PH.D. PROGRAM IN COGNITIVE NEUROSCIENCE AND PHILOSOPHY OF MIND

Scuola Universitaria Superiore Pavia (IUSS)

THE METAPHYSICS OF PREDICTIVE PROCESSING

A NON-REPRESENTATIONAL ACCOUNT

Doctoral Dissertation

Candidate: Marco Facchin (XXIV cycle)

Supervisor: Prof. Piredda Giulia
Co-Supervisor: Prof. di Francesco Michele

ACADEMIC YEAR: 2020/2021

Declarations

This dissertation expands upon some of the author's previously published works. More in detail:

> **Ch. 2 §2.2** reproduces material from Facchin, M. (2021a). Predictive processing and anti-representationalism. *Synthese*, *199*(3-4), 11609-11604.
>
> **Ch. 3 §4.1** reproduces material from Facchin, M. (2021a). Predictive processing and anti-representationalism. *Synthese*, *199*(3-4), 11609-11604.
>
> **Ch. 4** reproduces material from Facchin, M. (2021b). Are generative models structural representations?. *Minds and Machines*, *31*, 277-303.
>
> **Ch. 5** reproduces material from Facchin, M. (2021c). Structural representations do not meet the job description challenge. *Synthese*, *199*(3-4), 5479-5508.
>
> **Ch. 6** reproduces material from Facchin, M. (2021a). Predictive processing and anti-representationalism. *Synthese*, *199*(3-4), 11609-11604.

Thanks to the editors, presses and copyright holders for permission to freely reproduce my own work.

Figure 5 (p. 182) is a special case. The image originally appeared in (Bovet and Pfeiffer 2005b), and the rights belong to IEE. I have re-used the image with permission in (Facchin 2021a) and I'm reproducing it here in accordance with IEEE copyright standards.

Acknowledgements

Jelle Bruineberg's Ph.D. thesis opens asserting that everyone able to start a Ph.D. is able to finish one. I disagree - I was not able to finish mine; at least, not *alone*. So, thank you to: Elmarie Venter, Thomas van Es, Zuzanna Rucinzka, Luca Roccioletti, Andrea Raimondi, Bartosz Radomski, Matt Sims, Tobias Schlicht, Tobias Starzak, Nina Poth, Niccolò Negro, Erik Myin, François Kammerer, Adrian Downey, Krys Dołęga, Giulia di Rienzo, Bruno Cortesi, Silvia Bianchi, Francesco Beltrame and Arianna Beghetto for your feedback on my work (and the occasional kick in the butt). This dissertation would not exist if it weren't for you.

A special, *bell-sounding*, thanks goes to Marco Viola and Elia Zanin: writing together has been a lot of fun and we should definitely do it again - say, as soon as I manage to find a postdoc position.

A special thanks to Giacomo - I really cannot fathom *just how* you managed to put up with my "bizarre way of going about things" these three years. But you did, and I'm sincerely thankful for it.

A last, heartfelt thank you goes to Davide, Davide, Bea, and Fabri. There are no words to express my gratitude, so you'll have to try to imagine it. And yes, it is like that, *but a million times bigger.* Thank you.

Table of Contents

Index of Figures

Abstract

This dissertation focuses on generative models in the Predictive Processing framework. It is commonly accepted that generative models are structural representations; i.e. physical particulars representing *via* structural similarity. Here, I argue this widespread account is wrong: when closely scrutinized, generative models appear to be non-representational control structures realizing an agent's sensorimotor skills.

The dissertation opens (Ch.1) introducing the Predictive Processing account of perception and action, and presenting some of its connectionist implementations, thereby clarifying the role generative models play in Predictive Processing.

Subsequently, I introduce the conceptual framework guiding the research (ch.2). I briefly elucidate the metaphysics of representations, emphasizing the specific functional role played by representational vehicles within the systems of which they are part. I close the first half of the dissertation (Ch.3) introducing the claim that generative models are structural representations, and defending it from intuitive but inconclusive objections.

I then move to the second half of the dissertation, switching from exposition to criticism. First (Ch.4), I claim that the argument allegedly establishing that generative models are structural representations is flawed beyond repair, for it fails to establish generative models are structurally similar to their targets. I then consider alternative ways to establish that structural similarity, showing they all either fail or violate some other condition individuating structural representations.

I further argue (Ch.5) that the claim that generative models are structural representations would not be warranted even if the desired structural similarity were established. For, even if generative models were to satisfy the relevant definition of structural representation, it would still be wrong to consider them as representations. This is because, as currently defined,

structural representations fail to play the relevant functional role of representations, and thus cannot be rightfully identified as representations in the first place.

This conclusion prompts a direct examination of generative models, to determine their nature (Ch.6). I thus analyze the simplest generative model I know of: a neural network functioning as a robotic "brain" and allowing different robotic creatures to swiftly and intelligently interact with their environments. I clarify how these networks allow the robots to acquire and exert the relevant sensorimotor abilities needed to solve the various cognitive tasks the robots are faced with, and then argue that neither the entire architecture nor any of its parts can possibly qualify as representational vehicles. In this way, the structures implementing generative models are revealed to be non-representational structures that instantiate an agent's relevant sensorimotor skills. I show that my conclusion generalizes beyond the simple example I considered, arguing that adding computational ingredients to the architecture, or considering altogether different implementations of generative models, will in no way force a revision of my verdict. I further consider and allay a number of theoretical worries that it might generate, and then briefly conclude the dissertation.

Introduction

This dissertation focuses on the metaphysical status of generative models within Predictive Processing - a neurocomputational framework of increasing popularity. To anticipate, I argue, *contra* the prevalent structural-representationalist interpretation of Predictive Processing, that generative models are not structural representations. More in detail, I'm going to claim that they are not even *representations*, as, on a closer scrutiny, generative models are revealed to be nothing more than non-representational control structures that instantiate an agent's sensorimotor mastery.

Such a claim requires some setup to be expressed and defended properly. Hence, the dissertation is divided in two parts: one provides the setup, the other articulates my claim.

Part I: Predictive Processing, representations, and Predictive Processing and representations

Chapter 1 introduces the framework of Predictive Processing. The introduction will be reader-friendly, with little mathematical notation and many intuitive examples. Since the aim of the chapter is introductory, I bracket all the philosophical issues surrounding Predictive Processing, representationalism included. The latter will be thoroughly discussed throughout the rest of the dissertation.

The chapter is structured as follows. The first paragraph provides a bird's eye view of Predictive Processing. The second paragraph presents the core machinery described by Predictive Processing *twice*, first in homuncular terms, and then in a properly de-homuncularized connectionist fashion. Paragraph three and four expand the presentation of Predictive Processing covering expected precision/attention and active inference/action respectively. A fifth conclusive paragraph closes the chapter.

Chapter 2 provides the conceptual background needed to evaluate the representational status

of generative models. I introduce the issues surrounding representations in cognitive science, clarifying that the scope of my inquiry is restricted to *cognitive representations*; that is, representations as explanatory posits of specific cognitive theories. As it is customary, I will understand cognitive representations based on the template offered by *public* representations; hence, as triadic relations holding between a vehicle, a target, and some system "consuming" the vehicle. I will then briefly describe two further essential properties of representations; namely representational content and representational functional profile, and then close the chapter introducing four classes of representations (receptors, structural representations, superposed representations and input-output representations) that will play a major role in the upcoming argument.

Chapter 3 puts the conceptual resources introduced in Ch. 2 to use, introducing the structural-representationalist view of Predictive Processing. The chapter starts by presenting the representational functional profile of generative models, showing how the thinnest possible notion of model at play in the Predictive Processing framework has been strengthened so as to yield a *prima facie* robust representational functional profile. The chapter then deflects a number of objections aimed at shaking the representational credentials of such structures, showing how the structural-representationalist view can respond. In this way, the chapter presents a strong case in favor of the structural-representationalist view, providing a charitable reconstruction of it which also eases the examination of the representational content of generative models.

Part II: Generative models as instantiations of sensorimotor mastery

Chapter 4 examines whether generative models *actually* qualify as structural representations. In doing so, I will focus on the argument offered by Gładziejewski in his seminal *Predictive Coding and Representationalism*. This is because of two reasons. First, it

provides the standard understanding of "structural representations" used throughout the Predictive Processing literature. Secondly, it is, to my knowledge, the only argument explicitly aimed at establishing the metaphysical status of generative models as structural representations.

The structure of the chapter is straightforward. In the first paragraph, I briefly expose Gładziejewski's understanding of structural representations and unpack the relevant definitions. In the second, I sketch Gładziejewski's argument to demonstrate that generative models qualify as structural representations. In the third paragraph, I turn from exposing the argument to criticizing it, arguing that it does not substantiate the desired conclusion. In the fourth paragraph I examine a number of alternative arguments to the same effect, finding them wanting. They all either fail to establish the desired structural similarity or succeed to establish it only at the expense of some other feature individuating structural representations. I thus conclude that Gładziejewski's argument fails, and that the metaphysical status of generative models as structural representations is far from secured.

Chapter 5 begins where chapter 4 left off: what if some appropriate structural similarity were to be found? Wouldn't that turn the tide in favor of a (structural) representationalist reading of PP? I claim the answer to this question should be negative. This is because the relevant definition of structural representation provided does not spell out a *representational* functional profile, and items satisfying it do not meet the "Job Description Challenge"; that is, they do not function *as representational vehicles* within the systems in which they are deployed.

In the first section of the chapter, I introduce Ramsey's Job Description Challenge, and briefly show why receptors do not satisfy it whereas structural representations allegedly do. In the second section, I show that at least some receptors satisfy all the demands Gładziejewski poses on structural representations, thereby showing the structural profile of the two is not substantially different. Thus, if receptors do not satisfy the challenge, structural representations

don't satisfy it either. The third section replies at some foreseeable objections. The fourth section concludes the chapter spelling out the conclusion just reached: generative models definitely are *not* representations, structural or otherwise. But, then, what *are* they?

Chapter 6 tries to provide an answer. It considers generative models directly, trying to answer the question: "what could generative models be, if not structural representations?". To answer this question, I examine the simplest generative model capable of active inference I know of, in the form of a simple robotic "brain". I show that, as a matter of empirical fact, such a "brain" is manifestly sufficient to allow the robotic agent to achieve a certain degree of sensorimotor mastery. Yet, I also show that such a "brain" hosts no structure that can rightfully be identified as a representational vehicle. I thus conclude that such a generative model is a non-representational structure instantiating the agent's sensorimotor mastery; that is, the agent's practical and tacit knowledge of sensorimotor contingencies. I then consider whether my verdict generalizes to more complex Predictive Processing systems, concluding that, absent any compelling argument blocking the generalization, it does. In the conclusion of the chapter, I consider and allay some worries that my anti-representationalist verdict may rise.

A brief conclusion then closes the dissertation.

Part I: Predictive Processing, representations, and Predictive Processing and representations

Chapter one - An introduction to Predictive Processing

1 - What is Predictive Processing?

Predictive Processing is a model of cognition spanning Marr's (1982) three levels of analysis.

At the *computational* level, Predictive Processing (PP) suggests that the basic task of the entire brain is that of minimizing *prediction error*: the mismatch between expected and received sensory stimulation. In more rigorous mathematical terms, prediction error minimization can be cast as a form of approximated Bayesian inference; making PP a specific instantiation of the "Bayesian brain" hypothesis (Friston 2009; 2010, Hohwy 2013; Buckley *et al*. 2017).

At the *algorithmic* level, PP conceives the brain as a multilayer neural network hosting at least two (but often three) kinds of processing units, termed *prediction and error* units (but units for *precision weighting* are often added). These units are densely connected by two, non-overlapping, sets of connections, busy transmitting predictions and prediction errors according to a *predictive coding* processing regime (Rao and Ballard 1999; Friston and Kiebel 2009a; Clark 2013a; 2016).

At the *implementation* level, PP suggests that such a network can be readily seen in the mammalian cortical architecture, for example observing the *hierarchical structure* of the brain; or the *two non-overlapping sets of ascending and descending connections* tying together hierarchically stacked neural areas. (Friston 2005; 2008; Adams, Shipp and Friston 2013; Shipp 2016).[1]

PP is also a *unified* account of cognition, suggesting that cognition consists *entirely* in prediction error minimization. Hence the extensive explanatory reach PP seemingly boasts,

[1] I'm omitting lateral connections for the sake of simplicity. Notice, however, that they will be relevant in Ch. 6

spanning from simple sensorimotor coordinations to memory (Vecchi and Gatti 2020), emotion (Seth and Friston 2016), dreaming (Hobson and Friston 2012), curiosity (Friston *et al.* 2017c), social cognition (Friston and Frith 2015a, b) various form of reasoning (Spratling 2016) and even phenomenal consciousness (Seth 2021).

Here, I will introduce the basics of this account, focusing on how it accounts for perception and action at the algorithmic level.

