What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences, 16*(4), 715–736.
Wiese, W. (2018). *Experienced wholeness*. The MIT Press.

Wiese, W., & Metzinger, T. (2017). Vanilla PP for philosophers. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing: 1*. Frankfurt am Main: The MIND Group.
Williams, D. (2018). Predictive processing and the representation wars. *Minds and Machines, 28*(1), 141–172.
Williams, D. (2020). Predictive coding and thought. *Synthese, 197*(4), 1749–1775.