To be clear, this means that this introduction is *partial and idealized*; hence partially *distortive*. I shy away from currently debated topics in the PP literature, ranging from non-Bayesian forms of PP (e.g. Thornton 2017; 2020), issues concerning the correct form of the predictive coding algorithm (Spratling 2017) and network architecture (O'Reilly, Wyatte and Rohrlich 2014; Tani 2016; Matsumoto and Tani 2020; Ciria *et al*. 2021), as well as whether the evidence in favor of PP is conclusive (Keller and Mrsic-Flogel 2018; Walsh *et al*. 2020; Cao 2020; Millidge, Seth and Buckley 2021), and whether fitting the architecture described below to neuroscientific data will fore significant revisions (Spratling 2019; Millidge *et al*. 2020). Such controversies are best ignored in an introduction.

As it is customary, I will introduce PP starting from perception.

2 - Perception, and the core concepts of PP

PP suggests that perception is an instance of prediction error minimization, whereby the brain[2] inverts the generative model it encodes, thus mapping sensory inputs onto their most likely causes, thereby recognizing the former in terms of the latter. There's a lot to unpack here, starting from one of the PP core concepts; namely, the concept of *generative models*.

[2] Throughout the rest of the dissertation, I will use "brain" roughly as a shorthand for "cognitive machinery".

2.1 - Generative models

A generative model can be conceived as a *probabilistic mapping* between sources of inputs (or, more broadly, *hidden causes*) and inputs (or, more broadly *data or observations*) capturing a *generative process*; that is, how the former generates the latter (Foster 2019:1). Sampling from one such model *generates* new observations, which corresponds to what the model predicts, given its knowledge of the generative process (cf Danks 2014: 44).

Since sensory inputs are typically generated by nonlinear interactions of different causes operating at different spatio-temporal scales, a generative model capturing how our sensory input is generated must be *hierarchically structured*, so as to capture causes operating at progressively coarser spatiotemporal scales. Moreover, it must allow for nonlinear interactions among its variables.[3]

The model must also be *probabilistic* in nature. In ecologically normal contexts, causes do not *always* produce the same effects: for example, the inputs generated by a sound source vary in function of the perceivers distance relative to a source, and the light reflected by a surface changes as the illumination condition changes. For this reason, the model encodes the relevant knowledge in probabilistic terms. This is necessary to handle the perceiver's uncertainty about the states of the hidden variables at work behind every sensory observation (Tenenbaum *et al*. 2011: 1280). How, exactly, biological brains encode probabilities is not a settled matter and it will not be discussed here.[4]

One way to intuitively understand generative models is to use one. Enter *Artbreeder*: a collaborative art website providing a simple interface that allows its user to use a generative model to create pretty pictures, see **figure 1**.[5]

---

[3] See (Friston 2008; Kiebel, Danizeau and Friston 2008; Kibel *et al*. 2009).

[4] See (Knill and Puget 2004; Kersten, Mamassian and Yuille 2004; Aitchinson and Lengyel 2017). Indeed, the claim that the brain computes on probability distributions is contested (Sanborn and Chater 2016).

[5] It can be freely accessed at https://www.artbreeder.com/. Notice that in the present context Artbreeder is just used as an example: the neural network fueling it is actually quite distinct from the architecture PP envisages. But, for present purposes, their differences are irrelevant.

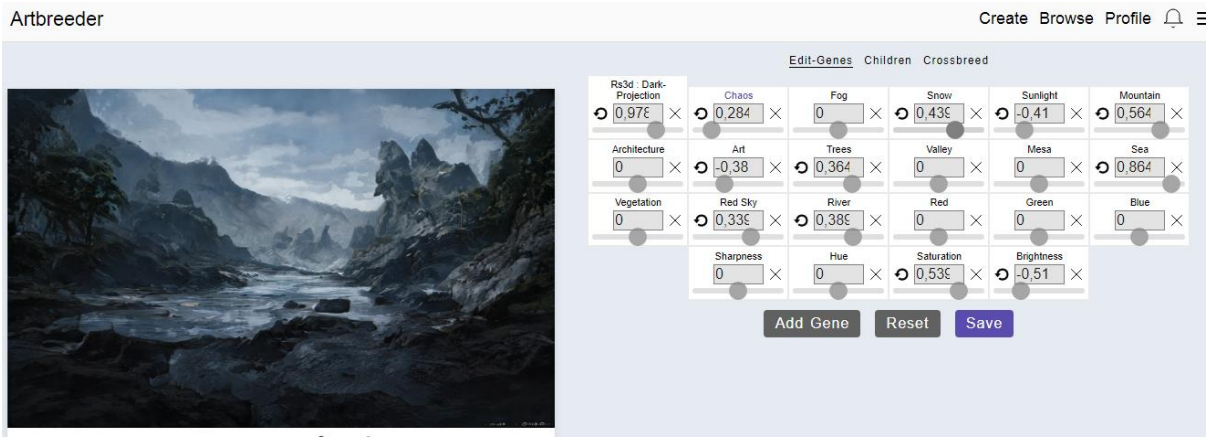

Figure 1. The interface of Artbreeded. Source: screenshot.

The interface displays the *hidden causes* (right) of the *generated image* (left). The user can assign a value (ranging from -2 to +2) to each cause, thereby modifying the generated image. For example, by incrementing the value of "snow", the image will be dominated by cold, muted colors, whereas decreasing its value will make colors more hot and vibrant. Each assignment of value to causes generates an image that Artbreeder's model *expects or predicts*, given the value of each cause.

This intuitively clarifies how generative models can *generate* predictable observations. It also clarifies in which sense *prediction* is relevant in the present context: predictions are not prophecies about the future, but (sub-personal) mechanisms of statistical estimation, closely aligned with mechanisms for pattern completion (Bubic, Shubotz and Von Cramon 2010; Falandays, Nguyen and Spivey 2021).

Artbreeder's interface, however, does not show its users the *hierarchical* nature of the model they're using. To intuitively visualize it, one must deploy a different method, embodying generative models in complex mathematical formulae. The following is a reasonably simplified rendition:[6]

- sensory signals = functions of the hidden causes + noise

[6] See (Wiese 2017; 2018) for the original proposal. See (Rao and Ballard 1999: Eq. 1; Friston *et al*. 2010; Eq. 8) for the non-hierarchical case, and see (Rao and Ballard 1999: Eq. 3) and (Friston *et al*. 2015: Eq. 3.3) for the hierarchical case.

or, equivalently:

- $s = f(c) + \omega$

If we ignore the noise term ω (it will come back in §3), the formalism roughly captures what is going on behind artbreeder's interface: the user sets the hidden causes at a value, and the program outputs the predictable image/sensory signals.[7] Now, it is fairly easy to expand this formalism to capture the hierarchical case:

- $c_2 = f_3(c_3) + \omega_3$
  $c_1 = f_2(c_2) + \omega_2$
  $s = f_1(c_1) + \omega_1$

Intuitively, the idea is that causes at hierarchically higher levels constrain the values of causes at hierarchically lower levels. This can easily be translated into Artbreeder's interface by imagining a *second* set of superordinate causes constraining the values of other causes: for instance, a superordinate cause such as "night time"could constrain the value of cause "direct sunlight" fixing it at very low level, and allow to increase to a significant level only the value of the cause "artificial illumination". Notice that this means that a hierarchical generative model just is a hierarchy of generative models: in fact, each hierarchical layer generates (and thus, models) what it observes at the layer directly below itself.[8]

Another way to conceive generative models is by means of *graphical models* - this will be important in chapters 3 and 4.[9] Graphical models are representational devices affording the concise representations of the relations (in this case, probabilistic relation) holding among a set of variables (in this case, hidden causes and the sensory signals they produce). A graphical model consists of two components: a set of (labelled) nodes, which represent the variables, and a set of (directed) connections, which represent relations among the variables, see **figure 2**.

[7] For the sake of completeness, notice that there is a sense in which the interface of Artbreeder lets its users access the noise term ω: it's the hidden cause "chaos", that governs the unpredictability of the image generated.
[8] See (Hinton 2005; 2007a;b; 2014; Eliasmith 2013: ch. 3; Simione and Nolfi 2015).
[9] See, for instance, (Friston, Parr and de Vries 2017). See also (Penny 2012: 4-7; Danks 2014: 39-64) for nice introductions and Koski and Noble (2009: 37-45) for a formal presentation of the relevant mathematical apparatus.

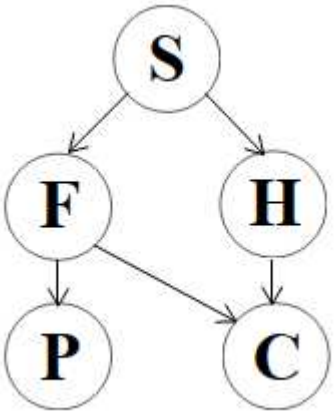

Figure 2. A simple graphical model, see main text for details (Drawing by the Author)

Let **S** stand for the variable *season*. Certain seasons (spring) raise the probability of *Heyfever* (**H**), and so there is a direct connection (→) between these two variables. Knowing that the other nodes stand for *flu* (**F**), *congestion* (**C**) and *muscular pain* (**P**), the graph becomes easily interpretable - for example, it makes clear that there is no direct connection between *season* and *muscular pain*: muscular pains are more likely in winter only because having a flu is more likely in winter. In this way, the topology of a graph comes to mirror the relevant relations holding among the elements that the graph models.

Notice that graphical models too allow their users to make predictions. Using the simple model in **figure 2**, one can easily predict that during certain seasons muscular pains will be more likely. Indeed, all graphical models are generative models - at least in the minimal sense that they can be used to generate typical data, given the knowledge they embody (Danks 2014).

Now, consider again the most intuitively manageable rendition of a generative model thus far encountered: Artbreeder's interface. It can be used to intuitively clarify in what sense a generative model is inverted by means of prediction error minimization. Suppose that an Artbreeder user is tasked with *recreating* a target image (here playing the role of the actual sensory input) via Artbreeder. A natural way to proceed would be that of modifying the values assigned to the causes in a way such that the image Artbreeder displays comes to progressively

resemble the target image. Notice that, by doing so, the user is *implicitly* mapping the target image onto its most likely causes (according to Artbreeder's model). In fact, in order to re-create the target image via Artbreeder, its user has to find the right setting of causes that, according to Artbreeder's model, would *generate* that image. Hence, by minimizing the discrepancy between the target image and the one generated by Artbreeder (which is the prediction error), the user is implicitly mapping the target image onto its most likely causes, *inverting* the mapping from causes to images encoded in the generative model.

Now, Artbreeder requires a human user. But human users must be eliminated by cognitive theorizing, on the pain of *homuncularism*. In the next two sections, I will clarify how to eliminate them, introducing the algorithmic core of PP.

2.2 - A connectionist architecture

PP is, at its core, a *connectionist* theory (Rao and Ballard 1999; Spratling 2016; Millidge *et al*. 2020).[10] And philosophers are familiar with a class of connectionist systems embodying statistical models in a homunculus-free way: feedforward networks used for classification all embody discriminative or recognition models. That is, they all output the most probable class, given an input (Skansi 2018). Thus, feedforward networks offer a natural starting point for the present discussion.

A feedforward network consists of a set of hierarchically stacked nodes (or units or neurons) systematically connected by means of weighted connections. The network is administered an input by setting the numerical value of the nodes at its input layer. Then, their activation spreads through the network, and it is dampened or bolstered depending on the weight of each

[10] Notice that this does not make it "less Bayesian". Both Bayesian and connectionist approaches to cognition take cognition to be a process of satisfaction of multiple, graded and probabilistic constraints (Kersten, Mamassian and Yuille 2004; Rogers and McClelland 2014). Moreover, some connectionist systems can be straightforwardly interpreted as implementing Bayesian probabilistic reasoning (e.g. Hinton and Sejnowsky 1983; McClelland 1998; 2013).

connection. Each hierarchical layer of units receives input from the layer below it, which determines the state of each of its neurons. This procedure is iterated through the layers until the output layer is reached, and its pattern of activation indicates to which class (among the classes the network has been trained to discriminate) the input most likely belongs.[11] As this simple description highlights, the computation of artificial neural networks massively (but not exclusively) depends on the *weights* attached to the connections between units. It is by training and adjusting these connections that a network comes to embody the relevant model (see Ripley 1996; Murphy 2012: Ch. 3): in the case at hand, a discriminative model mapping input patterns onto labels.

In feedforward networks, these weights are typically learned through an error minimization algorithm known as *error backpropagation*.[12] But it has some serious shortcomings. For one thing it is *supervised*, hence it presupposes the presence of a teacher who knows the right answers (cf Nolfi and Parisi 1993) as well as vast amounts of labeled data (Hinton 2014). Moreover, it is informationally wasteful: it forces the net to ignore all the information present in the data which does not pertain to the relevant discriminative model the network has to learn (Hinton 2007a: 340; McClelland 1998: 30). These shortcomings motivated researchers to find better ways to make networks learn discriminative models - and a powerful and potent way to do so is that of making the network learn a *generative* model of the input.

Consider Helmholtz machines[13] (Dayan, *et al*. 1995; Dayan and Hinton 1996): artificial neural networks deploying stochastic units uniformly connected by *two* sets of connections. The first set of *recognition* weights drives the machine in the familiar bottom-up way, allowing the propagation of activation from input to output layer. The second set of *generative* weights

[11] See (Haykin 2009: Ch 1 to 4; Kruse *et al*. 2016: Ch. 3-5 for a proper formal treatment) of feedforward networks.
[12] See (McClelland, Rumelhart and the PDP research group 1986) for its canonical formulation ans (Skansi 2018) for an abridged formulation.
[13] To be clear: Helmholtz machines do not provide the standard connectionist implementation of PP. I'm introducing them here to ease the exposition, just as I did with *Artbreeder*. This is far from an uncommon practice in the PP literature (cfr. Clark 2013a; Kiefer and Hohwy 2018; 2019).

drives the activity of the system from the top-down, allowing activity to spread from the output layer to the input layer. Both sets of connections are randomly initialized, and each set of connections acts as a teacher for the other set (Hinton, *et al*. 1995; Dayan 2003). This is the "wake-sleep" algorithm.[14]

In the "wake" phase, the network functions as a feedforward network, driven by its recognition weights, while the generative weights are amended so as to make the network more likely to generate the inputs just encountered. The amended generative weights are then put to use in the "sleep" phase of the algorithm, in which the network is driven "from the top-down" and the activation spreads from the *output* to the *input* layer through the generative weights. Under this processing regime, the network adjusts its *recognition* weights, making them more likely to map input patterns on the class producing them. Repeated iterations of this cycle will allow the network to generate *realistic* inputs, while discovering the *classes* that best account for its production in an entirely unsupervised fashion. For example, a network trained according to this learning procedure can discover that the *best* way to classify input digits is not based on their surface form (which would arguably led to reliably misclassify porely scribbled "3s" "5s" and "8s" as well as "1s" and "7s"), but rather based on the *motor programs* the network executed while generating numbers by scribbling with a virtual pen (Nair and Hinton 2006).

There are two lessons to draw from these examples. The first concerns the fact that a generative model and a discriminative model can *train each other* in an unsupervised manner, thereby jointly discovering the classes that best account for both the *production* and the *classification* of the input pattern. The second is that generative models can be easily de-homuncularized by simply letting the activation spread from input classes to the network's "sensory periphery".

[14] The wake-sleep learning algorithm is a tractable approximation of the expectation maximization algorithm (see Dayan and Abbott 2001: ch. 10), which belongs to a variety of techniques to estimate the priors directly from the data (see Murphy 2012: ch.5).

Now, Helmholtz machines (and similar networks) nicely illustrate these points. Yet, their *generative* weights operate only during training. This seems wasteful and biologically implausible (Rao and Senjowski 2002; Lee and Mumford 2003): wouldn't it be better if generative weights were able to operate *also* during online recognition? Predictive *coding* provides generative weights such a role to play.

2.3 - Predictive coding

Strictly speaking, predictive coding is just a data-compression strategy based on a simple idea: instead of transmitting an entire signal, just transmit the difference between the *expected* and *received* signal (i.e. the prediction error, Shi and Sun 2008: ch. 3; Spratling 2015; 2017). This can allow one to send *shorter* signals, containing only newsworthy information. Compare: if I'm not wearing socks, the sentence "I'm not wearing socks" is a much shorter description of my state than "I'm wearing only my shirt, jeans, belt, jacket, pants and shoes".

There are other reasons as to why predictive coding is attractive. In perceptual neuroscience, for instance, a common complaint is that:

> "[...], we barely understand the top-down mechanisms by which incoming sensory information invokes memories of past occurrences and activates our subjective prejudices and opinions" (Kandel *et al.* 2012: 471-472)

Predictive coding provides precisely such an understanding (cf. Bar 2007; 2009; Bar *et al*. 2006; O'Callaghan *et al*. 2017), as it provides top-down (generative) connections a role to play in online recognition.

As seen above (§2.2) a generative model can be understood as a multilayer neural network. Let each layer consist of two *functionally*[15] distinct sets of units, encoding predictions and prediction errors respectively. Let each layer L_n is connected with the adjacent layers L_{n-1} and

[15] Notice that functional separateness does not necessarily imply *physical* separateness. Tani's implementations of PP, for instance, have no separate set of error units (see Tani 2016), and his most recent models rely on a *single* set of connections (Matsumoto and Tani 2020).

L_{n+1} by two non-overlapping sets of connections. Descending connections, from L_n to L_{n-1} carry predictions, whereas ascending ones (from L_n to L_{n+1}) carry prediction errors. In each layer L_n, prediction units try to predict the activity of the layer below[16] by inducing[17], in that layer, the predicted pattern of activity. Their collective activity thus results in a *downward* flow of progressively spatiotemporally specified predictions. For instance, a relatively high-level layer might generate a state roughly corresponding to "face"; a middle-level layer a state corresponding to "eyes, nose and mouth arranged thus-and-so" and a low-level layer something like the complex set of colored pixels corresponding to the image of a face. This signal is then subtracted from the *actual* activity of the layer receiving it (or, at the bottom most layer, the actual input). The remaining signal (prediction error) is then sent *upwards* into the hierarchy. The receiving layer uses it, to update its guess, so as to minimize the incoming prediction error. As a global minimum of error is reached, the sensory input has been recognized, and interpreted in terms of its most probable underlying cause (according to the model). As Tani clearly explains:

> "[...] the process of recognizing a target perceptual sequence can be formulated as a process of searching for an optimal intentional state by which the target sequence can be generated with a minimum error while the learned weight parameters W remain fixed." (Tani 2014: 589; note that W denotes the weights of the connections)

Notice that this is precisely what the imaginary user of Artbreeder tasked with *copying* an image did (§ 2.1): to copy it, the user searched for the configuration of causes[18] that generated the image most similar to the target one; that is, the configuration that generates *that* image with the least prediction error possible.

---

[16] Or, in the bottomost layer, the incoming input signal

[17] The term "inducing" in this context is tricky - in fact, from a purely technical point of view, top-down influences are often described as inhibitory (cf Friston 2005). One might more properly think of each layer as passing *a negative* of the expected activity to the layer below, in a way such that, were the negative (i.e. the prediction) correct, the inferior level would stay silent because expected and actual activity cancel each other out.

[18] Tani refers to it as an "intentional state" due to his strong phenomenological influences (cf Tani 2016: Ch. 3). I prefer a more sober terminology.

To see predictive coding operate in artificial neural networks, consider the network described in (Rao and Ballard 1999). The network consists of a hierarchy of stacked prediction and error units, connected as sketched above, which was trained to recreate naturalistic images.

The network learned to both recreate *and recognize* simple visual features (such as edges and bars) at the first level, as well as their *complex combinations* at the second level, in a way that fits the statistics of the natural images used to train the model. It also exhibited *non classical* receptive fields effects, such as end-stopping. End-stopping (see Rao and Ballard 1999: 79) occurs when a neuron vigorously responds to a feature (e.g. a bar) presented in its classical receptive field, but stops responding when the same feature is also present in the surrounding of the neuron's classical receptive field (e.g. if the bar prolongs outside the neuron's receptive field).

The fact that Rao and Ballard's net displays non classical receptive fields effects suggests a plausible *functional interpretation* of such effects in terms of fulfilled predictions. In the case at hand, end-stopping ensued because the activity of the relevant units can be well predicted by the activity of the units surrounding it. As that activity is easy to predict, little to no prediction error is generated, and the unit gets "turned off". To test this conjecture, Rao and Ballard (1999: 81-83) tested the network inhibiting the message passing from higher to lower level, thus impeding the propagation of the prediction signal. In this condition, the neurons' "end-stopping behavior" vanished: they responded equally vigorously to bars, regardless of the number of receptive fields the bar crossed.

The fundamental pieces of the PP machinery have been introduced. The next subsection shows them in action through one paradigmatic example.

2.4 - The core machinery in action: the case of binocular rivalry

Consider binocular rivalry: a perceptual phenomenon ensuing when each eye receives a

different stimulus from one of two different visual targets. Subjects report that, in such conditions, they see only *one* visual target at a time, and that their perception alternates between the two targets. Thus, if the two targets are a face and a house, subjects do not see a "face-house", but rather a face, then a house, then a face (and so on).

Thus, binocular rivalry poses *two* explanatory challenges. The first is that of providing an account of *percept selection*; that is, providing an account explaining why subjects report seeing coherent percepts rather than "mashups" of the two stimuli. Secondly, one must account for the *alternation* of the percepts.

PP easily meets both challenges (Hohwy, Roepstorff and Friston 2008). According to PP, the perceiver's brain constantly tries to "guess" the incoming input. Suppose the guess corresponds to one of the stimuli (say, pattern 1). Since one eye is actually exposed to pattern 1, the corresponding half of the visual cortex will generate only little prediction error. But the *other* half will generate a strong burst of error, for the eye connected to it not stimulated by pattern 1, but rather by pattern 2. To minimize that error, the brain revises its prediction in favor of pattern 2. This minimizes the previous burst of prediction error, but also generates a new one, this time coming from the half of the cortex "seeing" pattern 1. So, the prediction has to be revised again in favor of pattern 1, eliciting a burst of error from the half "seeing" pattern 2 - and so forth.

But why does the brain "select" coherent percepts, rather than a mashup of the two stimuli? The answer is straightforward: generative models can generate (and classify) the data *they encounter*, and we rarely (if ever) encounter such mashups in *ecologically normal* contexts. So the brain simply does not generate a "mashup" prediction, because it is not part of the data it has learned to generate and recognize.[19]

[19] Statistically speaking, the "mashup" percept (if at all present) has too low of a prior probability to be the "winning class" of the classification

The core functioning of the PP machinery should now be reasonably clear. It is thus time to introduce two important additions.

3 - Expected precision and attention

Binocular rivalry illustrates nicely how the knowledge encoded in a generative model shapes perception: one cannot see the "mashup" percept because it's not part of the data the generative model has learned to generate and recognize. Such knowledge accounts for numerous cases of perceptual illusions (Brown and Friston 2012; Weiss, Simoncelli and Adelson 2012), suggesting that perceptual illusions are *optimal* percept, given what the model knows.

Consider now a *prima facie* similar case (Merckelbach and van der Ven 2001). A number of undergraduate students were made to listen to a short audio file containing just wite noise. However, participants were informed that the audio file contained a barely audible version of "White Christmas" buried under the noise, and one third of them actually reported hearing the song.

Prima facie, this seems just a simple perceptual illusion, accounted for the participants' expectations concerning White Christmas. But, *why* hadn't the participants *revised* their expectations in the light of the incoming prediction error? After all, what they *really* heard was just white noise, which must have generated at least some prediction error. So, why didn't it force a revision of the prior expectation? To answer this puzzle, recall the noise term "ω" introduced in §2.1. It is now time to integrate it in the machinery of PP.

Consider again the "White Christmas" scenario. Participants were informed both on what to expect and *how* to expect it (i.e. White Christmas *buried in the noise*). Given this expectation on the *quality* of their sensory inputs, small snippets of noise casually resembling White Christmas confirmed the prediction *well enough*, allowing one to ignore the prediction error generated by the rest of the singal. Indeed, participants expect their sensory signals to be *low*

quality. They *expect* it not to match their predictions. Thus, it seems right *not* to use the prediction error to revise the prediction (cf Clark 2016, Ch. 2).

PP suggests considering this expectation on the quality of the data as an estimate of the signal-to-noise ratio, or *precision*, of sensory signals and the prediction error they generate (Friston 2009; Feldman and Friston 2010). Notice the metacognitive character of these predictions, as they do not concern the data itself, but rather their *reliability*, determining to what extent the prediction errors they generate should be allowed to impact on our prior expectations about the *causes* of the signal (Adams *et al*. 2013: 2). Prediction errors generated by signals expected to be highly precise sensory signals are considered reliable, hence they should be allowed to force a revision of the predictions. Conversely, prediction errors with low expected precision are deemed an *unreliable* source of information, and so are not "trusted over" the predictions generating them.

Here's an intuitive example. Suppose I expect point A and B to be 6 steps apart. I then repeatedly count the steps I take to move from A to B. This isn't a *reliable* measuring procedure: sometimes I will count 7 steps, other times I will count 5. Knowing this, it is rational for me to ignore the prediction error (i.e. the discrepancy between expected and actual distance) and keep believing A and B are roughly 6 steps apart.[20] Had I used a laser rangefinder, I would have revised my belief: this is because rangefinders are very precise tools - and so if their measurement contradicts my estimate, I better trust the measurement.

As the example intuitively shows, expected precision acts as a *weight* on the prediction error. If the weight (i.e. the expected precision) is low, prediction errors will be "silenced", and predictions will dominate. Conversely, if their expected precision is high, they will fore predictions to be revised:

> "[...] when the bottom-up noise variance is high (for instance, due to occlusions),

[20] Of course, it is rational to ignore the prediction error *up to a point*: had I measured the distance from A to B to be of hundreds of steps, it *would* be rational for me to revise the estimate. The same idea is actually present in PP: large enough prediction errors have an increased expected precision (cfr. Feldman and Friston 2010: 23).

> the bottom-up term is given less weight in the state estimation step [...] and the estimate relies more on the top-down term and the system prediction[s]. On the other hand, when the top-down noise variance is high (for instance, due to ambiguity in interpretation by the higher-level modules), the estimate relies more heavily on the bottom-up term [...]. The dynamics of the network thus strives to achieve a delicate balance between the current prediction and the inputs from various cortical sources by exploiting the signal-to-noise characteristics of the corresponding input channels" (Rao and Ballard 1997: 733, formalism deleted)

In biological brains, PP suggests that this delicate balance is achieved by a variety of mechanisms regulating the firing rates of neurons, so that neurons reporting highly precise prediction errors will fire more, whereas neurons reporting less precise errors will be inhibited.[21]

Importantly, these mechanisms collectively implement a form of *attentional control* (Feldman and Friston 2010). Intuitively, the idea is extremely appealing. In fact, "boosting" the error coming from one source while inhibiting the error coming from other sources[22] *just is* to attend to a specific source, letting the signal it produced dictate one's neural processing. This roughly captures *endogenous* attention; that is the kind of attention we allocate "from the top-down" to what we deem relevant. Exogenous attention is instead captured by the fact that large bursts of prediction error are always highly weighted *because* they are large (Feldman and Friston 2010: 23). Indeed, large deviations from our expectations *do* capture our attention, as when a sudden movement, or an unexpected booming sound, force us to turn in the direction of their source.

According to PP, expected precision is also related to the notion of *salience*. Salience denotes a quality of sensory signals which have an high expected precision but whose causal origin is still ambiguous (Friston *et al*. 2012b; Parr and Friston 2017; 2019). The sensory signals obtained when opening the door after the doorbell rang are salient in this sense: we do

[21] These mechanisms canonically included post-synaptic responsiveness modulation, synchronization of firing rates, and the release of neurotransmitters (Feldman and Friston 2010; Friston 2010; 2012a; Friston *et al*. 2012a).
[22] Of course, the weighting due to expected precision must be differential, just as in a weighted sum addends must have different weights (otherwise, that wouldn't simply be a *weighted* sum!).

not know what is their causal origin (i.e. who rang the doorbell), but we know that opening the door will provide us high-quality information on that matter.

As a more concrete example, consider the learning algorithm engineered by Bongard, Zykov and Lipson (2006). The algorithm was tailored to enable a simple robot to infer its bodily morphology from its actions. The robot starts its learning cycle acting at random, only to collect some data relevant to infer its body morphology. Then the robot computes a number of competing body models, which are examined to determine the course of action that yields *the most different sensory signals* according to each model. The robot then enacts the action thus selected, eliciting some propriocetive signals. This enables the robot to "guess" its morphology: the body model that best predicts the actual signal is the body model that most closely approximates the robot's actual morphology. The path of action the robot choses is *salient* in the relevant sense: its causal origin (actual morphology) is still unknown, but it is expected to yield the relevant data to guess it the best.

The example also nicely highlights the *prescriptive role* of salience. Salience suggests generative models where to look and what to touch. But how do generative models look and touch? And, more generally, how do generative models *do* things?

4 - Letting predictions do by their their fulfilling: action and active inference

As seen in (§2) PP casts perception as a process in which predictions are aligned to the incoming sensory inputs, thereby minimizing prediction error. But prediction error could be minimized also the other way around; that is, by changing the inputs so as to make *them* fit the predictions. Roughly, this is the PP account of action (*active inference*). On the view PP offers, actions make our predictions *self-fulfilling*: we move so as to encounter the sensory inputs we expect to perceive. There's thus a sense in which the model of cognition PP offers:

> "[is] *concerned with, and only with, perception*. Action per se, was a result of movements that conformed to the proprioceptive predictions of the joint angles.

> This means that perception and action were both trying to minimize prediction errors throughout the hierarchy, where movement minimized the prediction errors at the level of proprioceptive sensations." (Namikawa, Ryunosuke and Tani 2011: 4; emphasis added).

Action (active inference) and perception are thus complementary sides of the same prediction error-minimizing coin, distinguished only by the *way in which* error is minimized. In perception, error is minimized by letting reality control and correct the model's prediction. In action, the model is let free to control reality.[23]

One way to approach active inference is by noting that living bodies are rich sources of proprio- and viscero- ceptive signals, which predictive brains strive to predict and model.[24] Error relative to these predictions can be easily corrected through bodily motions that *bring about* the predicted sensory state: if I predict the suite of sensory signals caused by my hand being closed, a good way to eliminate the error these predictions generate is just that of closing my hand.

Notice that, in order for such a prediction regime to function properly, predictions in all modalities must march in step. Otherwise, minimizing prediction error in *one* modality might actually increase the prediction error of others. For example, if predictions about my head position and visual stimulations do not march in step, each head movement would bring about *unexpected*, hence prediction-error-inducing, visual stimulation.

A simple way to quite literally see predictions in all modalities marching in step is that of closing an eye, while gently pushing the open one, which typically causes one to perceive the objects moving in the direction opposite to the direction of the push.[25] This might seem

---

[23] One could say that perceptual and active inference differ in their *direction of fit*. Yet, the profound neurocomputational similarity of perception and action should make us resist that characterization (Wiese 2018; Clark 2020). This is why, in the following, I will largely be silent on *imperative* representations (i.e. representation with a world-to-mind direction of fit).

[24] See (Seth and Critchley 2013; Seth 2015; Barrett and Simmons 2015; Seth and Friston 2016; Seth 2021). According to PP, these bodily predictions are also linked to emotion and our sense of ourselves - two themes I won't explore here.

[25] This discovery, as well as the discovery that the same shift is preserved when the eye is paralized but one intends to move it, are typically credited to Helmholtz's self-experimentation.

unsurprising: after all, the push physically displaces the eye, thereby changing the retinal image. But saccadic eye movements physically displace the eye too, and yet we do not see objects constantly "jumping around". Moreover, *the same* shift is also observed when the eye is paralized, and one attempts to move it (Gallistel 1980: 175), and surely the *attempted* movements of a *paralyzed* eye do not cause change in the retinal input.

So, what accounts for these phenomena? A long-standing account (cfr. Sperry 1950) suggests that these phenomena are due to the match (or lack thereof) between visual signals expected after the movement and actually received ones. In normal contexts (i.e. normal saccading eye movements) the two "march in step", and so the expected signals compensate for the apparent motion caused by the physical displacement of the eye. But when the two mismatch (either because no motor command is ensued, or because the physical displacement of the eye is prevented), the prediction fails to compensate for the incoming (unexpected) sensory signals, and so apparent motion is perceived.

Computationally, the visual signals expected after movement are produced by a *forward model*. (Blackmore, Wolpert and Frith 1998; 2000; Pickering and Clark 2014). Forward models are *special purpose* generative models, tasked with converting their inputs (motor commands) into the *expected sensory consequences* of movement. As the example above demonstrates, these models "polish" our perception, filtering out the predictable (hence uninformative) input.

Forward models also *enable* fast and fluent action. They do so in two ways. On the one hand, they allow for the identification of a "redundant subspace" of motor parameters, the fluctuations of which do not hinder the success of actions. In this way, motor control is enormously simplified, as the system has to control only a few, well selected parameters (cf. Todorov and Jordan 2002; Todorov 2009a). On the other hand, they allow to circumvent the delay of reafferent signals (e.g. Clark and Grush 1999; Grush 2004). It estimated that proprioceptive feedback is delayed from 80 to 150 ms if compared to visual feedback (see

McNamee and Wolpert 2019: 352), and so it comes *simply too late* to guide on-line quick actions. Indeed, when it comes to motor control:

> "We effectively live in the past, with the control systems only having access to out-of-date information about the world and our own bodies, and with the delays varying across different sources of information." (Franklin and Wolpert 2011: 425-426)

Predicting the sensory consequences of our movement, thus, allows us to *act in the present*. As a nice example of this, consider so-called "waiter tasks".[26] In these tasks, an experimental subject holds firmly an object within one hand. A weight is then added to the object, either by the subject or by the experimenter. When the additional weight is added directly by the experimental subjects, they are able to *proactively* modify the force of their grip, avoiding almost any slippage of the object. This is because they can predict the sensory consequences of their movements (i.e. the increase of the load carried) and thus act so as to counteract the slippage of the object. Conversely, when the additional load is *not* added by the subjects, the adjustment of the grip force *lags behind* the addition of the load. Since subjects are unable to precisely predict how (and when) the load will increase, they are forced to adjust their grip reactively, with potentially disastrous results (i.e. letting things drop off) (Flanagan and Wing 1997; see also Wolpert and Flanagan 2001 for a nice review).

Notice (as it will be important in Ch. 6) that, by having to predict the sensory consequences of movement in all modalities, forward models are forced to learn *sensorimotor contingencies*: the law-like ways in which bodily movements alter sensory stimulation (O'Regan and Noë 2001; Maye and Engel 2013; Brette 2016; Pezzulo *el al*. 2017). Sensorimotor contingencies come in two basic kinds: modality-related and object-related (O'Regan 2011). Modality related sensorimotor contingencies depend on the features of an agent's perceptual system - for instance, only systems with eyes must compensate for the optic flow caused by head movements. Object-related ones depend on the features of the source of the sensory signal - for

[26] I take this nomenclature from (McNamee and Wolpert 2019, p. 352).

instance, circling around an object will not change its retinal projection only if it is round or spherical.

Now, according to a well-established theory of motor control, forward models operate in tandem with *inverse* models: computational structures converting goal-states into motor commands. Each pain of inverse and forward model forms a task-specific module for motor control (Haruno, Wolpert and Kawato 2001), and these modules can be hierarchically staked so as to recombine motor primitives into novel motor actions (Haruno, Wolpert and Kawato 2003), and might be deployed *offline* (i.e. without directly controlling actual bodily movements), smoothing social cognition (Wolpert, Doya and Kawato 2003).

PP suggests instead that there is *no inverse model*, and proposes a more economic solution to the problem of motor control.[27] Leveraging the formal identity between motor command and Bayesian inference (Todorov 2009b; Botvinik and Touissant 2012), it simply suggests that the problem of deciding what to do can be reduced to the problem of *what to predict in all modalities*. In a sense, thus, *the entire brain is the forward model*[28], in the task of predicting the incoming flow stimulation based on its knowledge of modality-related and object-related[29] sensorimotor contingencies (Seth 2014; Pezzulo *et al*. 2017; Pio-Lopez *et al*. 2017; Baltieri 2019: 85-100).

Notice that, thus framed, the problem of motor control ceases to exist *as such*. It becomes instead a *perceptual* problem: to control actions, the brain needs only to "decide" what to perceive in all modalities. And, in fact, according to PP:

> "The primary motor cortex is no more or less a motor cortical area than striate

[27] However, PP argues that the way in which spinal alpha motor neurons innervate the muscles constitutes a sort of *implicit* inverse model. See (Friston 2011: 491; Friston *et al*. 2010: 254). So, strictly speaking, PP only denies that *inverse models are explicitly encoded in the* (motor) *cortex*. The reasons for this denial are complex and multifaceted, and, in the present context, examining them would be prohibitive.

[28] Or, better, forward models are no longer *special purpose models*, but are integrated in the overall generative model realized in the brain (Pickering and Clark 2014).

[29] Notice that object-related sensorimotor contingencies *seem* capable of capturing the kind of "knowledge" static (i.e. without agency) generative models encode. See (Hemion 2016; Laflaquiere 2017; Le Hir, Sigaud and Laflaquière 2018) for some evidence in this regard.

> (visual) cortex. The only difference between the motor cortex and visual cortex is that one predicts retinotopic input while the other predicts proprioceptive inputs from the motor plant" (Friston, Mattout and Kilner 2011: 138).

So, according to PP, "motor cortices" are not a "special" kind of cortices issuing *motor commands*. Rather, they are sensory areas issuing *proprioceptive predictions* (Adams, Shipp and Friston 2013; Shipp, Adams and Friston 2013) that happen to trigger bodily movements only because of how the brain is wired to the body:

> "If motor neurons are wired to suppress proprioceptive prediction errors in the dorsal horn of the spinal cord, they effectively implement an inverse model, mapping from desired sensory consequences to causes in intrinsic (muscle-based) coordinates." (Friston 2011: 491)[30]

Meaning that:

> "[...] the inverse problem becomes almost trivial—to elicit firing in a particular stretch receptor one simply contracts the corresponding muscle fiber. In brief, the inverse problem can be relegated to the spinal level, rendering descending afferents from M1 predictions as opposed to commands— and rendering M1 part of a hierarchical generative model, as opposed to an inverse model" (Adams, Shipp and Friston 2013: 25)

Motor control thus appears as an emergent property of generative models, which depends on how they are "wired" to the organism they control (see Friston 2009: 300). This is why active inference is sometimes described just as "predictive coding equipped with simple reflex arcs" (e.g. Friston 2012b; Friston *et al* 2010). What imbues action (i.e. spinal reflexes) with goal directness are the *prior expectations* of the generative model, which force proprioceptive predictions towards desired sensory states (cf Van de Cruys, Friston and Clark 2020). This means that *active generative* models suffer, so to speak, of an *optimism bias*: they tend to predict sensory streams that conform to their prior preferences (Friston *et al*. 2010: 256; Friston *et al*. 2017b; Smith, Ramstead and Kiefer 2021), letting the action of spinal reflexes cancel out the prediction error ensuing from these not-yet-true optimistic prediction.

[30] Hence notice that active inference does not eliminate inverse models: it only simplifies them to the extreme, displacing them into the spinal cord (Pace Clark 2016). In fact, active inference posits "an inverse model so simple evolution could have hardwired it" (Friston et al 2010: 254).

Noticing how action and perception are closely intertwined, however, raises an important puzzle. As seen in §2, perception is a process of error minimization, in which a generative model *changes its predictions* to fit the incoming sensory input. Action, active inference suggests, is also a process of error minimization, in which *the incoming sensory inputs* are changed to fit the predictions. Isn't this a *tension* between perception and action? Why, when an agent ensues a (proprioceptive, movement inducing) prediction the incoming sensory evidence that the agent *is not* moving does not force the agent to *revise* its proprioceptive expectations and stay put? The answer to these questions lies in the notion of expected precision. Briefly put, the idea is the following: in order to avoid having to revise action-ensuing proprioceptive predictions, the model *temporarily downweights* the incoming (proprioceptive) prediction error, so to avoid that its accumulation will lead to a different prediction (Brown *et al.* 2013). Such a mechanism should operate very early in the prediction hierarchy, as changing even very early-level proprioceptive predictions would factually impede movement. Some experimental data do confirm this picture (Brown, Friston and Bestmann. 2011). Thus action, if the PP here sketched in on the right track, appears to be a special case of non-attentive perception of one's own body.

5 - Pointing forward

This concludes my introduction of PP: I have introduced all the key concepts that will be relevant later in the text.

Notice how my introduction of PP is riddled with representational terminology: the brain stores a generative *model* (which, during perception, sometimes is inverted in a recognition model); and, *prima facie*, models are representations: intuitively, a model of gas motion *represents* gas motion, and a model of a train station represents a train station. Moreover, the model is leveraged to make *predictions*. And, again, predictions seem to be representations:

they are about something which is estimated. These predictions are then corrected based on an error signal (prediction error) - but, again, the class of things that can be in error is the class of representations. It is thus natural to hold that PP is a representational theory.

In the following, I will argue that, natural as it may be, holding that PP is a representational theory is a mistake. But before doing so, I need to clarify what representations are and in what sense one can think that PP is a representational theory. I will do so in the next two chapters.

Chapter two - A field guide to representations[31]

1 - Representations: what are they?

Representations are so central in cognitive science they seemingly define it. For instance, according to Chomsky:

> "In the light of the work of the past twenty years, *it is fair to define cognitive psychology as the study of mental representations* – their nature, their origins, their systematic structures, and their role in human action" (Chomsky, 1983: 2; emphasis added)

Although this view is widely accepted, cognitive scientists often admit having no idea about what representations are:[32]

> "We, as cognitive psychologists, do not really understand our concepts of representation. We propose them, talk about them, argue about them, and try to obtain evidence in support of them, but we do not understand them in any fundamental sense." (Palmer 1978: 259)

Cognitive scientists shouldn't be embarrassed by this. They use representations as explanatory primitives: *ready-made* building blocks to be deployed in the explanation of cognitive phenomena and intelligent behavior. *Philosophers of cognitive science*, however, aim at dissecting the explanatory primitives of cognitive science, providing a rational account of them which is at least consistent with the empirical practice of cognitive science (e.g. Cummins 1991a; Ramsey 2007; Shea 2018; Rupert 2018).

Whilst there is no agreed-upon philosophical account of representations, one can point out to some features representations are (close to) universally supposed to have. In the following, I will highlight them, and then take a closer look at some paradigmatic instances of representations in cognitive science.

---

[31] Part of §2.2 reproduces material from Facchin, M. (2021a). Predictive processing and anti-representationalism, *Synthese*, https://doi.org/10.1007/s11229-021-03304-3

[32] In all fairness, however, cognitive scientists are now starting to adopt a philosophically informed conception of representation, see (Webb 2006; Poldrack 2020; Backer, Lansdell and Kording 2021).

Three *caveats* before I move on. First, I will only discuss *cognitive* (as opposed to mental or public) representations. By "cognitive representations" I designate the explanatory posits of cognitive science (e.g. the internal representation of a syntactic tree), which are sub-personal and *need not* be introspectable by a subject. In contrast, by "mental representations" I mean personal level, typically introspectable, representational states. The unqualified term "representation" refers only to cognitive representations.

Secondly, by "computation" I will always mean *generic computation*: the processing of vehicles (see below) according to rules sensitive only to specific vehicle properties (Piccinini and Scarantino 2011). I do so to avoid any commitment to specific computational styles (e.g. digital computation) and specific accounts of computational implementation.

Lastly, I aim only at providing an understanding of representations servicing my analysis of the representational commitments of PP. Thus, I will be silent on a number of important issues that are simply not functional to my aim (for instance, I will be almost silent on *imperative* representations). The following is meant as a tool for later use, rather than the ultimate truth about representations.

2 - Some features of representations in general

Here I highlight some features that representations in general (public and mental representations included) are typically supposed to have, to then focus on the content and functions of representations in two dedicated subsections.

2.1 - Representation in general: some remarks

Here, I follow (Bechtel 2008; Godfrey-Smith 2009; Mollo *forthcoming*), and consider cognitive representations as a particular variety of *representations in general*, to be understood on the template offered by *public* representations such as models, sentences, or graphs. I take

representations in general (i.e. both public and cognitive) to be *triadic relations.*[33] Something (1st *relatum*) represents some other thing (2nd *relatum*) only in some context, or to someone, or in some system (3rd *relatum*). The first *relatum* is the *representational vehicle* or simply *vehicle*, and I will indicate it with "V". Following the mainstream[34], I take vehicles to be always *concrete particulars*. The second *relatum* is the *representational target* or *target*, which I will indicate with "T". Since everything can *in principle* be the target of a representation, there are no restrictions to what targets could be.

Vehicles and targets can be *complex*. So, they can have *constituents*. For instance, a sentence (vehicle) is made out of words and represents a state of affairs ("made out" of objects and relations). I indicate constituents with an uncapitalized letter and a subscript (indicating whether I'm referring to some *specific* constituents or to constituents in general). For example, "v_x" indicates any constituent of V, and "t_a" indicates a specific constituent *a* of T.

The third *relatum* varies depending on the type (public, mental or cognitive) of the representation. The sentence "Marina mangia le mele" represents the fact that Marina eats apples *in some linguistic context*; my belief that Marina is an apple-eater represents Marina as being in a certain way *to me*, and an appropriate series of 0s and 1s represents "Martina mangia le mele" *in my personal computer*. When it comes to cognitive representations, the relevant "third" relatum is either the entire system S in which the representation is tokened (e.g. Shea 2018) or some sub-system of S, typically called a *consumer* (e.g. Millikan 1984).[35]

Now, back to cognitive representations.

---

[33] For defense, see (Peirce 1938-51; Millikan 1984; Von Eckart 1996; Menary 2007; Neander 2017).

[34] See, for instance (Egan 2019, 2020; Shea 2018; Ramsey 2020).

[35] Although, to be fair to Millikan, it should be noticed she concedes the relevant consumer might be physically located outside the system, and might even be an entirely separate system (e.g. another human listening to my words).

2.2 - Representational content

Representations belong to an *intentional kind*: their vehicles refer to their targets, or "are about" their targets in some other way. This is what I mean when I say that a vehicle bears some content (or that a representation has content). The metaphysics of content is complex and lively debated[36] - but luckily, given my purpose here, I can simply gloss over a number of contested matters.

First, contents are semantically evaluable abstract objects.[37] Hence, they must somehow relate to normative conditions of satisfaction. Conditions of satisfactions are sets of conditions the fulfillment of which determines whether the representation is successful or not (e.g. whether it represents accurately or veridically). These conditions are normative, and, however determined, it must always be possible for them *not* to be fulfilled: vehicles can always, in principle, mis-represent their targets (e.g. Dretske 1986; 1988). In other words, the possibility of misrepresentation (partially) *individuates* representational vehicles.

This entails that contents have at least the following two features: *determinacy* and (in the case of cognitive representations) *distality* (e.g. Egan 2012: 256). The relevant senses of "distality" and "determinacy" are the ones at play in the horizontal disjunction/stopping problem; that is, the problem of providing a theory of content such that the contents it delivers are neither arbitrary disjunctions of targets, nor the most proximal causes of the tokening of a vehicle (see Dretske 1986; Godfrey-Smith 1989; Neander 2017). Notice that albeit the label "horizontal disjunction problem" ties them together, distality and determinacy are two *logically independent* requirements, which can *independently* fail to obtain: non disjunctive, but purely proximal, contents are possible (Artiga and Sebastián 2018; Roche and Sober 2019).

To see why the fact that conditions of satisfaction must be such that they might fail to obtain

[36] For a review, see (Ryder 2009b).

[37] Notice that this is relatively uncontested (e.g. Fodor 1987: 10-11; Hutto and Myin 2013: X; Lee 2018: 6; Egan 2019: 247). What is contested, however, is the kind of abstract objects contents are (e.g. possible worlds, modes of presentation, propositions, *etc.*).

entails that content is distal and determinate, consider Fodor's (1987: 99-102) crude causal theory of content. Bluntly put, the crude causal theory says a vehicle V represents whichever target T causes its tokening within a system. Thus, if dogs cause the tokening of V, then V represents dogs. Suppose now a sheep causes a "wild" tokening of V. It seems intuitively correct to say that V *misrepresents* the sheep as a dog. Yet, the crude causal theory does not allow us to say so. For, if V represents whatever causes its tokening, and its tokening can be caused by dogs *or* by sheeps, then V represents dogs or sheeps; hence, its content is the disjunction (dog or sheep). So the system is *correctly* representing dogs or sheeps and the "wild" tokening of V is not a misrepresentation. Further, it could be argued that the tokening of V is not really caused by dogs (or sheeps), but by some more proximal condition, such as quadruped-shaped retinal images, or a pattern of activation *p* in the early visual cortex. Again, in this case, it seems that "wild" tokenings of V do not misrepresent dogs as sheeps, but correctly represent some more proximal condition, which happen to be disjunctively caused by both dogs and sheeps. Hence, contents that are not appropriately distal or determinate make it impossible for representations to mis-represent. Since representations are obviously capable of mis-representing, it follows that their contents are neither proximal nor disjunctive.

Notice that this conception of representational content is embedded in the explanatory practice of cognitive science. Representations, as cognitive scientists conceive of them, are about *worldly targets*[38], rather than the proximal conditions by means of which worldly "things" are encountered. Moreover, cognitive representations are about *well-specified* wordly targets, rather than arbitrary disjunctions thereof. For instance, vision scientists say (and assume) such-and-such an activation of V1 represents a *straight bar*, or that such-and-such an activation of the fusiform face area represents *a face*. They do not say that activations in V1 represent *patterns of retinal stimulations*, or that activations of the fusiform face area represent

[38] Agent's body included.

faces or face-shaped patterns of stimulation. Thus, if philosophers wish to account for representations as cognitive scientists use them in their explanatory practices, content must be distal and determinate.

I take contents to be determined by an appropriate *content-grounding* relations holding between vehicle types and target types. Some content-grounding relations will be briefly examined below (§§ 3.1, 3.2 and 3.4). As for now, I just wish to lay down some constraints on what counts as an appropriate content-grounding relation.

First, the relevant relation of relations must be *asymmetric* and *non-reflexive*. A vehicle V *is about* a target T, but not *vice-versa*; and a vehicle V is *not* about itself. A picture of an apple is about an apple, but apples are not about anything, and the picture is not about itself (Goodman 1969: 3-4).

Secondly, the relevant relation must be *reductive* (e.g. Fodor 1987: 97): it must account for what "makes" tokens of V *about* T in a way that does not presuppose aboutness (or other semantic or intentional notions). This, I believe, is essential, at least insofar cognitive representations are taken to *explain* the intelligent (that is, intentional) behavior of an agent. Accounting for the intentionality of cognitive representations in terms of intentionality would make the account suspiciously circular (e.g. Cummins 1996: 3).

2.3 - Representational function (the Job Description Challenge)

Above I said that representations form an intentional kind. This provides the *thinnest* possible notion of representation; namely, that of representations as semantically evaluable "things" (Ryder 2009a: 234-235). There are reasons as to why such a thin notion of representation is inadequate. For instance, our best accounts of the relevant content-grounding relation(s) massively overgeneralize (e.g. Ramsey 2007; Orlandi 2014; Morgan 2014). Were such a thin notion of representation accepted, panrepresentationalism would follow: everything

(or, at least, an inordinate amount of things) would count as representations. But panrepresentationalism is intuitively unappealing to most. Moreover, intuitive appeal (or lack thereof) aside, if the relevant task philosophers of cognitive science face is that of accounting for cognitive representations as they figure in the standard practice of cognitive science, then panrepresentationalism must surely be avoided: cognitive representations are *not* posited to account for the inner workings of *every* system (see Webb 2006; Tani 2007; Backer, Lansdell and Kording 2021).

For this reason, a more robust notion of representation is needed. This is why representations are ordinarily considered also to form a *functional* kind.[39] There is something for a vehicle to function *as such*; namely, as a representation of a given target in a given context or system. Compare: a newspaper surely "carries", in the relevant sense, some semantically evaluable contents: the news it reports might be *true or false*. Its pictures might be *accurate* or *inaccurate*. Yet, if I roll the paper up and use it to kill flies I'm not *using it as a representational vehicle*, and its contents are simply irrelevant for its functioning. So there really seems to be functioning like a representation (or, more properly, as a representational vehicle) as opposed to functioning as anything else. Hence, there is a specific *functional profile* of representations. To specify such a functional profile (for a given class of representational posits) is to meet what Ramsey (2007) called the "Job Description Challenge". To the best of my knowledge, such a functional profile has not yet been spelled out in its entirety. There are, however, some features that typically characterize it.

One is *decouplability*.[40] Decouplability is a hard-to-define notion, but it is typically characterized in terms of *absence of causal contact* between vehicle and target (e.g. Chemero 2009, pp. 48-49; Gładziejewski 2015b). The idea behind it is reasonably straightforward:

[39] See (Haugeland 1991; Ramsey 2007; 2016; Ryder 2009a; O'Brien and Opie 2010; Bechtel 2008: 160-161 Godfrey-Smith 2009; Williams and Colling 2017; Lee 2018: 2; Williams 2017; 2018a, b; Millikan 2020).
[40] See, for instance: (Haugeland 1991; Clark and Toribio 1994; Grush 1997; Clark 1997; Clark and Grush 1999; Webb 2006; Rowlands 2006; Pezzulo 2008; Ryder 2009a; Orlandi 2014; 2020).

prototypical instances of representational vehicles (e.g. the utterance "Mary is tall") can be tokenized even when their representational target (Mary) is not present, and thus does not causally affect the representational vehicle (or the system tokening it) in any way. Moreover, representational vehicles can represent non-existent or abstract targets, which lack any causal power. Hence, representational vehicles are surely decouplable from their targets. Lastly, cognitive representations are typically posited to account for behaviors "directed at" non non-present targets (e.g. Haugeland 1991; Orlandi 2020). When the target is present (and available to the system through an appropriate signal), the explanation of the relevant behavior does not force one to posit representations (e.g. Haugeland 1991; Clark and Toribio 1994; Clark 1997). For this reason, decouplability seems a necessary feature of representational vehicles, which sets them apart from non-representational things.

Another functional feature that separates representational vehicles from non-representational "things" is that the content of representational vehicles is *causally relevant* in their mechanical functioning.[41] Again, the reason behind this requirement is straightforward: the fact that a representational vehicle represents a given target T rather than T* is supposed to account for why a system behaved the way it did. The relevant content of the vehicle *accounts for* the production of the relevant behavior under investigation, and it is what gives the representational account of said behavior its explanatory bite (Shea 2018).

These two functional features are commonly (if not universally) considered to be functional features of representational vehicles. Yet, most likely, they do *not* exhaust the functional profile of representational vehicles.[42] So, how should one judge whether some "thing" is actually

[41] See, for instance: (Ramsey 2007; Sprevak 2011; O'Brien 2015a; Gładziejewski and Miłkowski 2017; Williams and Colling 2017).

[42] And, in fact, some accounts of the functional profile of representational vehicles impose further conditions. For instance, some require that genuine representational vehicles might generate system-detectable representational error (e.g. Bickhard 2009), or that they can account for the *proactive* behavior of a system (e.g. Pezzulo 2008; Williams 2017a; Gładziejewski and Miłkowski 2017), or that they are identifiable with discrete states of a system (e.g. Ryder 2009a: 235-238), or that representational vehicles can be productively re-combined (e.g. Rowlands 2006). None of these requirements, however, is uncontestedly accepted in the literature.

tailored to play the role of a representational vehicle? A common answer is: by judging how that "thing" is analogous to some *paradigmatic* example of a public representational vehicle.[43] The procedure is roughly the following: first, one identifies some relevant representational prototype by looking at some paradigmatic instance of a *public* representation (e.g. a map, a model, a sentence, *etc.*). Then one abstracts from the prototype the relevant functional features of the vehicle, that is, the functional features in virtue of which the representational vehicle functions *as such*. Having done so, one contrasts this "core functional profile" (see Williams 2018a: 21) with the relevant functional profile of the candidate vehicle of a cognitive representation. If the two match, then the candidate vehicle *really is* a vehicle, whose functional profile is now well understood: it functions as a representation by *functioning as* a given representational prototype (a map, a model, a sentence, *etc.*).

3 - Cognitive representations: four examples

Thus far, I have only offered a relatively general conception of cognitive representations. I believe it is now time to look at some concrete examples. I'll do so in the following, introducing four different kinds of representations which will be relevant in the rest of the dissertation.

3.1 - Receptors (and teleo-informational theories of content)

Intelligent behavior often depends crucially on a number of different environmental contingencies. A termite, for instance, might push a ball of mud following a chemical gradient, contributing to the building of the termite nest. In these cases, the relevant behavior seems to be guided by the agent's sensitivity to some environmental magnitude. It is thus tempting to think at the brain[44] as a complex measurement system, whose task is that of indicating the

[43] See, for instance, (Ramsey 2007; 2016; Gładziejewski 2015a; 2015b; 2016; Williams 2018a; b; Downey 2018).
[44] Or, more broadly, cognitive systems. Since my aim is to "get to" the structural representationalist reading of PP, and PP is a neurocomputational theory, I will often talk about brains. But the relevant claims about representations can be easily expanded so as to cover cognitive systems in general.

presence (and magnitude) of the environmental contingencies salient to the agent's conduct (e.g. Ryder 2009b: 256). The internal state of such a system *indicates* the magnitude of the relevant environmental contingency, and the whole structure is a *receptor* for that contingency.

Bluntly put, the idea behind the receptor notion of representation is this: if, in a system S, a structure V tends to respond robustly to some environmental parameter T being in state t_x (of a range of states $t_a...t_n$) by entering in a corresponding state v_x (or a corresponding range of states $v_a...v_n$), then V represents T, and each state v_x of V (in the relevant range) represent the corresponding state t_x of T (in the relevant range) (Ramsey 2003; Sullivan 2010; Boone and Piccinini 2016). Thus, according to the receptor notion of representation, cognitive representations are like (hyper-sophisticated) thermometers. Just like a thermometer represents the temperature and each *individual* state of the thermometer indicates a specific temperature, a group of neurons in the early visual cortex represents the features of a perceptual scene, and the firing of each individual neuron indicates the presence of a specific feature (e.g. Hubel and Wiesel 1962; 1968). Similarly, it is often said that the activation of individual units in connectionist architecture *indicates* the presence of a relevant feature in the input pattern, the detection of which crucially influences the output of the network (e.g. Goschke and Koppelberg 1991).

How do these responses acquire content, however? What are the facts *in virtue of which* the activation of a neuron represents (as opposed to merely causing some other neural goings-on)? Since the content of a representation is *essential* to its functioning, and since receptors function by *indicating*, it is reasonable to suppose that their content must be grounded in indication. And, in fact, receptors are often closely tied to *teleo-informational* accounts of content, according to which content is grounded in indication. Indication can be understood in a number of ways (e.g. Eliasmith 2000; Usher 2001; Neander 2017; Rupert 2018).[45] Here, I employ the

[45] To be clear: not *all* these accounts of indication are *teleo*-informational accounts of content (e.g. Eliasmith

following definition, which is intended to capture the idea of indication in terms of Shannon Information (Shannon and Weaver 1949):

> **Indication**: For all the states of V and T in a relevant range of states, v_a indicates t_a if, and only if, $P(t_a|v_a) > P(t_a)$; that is, the occurrence of v_a increases the odds of t_a being the case (e.g. Dretske 1988; 1994; Shea 2018: 76)

where, the relevant probabilistic relation between v_xs and t_xs is taken to *determine* the content of each v_x, and the *obtaining* of the relation is that in virtue of which v_xs carry the content they carry.[46]

Thusly defined, however, indication is insufficient to determine content. For one thing, the tokening of any state v_x makes *a number of things* more likely: observing that the mercury bar in a thermometer has reached the "38°" mark makes *both* the fact that the environmental temperature is 38° *and* that the pressure is *n* (where *n* is an appropriate number of bars) more likely. So content is not appropriately determinate. It might also not be distal: perhaps the mercury bar reaching the "38°" mark makes only more likely that *the temperature inside the thermometer's bulb* is 38°. In such a case, the thermometer would *not* indicate something "appropriately out there" - it would not be indicating an *environmental* contingency (see Dretske 1986; Artiga and Sebastián 2018).

To obviate these problems teleo-informational accounts of content add a *teleological* component to the relevant content-grounding relation. Given this teleological component, for V to represent T (i.e. for a thermometer to represent the temperature) and for v_a to represent t_a (i.e. for a given height of the mercury bar to represent the corresponding temperature), V must be *supposed to* indicate T, and v_a must be *supposed to* indicate t_a. Here, the "supposed to" part should be unpacked in terms of proper functions; namely the outputs the production of which

2000; Usher 2001). Yet, these theories suffer from serious problems (Artiga and Sebastián 2018) and will not be considered further here.

[46] Another way to say this is that the obtaining of that relation is the truth-maker of claims with the form "The content of v_x is t_x".

accounts for the continued reproduction of devices of a given type in spite of some selective pressure (cfr. Millikan 2017: 6).[47]

Hearts, for instance, are constantly reproduced in spite of selection pressures because they pump blood (rather than making "thump" sounds); hence they are *supposed to* pump blood (rather than making "thump" sounds). Similarly, Vs must constantly be reproduced (in spite of selection pressures) because they indicate Ts in order to be *supposed to* indicate Ts - indicating Ts must be their *raison d'etre*. Here, it is not immediately relevant to discuss what sort of selection processes can determine proper functions of representational devices, but natural selection, individual learning (with feedback) and explicit design are typically held responsible for their production (e.g. Shea 2018: Ch. 3). What is relevant to notice, however, is that such processes relativize indication to *a certain type* of organism or systems (or individual organisms or systems), thereby making the relation triadic (Nirshberg and Shapiro 2020). It is also able, at least intuitively, to "chunk down" the space of possible targets a vehicle can be supposed to indicate, thereby contributing to determine content.

To get an intuitive grip on how this might happen, consider the case presented in (Levittin *et al.* 1959). Levittin and colleagues found that a number of cells in the frog's (*rana pipens*) visual cortex vigorously respond to *net convexities* entering in these cells' visual fields. Thus, strictly speaking, the activation of these cells increases the odds that *net convexities* are present in the frog's visual field. However, given that typically the only net convexities entering in the visual field of frogs are *bugs*, and given that it is evolutionary advantageous for frogs to see (and catch) bugs, Levittin and colleagues suggested that the activity of these cells detect the presence of *bugs*, rather than net convexities in general.

[47] To be clear: the notion of proper function hinted to here is that of *etiological* proper function. It is not the only notion of proper function that has been proposed, nor the only notion of proper function that has been tied to teleo-informational accounts of content (e.g. Piccinini 2020). Regardless, it is the notion of proper function most commonly deployed in teleo-informational accounts, and the only one that will be relevant for the purpose of this dissertation.

3.2 - Structural representations (and structural similarity-based theories of content)

The receptor notion of representation takes brains (or at least some parts of them) to be measuring devices. The structural notion of representations offers a different view: brains (or at least some parts of them) are *internal models* of the environment, which control animal-environment interactions by "internally testing" the outcomes of these interactions through *simulations* (e.g. Craik 1943; Dennett 1996; Williams and Colling 2017). Rather than being seen as hyper-complex thermometers, neuronal structures are seen as hyper-complex *orreries* (Williams 2018a: 62-63): structures that recapitulate (or otherwise capture) some relevant environmental structure, and whose activity *simulates* the activity of the represented target (Cummins 1991; O'Brien and Opie 2010).[48]

In the cognitive sciences, an often used example is that of the place cells in the rat hippocampus (e.g. Shea 2018: 113-116). In the rat's hippocampus, a number of place cells is remarkably sensitive to the rat's location in space. Place cell v_a fires only when the rat is in position t_a (or close to it), and so does place cell v_b in regard to position t_b. Thus presented, place cells look simply like position-detectors; each place cell merely indicating the rat's current location. However, place cells do not *just* indicate the rat's position. Their reciprocal connections are such that place cells indicating nearby places tend to *co-activate* each other. In

[48] In the following, I will consider the *entire vehicle* V a structural representation of an *entire target* T. This contrasts with some definitions of structural representation, according to which structural representations are collections of vehicles (see Swoyer 1991; Ramsey 2007). The contrast is, however, superficial, as vehicle constituents are vehicles too. Compare: a part of a model represents a part of the modeleed phenomenon; hence the *physical part* of the model is a vehicle too. Notice there is nothing problematic in nesting representational vehicles in such a way: "John loves Mary" is a vehicle, but also some of its constituents (e.g. "John; "Mary") are doubtlessly vehicles too. That being said, I think that there is a good reason to privilege a definition of structural representations in terms of the whole vehicle V rather than its constituents v_xs. The reason is the following: when it comes to grounding the content of structural representations, it is often assumed that their content is grounded in the relevant similarity holding between V and T, and that the constituents v_x of V come to represent constituents t_x of T *in virtue of* the overarching similarity holding between V and T (e.g. Cummins 1996: 96-97). Given that representational vehicles (as such) necessarily have content, and given that (typically) the content of vehicle constituents depends on the overarching similarity between the *whole* V and T, it seems to me correct to privilege V when defining structural representations. Yet, as far as I can see, very little hinges on this matter.

this way, the spreading of activation among place cells comes to "mimic" possible journey through the environment, as in "preplay", a pattern of activation that spreads so as to *anticipate* the route the rat will follow, and "foreplay", a pattern of activation that "mimics" the route the rat just took (see Moser, Kropff and Moser 2008). In both cases, the spreading of activation seems to "simulate" the represented target; namely, the rat's route through the environment. Hence it seems that place cells do not represent just by *indicating* environmental locations; they can also *simulate* trajectories through them. The entire structure they compose does not just track the rat's actual position; rather, it models the route the rat is following.

Now, just as it was the case for receptors, structural representations are closely tied to a specific family of theories of content, namely structural-similarity based theories of content. According to these accounts, the content of a vehicle V is determined by a special kind of resemblance it bears to its target; for instance *second-order structural resemblance*.[49] According to O'Brien and Opie:

> **Second-order structural resemblance (Original):** Suppose $S_V=(V,\Re_V)$ is a system comprising of a set V of objects, and a set $\Re_V$ of relations defined on the members of V. The objects in V may be conceptual or concrete; the relations in $\Re_V$ may be spatial, causal, structural, inferential, and so on. […] We will say that there is a second-order resemblance between two systems $S_V=(V,\Re_V)$ and $S_O=(O,\Re_O)$ if, for at least some objects in V and some relations in $\Re_V$, there is a one-to-one mapping from V to O and a one-to-one mapping from $\Re_V$ to $\Re_O$, such that when a relation $\Re_V$ holds of objects in V, the corresponding relation $\Re_O$ holds of the corresponding objects in O (O'Brien and Opie 2004:11)

Sadly, O'Brien and Opie's formalism is different from the one I chose to adopt. Hence, I rewrite their definition as follows:

> **Second-order structural resemblance (rewritten):** V is structurally similar to T if and only if:
>
> (i) there's a one-to-one mapping from at least some vehicle constituents

[49] In all fairness, the relevant (i.e. content-grounding) structural similarity is typically unpacked in terms of V being *homomorphic* to T (e.g. Bartles 2006). That being said, it is hard not to notice that, most of the times, the requirement of a strict homomorphism is relaxed in order to allow from *approximate* instantiation of the relevant resemblance relation (e.g. Swoyer 1991: 470-476; Shea 2018: 140-142). Since these approximate instantiations *just are* second-order structural resemblances, I think it is convenient to spell out the relevant content-grounding similarity directly in terms of second-order structural resemblance, glossing over the problem of approximate instantiation.

(v_xs) onto at least some target constituents (t_xs); &

(ii) there is a one-to-one mapping from at least a relation *R* holding among the vehicle constituents onto at least a relation *R'* holding among the target constituents; &

(iii) For all the vehicle constituents satisfying (i), $v_a R v_b \rightarrow t_a R' t_b$ (i.e. the same *pattern* of relations hold in V and T)

Notice that in my definition I have assumed that the relevant mapping in (i) is "subscript preserving": v_a maps onto t_a. v_b maps onto t_b and so forth. This is purely a conventional notation I adopt to simplify the discussion. A schematic illustration might further simplify the understanding of second-order structural resemblance; hence, see **figure 3**

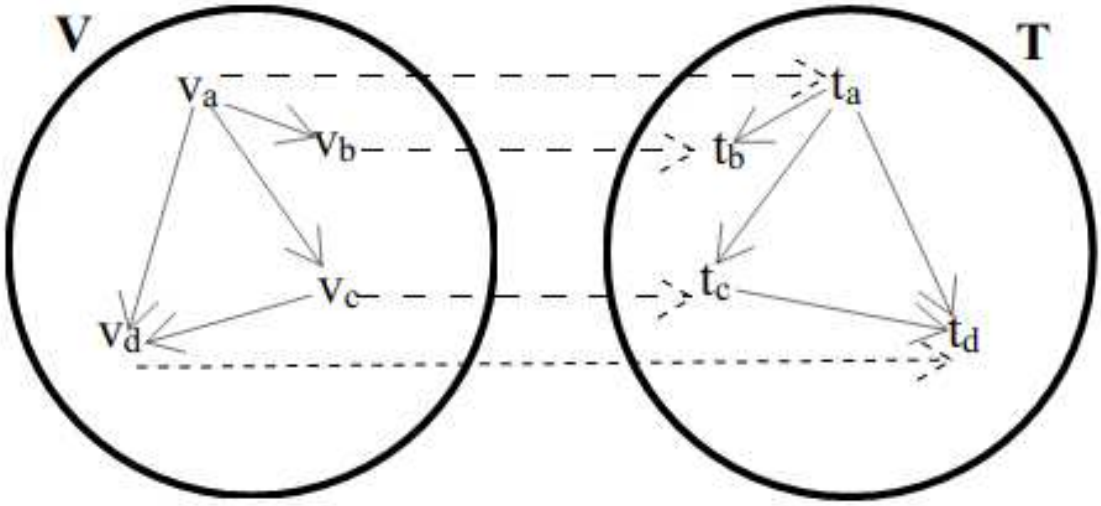

Figure 3. A schematic rendition of second-order structural resemblance. Vehicle constituents map onto target constituents (Black dashed arrows) in a way such that the same abstract pattern of relations (solid gray arrows) is preserved on both sides of the mapping. Vehicle constituents, target constituents and relations *R* and *R'* not relevant to the structural similarity have been omitted for the sake of clarity (drawing by the author).

Just as it was for receptors, this relation determines the content of V (and its constituents), and the obtaining of such a relation is what "makes" V be about T. Yet (again, just as it was for receptors) this relation is manifestly insufficient to deliver *on its own* the relevant content (McLendon 1955; Goodman 1969). To start, it is a *reflexive* relation: any vehicle bears a second-order structural resemblance to itself. It is also symmetric: if, for at least two

constituents v_x of V in a relation *R* there are at least two constituents t_x of T in a relation *R'*, then it it obviously true that for at least two constituents t_x of T in a relation *R'* there are at least two constituents v_x of V in a relation *R*. This surely is problematic: a photo of me "is about" me, but I'm *not* "about" any photo of myself. Second-order structural resemblances are also easy to come by: since the relevant relation quantifies only over "at least two", it is sufficient for two systems to have *only two* components in corresponding relations to be structurally similar. But if this is the case, then one vehicle will be structurally similar to *many* targets; hence its content will be underdetermined.

All these problems are typically faced by pointing out that the relevant (i.e. content grounding) structural similarity must be *exploitable* (e.g. Shea 2014). Informally, exploitability requires two things. First, that the relevant (i.e. similarity constituting) relation holding among the constituents of V must influence, in some systematic way, the behavior of the system. Secondly, the system must be supposed to operate (in the sense sketched in the previous paragraph) on the relevant target and its constituents. That is, the target at the end of the structural similarity relation must *matter* to the system's functioning. The conjunction of these two requirements is taken to be sufficient to avoid the problems raised above (e.g. Williams and Colling 2017).[50]

Notice two things about exploitable structural similarities, which will be relevant in the following chapters.

First, the relation holding among vehicle constituents *need not* be the relation holding among target constituents. It is thus possible to represent, for instance, relative weights with frequencies, or frequencies with gradients. This bolsters the expressive powers of structural representations, and it is especially important when it comes to *neuronal* representations, for

[50] I postpone any further analysis of exploitability to Ch. 4, when the relevant notions needed to analyze it will be put to use. There is no need to introduce them here, only for the reader to forget them.

there are not that many *exploitable* relations holding among neurons (chiefly, excitation and inhibition). The relevant point to notice is that what makes V and T structurally similar in the relevant sense is that they share a *pattern* of relations; not the fact that some *identical relation* holds among their constituents.[51]

Secondly, exploitable structural similarity entails what I will call *semantic unambiguity*. The idea is simple: if V is structurally similar to T, then it is always in principle possible to determine, for each constituent v_x of V participating the relevant structural similarity, what is the constituent t_x of T onto which v_x maps. This follows straightforwardly from (i). And the same is true for the relations participating in the structural similarity because of point (ii). Hence, it is always possible to say what each "bit" of the structural representations means in an unambiguous way: v_a, unambiguously means t_a, and $v_a R v_b$ unambiguously means that $t_a R' t_b$.

3.3 - Superposed representations

In the previous two subsections, I have briefly sketched two different kinds of representations. Importantly, both kinds have a well-determined functional profile (receptors *function as* measurement tools, structural representations *function as* simulations) and are typically linked with a specific account of content; that is, they are linked to a specific content-grounding relation.

Superposed representations, in contrast, have no well-determined functional profile, and are tied to no specific account of content.[52] This is because, unlike structural representations and receptors, superposed representations have no philosophical pedigree. In fact, they have been

[51] A concrete example in service of clarity. Suppose a vehicle V is made up, among others, by three constituents ordered by their relative *magnitude* in the following triplet (v_a, v_b, v_c) and T is made up, among others, of three constituents ordered by their relative *weight* in the triplet (t_a, t_b, t_c). Given that constituents with identical subscripts map onto each other, V and T are structurally similar: an identical *pattern* of relations holds among at least two of their constituents.

[52] So much so that some have argued they are not representations at all (Ramsey 1997; 2007). I find these arguments persuasive, but they will be irrelevant throughout the entirety of this dissertation.

introduced by connectionist researchers as *explanatory primitives*, to name the distinctive (and counterintuitive) way in which weight matrices of connectionist systems represent.[53]

The fact that superposed representations have been introduced as explanatory primitives, however, does not entail that they cannot be characterized in any way. Andy Clark, for instance, proposed the following functional characterization:

> "If a network learns to represent item 1 by developing a particular pattern of weights, it will be said to have superposed its representations of items 1 and 2 if it then goes on to encode the information about item 2 by amending the set of original weightings in a way which preserves the functionality (some desired input-output pattern) required to represent item 1 while simultaneously exhibiting the functionality required to represent item 2." (Clark 1993: 17)

The basic idea, then, is that superposed representations are representational vehicles that store *multiple* contents at the same time, thereby enabling a system to cope with a variety of task domains using a limited number of resources (Rogers and McClelland 2004: Ch. 3).

This way of representing can be made more precise in terms of vehicles being conservative over contents (see Van Gelder 1991; 1992). The basic idea is straightforward: a vehicle V is conservative over a content C just in case that the representational resources needed to render C (roughly, to represent T in some way C) are equal to V. Thus "John" is conservative over John: to represent John as John I need to token "John" and "John" has no representational credit left to spend (so to speak) to represent something in addition to John. Conversely, "John loves Susan" is not conservative over John: to represent John as John I need not token the whole "John loves Susan" and "John loves Susan" *has* representational credit left to spend to represent something other than John as John; namely Susan (as Susan) and the fact that John loves her.

Superposed representations can therefore be defined in terms of conservativeness as follows: a vehicle V is a superposed representation of a series of contents $C_1...C_n$ if, and only

[53] Strictly speaking, activity vectors *can* be superposed representations too (e.g. Smolensky 1990). Yet, throughout the entirety of the dissertation I will only refer to superposed representations while dealing with connections, so it seems reasonable to focus this short introduction on weight matrices. Notice, all the arguments I will put forth (in Ch. 4, 5 and 6) concerning weight matrices can be applied, *mutatis mutandis*, to activity vectors.

if, V is conservative over each member of the series (Van Gelder 1991; 43). Hence, the tokening of V is sufficient to *simultaneously render* the whole series of contents, and V's representational credit is, in a sense, overspent: its representational credit has been spent to buy multiple content at once. Another metaphor (proposed by Van Gelder himself, see his 1991: 45-46) is that of a point in a vehicle space onto which multiple contents simultaneously collapse.

What, however, is the relevant content-grounding relation? To my knowledge, no proposal has been advanced on this front.[54] Moreover, to my knowledge, their functional profile is unlike the functional profile of any public representation. Haugeland (1991: 66) compared them to holograms. But Haugeland himself found the similarity between holograms and superposed representations quite superficial: after all, there is a clear sense in which holograms are just regular images, whereas superposed representations are not.[55] Haybron (2000) suggested two other comparisons. One is the compression of multiple contents encoded in different sound-waves in a single sound-wave. Yet, as Haybron himself noticed, this won't do: in the compressed sound-wave, content specific frequencies can still be isolated. However, in the case of genuinely superposed representations, no vehicle part can be singled out as a conservative representation of a single content (see McClelland and Rumelhart 1986: Vol I, 176). The other comparison concerned a single numerical value stored in the memory cell of a computer.[56] This

[54] This is not exactly true: O'Brien and Opie (2006) suggest that weight matrices and activation vectors (hence, the two kinds of vehicles that can be vehicles of superposed representations) are contentful in virtue of a second-order structural resemblance relation. This proposal has never caught up in the literature, and I will criticize it in (Ch. 4: § 4.3)

[55] He also suggested that a superposed representation can be compared to a chord played by a piano, in which each individual note carries some specific bit of information about the represented target (Haugeland 1991: 67). I see two deep flaws with this suggestion. First, piano chords are not paradigmatic cases of public representations. Compare: whereas it makes intuitive sense to say that something functions as a representation by functioning *as a map*, it is very hard to grasp what is meant by saying that something functions as a representation by functioning *as a piano chord*. In fact, there is no common usage of, say, A major and C minor as representational vehicles. Secondly, it seems that in Haugeland's (underdescribed) example the note-to-target matching is *arbitrary*, and established by convention. But if this is the case, then each note is a symbol in an entirely unproblematic way, and the chord is just the simultaneous tokening of different symbols.

[56] Grush and Mandik (2002) might be read as improving on this example; but insofar as I can see the same line of criticism applies.

numerical value enters in different computations, and has a different "meaning" in each computation. This example too, however, is not adequate. To begin with, pretty much every representational vehicle can render different contents in different contexts: consider, for instance, the word "lioness" in "Nora is a lioness" and "The lioness hunts". Pretty clearly, the token "lioness" has different contents in the two cases. But surely linguistic tokens are *not* superposed representations. Secondly, in Haybron's example, the vehicle token does not render different contents *simultaneously*. It renders content C at time *t* when involved in a process *p* and *then* renders content C* at time *t** when involved in process *p**. But superposed representations *do* render multiple contents simultaneously; indeed, their ability to render different content simultaneously *identifies* them as a specific representational kind.

So, to conclude, superposed representations have been introduced as explanatory primitives. What sort of relation grounds their content (if any) and how they function as representations within the systems in which they are tokened is still undecided. Hence, their representational credentials are far from secured.

3.4 - Input-output representations (and mathematical content)[57]

In contrast to superposed representations (which are explanatory primitives that have been introduced by cognitive scientists), input-output representations have been introduced by philosophers (see Ramsey 2007) to account for the empirical practice of cognitive science. The basic gist behind them seems to be the following: computational theories in psychology suggest that brains (and other devices) *compute* mathematical functions.[58] Yet mathematical functions

[57] Strictly speaking, "mathematical content" refers to Egan's (2014; 2019) account of content in cognitive science; whereas "input-output representation" refers to a specific account of representation alayzed in (Ramsey 2007). Both notions trace back to Cummins's (1991) seminal book, and Ramsey (2020) has recently conceded that Egan's account of mathematical content is an improvement over his original account of input-output representations. For this reason, I here clump the two together.

[58] To give just one example: it is commonly assumed that prediction error is computed by *subtraction*. Hence, the physical device or devices generating prediction error must be able to compute the subtraction function.

are relations defined over mathematical objects. And computational devices (e.g. adders) do not seem to have any access to them. So how can a device compute? The answer input-output representations provide is the following: a device can compute if we allow its physical states to *stand-in for* the relevant mathematical objects; that is, the arguments and values of the function computed (Cummins 1991; Ramsey 2007; Egan 2014; 2019; 2020).

This can be done through an interpretation function which maps computational state types onto numerical values, in a way such that the physical interactions among computational states will march in step with the function to be computed. An example might clarify this point. Thus, let *p* be the prediction signal, *i* the actual incoming input signal and *pe* the prediction error signal. Saying that prediction error is computed by subtraction is saying that there is a realization function *f* such that, if $f(p)=x$ and $f(i)=y$, then the *pe* signal generated by the physical interaction between *p* and *i* is such that $f(pe)=(x-y)$, for all the instances of *p*, *i* and *pe*. When this is the case, *p*, *i* and *pe* can be said to be representational vehicles having *x*, *y* and $(x-y)$ as targets, respectively.

Following Ramsey (2007; 2020), I will call these representations input-output representations, as they represent the arguments and values that a device computing a function takes as input and yields as outputs. Moreover, I will follow Egan (2014; 2018; 2020) and call the specific content of these representations "mathematical contents", to distinguish it from the more familiar representational content of the species of representations introduced above. Input-output representations are in fact very different from the other kinds of representations I've sketched in many regards.

First, Input-output representations (Ramsey 2007) and mathematical contents (Egan 2014; 2020) are explanatory posits of computational[59] theories. Conversely, receptors and structural

[59] Ramsey (2007) originally presented input output representations as explanatory posits of *classical* computational theories, seemingly suggesting that non-classical computational theories (e.g. analog computation, connectionism and probabilistic theories of computation) do *not* posit input-output representations. I believe this is a mistake: insofar these theories are committed to the claim that cognitive systems compute functions, they

representations have no *necessary* tie to computational theories. John Locke, for instance, might be seen as proposing an account of mental representations as detectors (see Cummins 1991), and Tolman (1948) suggested that something quite like structural representations (as sketched above) were implicated in numerous cognitive tasks. Yet, neither Locke nor Tolman were computationalists. Something similar might also be said about superposed representations: albeit the term "superposed representation" was introduced by connectionist researchers (which are computationalists), something *like* superposed representations had already been proposed by Lashley (1929), who was not a computationalist.

Secondly, as noticed by Miłkowski (2017), a vehicle can be the vehicle of *both* an input-output representation *and* of a receptor/structural representation/superposed representation. A single "thing" can play the role of an input-output representation (in virtue of the role it plays in the computational economy of the system) *and* function as a measurement tool, or as an inner model. This also means that a single representational vehicle can in principle carry two contents: a *mathematical content*, which is tied to it being the vehicle of an input-output representation, and a *representational content*, grounded in its indicator capacities or in a relevant structural resemblance.

Notice that mathematical and representational contents are quite different. The former is narrow (Egan 2014), the latter is typically wide. The former is determined *only* by the computations a system performs, the latter is determined by some privileged naturalistic relation. The former can represent only mathematical objects, the latter has no such restriction. This invites one to wonder how these contents might be related. On Egan's (2014; 2020) account, the relation is loose to non-existent: she contends that representational vehicles have mathematical content intrinsically, whereas representational contents are just a matter of

have to posit that certain physical states "stand-in for" the arguments and values of the function computed. Thus, it seems to me that both classical and non-classical computational theories posit input-output representations. See also (Cummins 1991, Ch.11).

interpretation.[60] Others (e.g. Wiese 2017; 2018) contend instead that these two kinds of content are deeply related. But since these proposals form the backbone of the structural-representationalist reading of PP, they will be discussed in the next chapter.

Lastly, a point on how mathematical contents are determined. What is the relation grounding them? Cummins (1991) only speaks of an "interpretation function", which systematically maps vehicles to contents according to a rule, and so does Egan (2014). Yet, they both leave the relevant interpretation function almost unanalyzed.[61] It is thus not clear whether finding an appropriate interpretation function (whatever it might be) will provide appropriately *naturalized* mathematical contents. Egan (2019) is explicitly skeptical in this regard. In her view, it is unlikely that mathematical contents will be naturalized, as there can be no naturalistically respectable relation holding between representational vehicles and numbers.

Yet, in other publications (Egan 2010; 2020) her account of mathematical contents seems to entail that mathematical contents are determined by *computational implementation*.[62] In these accounts, the interpretation function is accompanied by a *realization function*, which specifies how physical states of computational devices should be clumped together in vehicle types. For instance, the realization function might specify that firing rates below a certain threshold x all are instances of the same vehicle type, whereas firing rates above x are all instances of a different vehicle type. In this way, the realization function allows us to see the neurons as computational devices with two well-defined internal states relevant to the

[60] In Egan's view, representational contents are assigned by researchers "from the outside" of the system based on pragmatic considerations. This is because, in her view, representational contents are not part of cognitive-scientific explanations proper; rather, they form a non-explanatory gloss that allows researchers to connect the computational formalism (the *explanans*) with cognitive phenomena and intelligent behaviors (the *explananda*) in an intuitively pleasing way.

[61] This is almost spectacular in Cummins (1991: 102-108). In that passage, Cummins introduces the notion of a *direct interpretation* of a device; that is, the interpretation that assigns, to each vehicle, the mathematical content it *actually* carries. But then he candidly admits: "But I must confess I don't know how to define directness" (Cummins 1991: 104).

[62] Dołega (2017: 15) noticed this too. In his words, the mapping assigning mathematical content "is supposed to obtain between the computational description and its physical vehicles manipulated by the computational mechanism". But to specify such a mapping *just is* to specify (at least in part) the relation of computational implementation.

computations the neuron performs, thereby unraveling how neurons (or other devices) *physically implement* the relevant computations they execute.

Crucially, in her most recent account, Egan (2020: 26) characterizes the vehicle types at one end of the realization function as *numerals*; that is, as *representations of numbers*. But to conceive something as a representation of a number *just is* to assign it a mathematical content. Hence, if it is correct to say that the realization function is an account of implementation, and that such an account maps physical state-tokens onto vehicle types which *are numerals*, then it seems correct to say that the account of implementation determines the relevant mathematical content. In a less convoluted way, the idea is this: if it is literally true that a device computes the subtraction function, then, *in virtue of the fact* that the device literally computes subtractions, it is literally true that certain states of the device will represent the arguments and values of the subtraction function.[63]

Thus, if an appropriate (i.e. naturalistic) account of computational implementation mapping physical states onto *numerals* can be provided, then mathematical content can be naturalized. But, as signalled by the second caveat in §1, I do not wish to take a stance on computational implementation here. Hence, for present purposes, I will *assume* that such an account of implementation can be provided, and conclude that input-output representations bearing mathematical contents have strong representational credentials.[64]

[63] Cummins (1991: 93) apparently concurs: "There is a sense in which an adding machine adds because it represents numbers, but there is a more important sense in which it represents numbers because it adds".

[64] Upon further reflection, I discovered that this is a *huge* concession: in fact, no naturalistically acceptable theory of computational implementation seems able to deliver well-determined mathematical contents. See (Facchin *submitted*) for the argument.

Chapter three - The structural-representationalist view of predictive processing[65]

1 - Representations in the predictive processing framework

PP claims brains are fundamentally in the task of minimizing prediction error, thereby realizing a form of Bayesian inference (Friston 2009; 2010). But inferences, at least *prima facie*, are processes that manipulate representations. Moreover, prediction error is computed relative to the *predictions* issued by a (generative) *model* of how sensory states are caused. The explanatory lexicon PP leverages is thus ripe with representations, so much so that many philosophers[66] argue that PP would not be *intelligible* without them (e.g. Clark 2015a, 2016: 291-294). PP is also said to be the "last word" of representationalism, since its representational posits allegedly put an end to the philosophical debate on the role of representations in cognitive science (Clark 2015a; Williams 2017; Constant, Clark and Friston 2021).

Here, I illustrate the most prominent representationalist interpretation of PP, namely the structural-representationalist interpretation. It claims that PP posits one fundamental kind of representational posits; namely generative models, which, *being models*, are naturally understood as structural representations. (e.g. Gładziejewski 2016; Williams 2017; 2018a; 2018b; Kiefer and Hohwy 2018; 2019; Wiese 2018).[67]

As clarified in the previous chapter, representations have both a specific functional profile and some content. I will now explain what those are in the case of generative models interpreted as structural representations, starting from the former.

---

[65] Part of §4.1 reproduces material originally presented in (Facchin 2021a).

[66] There are exceptions (e.g. Orlandi 2014, 2016; Downey 2017, 2018), but these are few and far between. See (Pezzulo and Sims 2021) for a nice survey of the conceptual landscape.

[67] Since this view is shared by both "radical" or action-centric (Clark 2015b) and "conservative" or inference-centric (Hohwy 2013) accounts of PP, I here clump the two together. I feel safe to do so, because these two accounts are closely aligned on a number of issues, representations included (Gładziejewski 2017).

2 - Generative models as structural representations: the functional profile

As previously seen (Ch. 2: § 3.2) structural representations function as models or simulations of their targets. How do generative models fulfil this functional role? The answer the structural-representationalist reading provides is that generative models function as models by being *effective control structures* that are *decouplable* from their targets and that allow for *error detection* (e.g. Gładziejewski 2016; Williams 2017; 2018a). Let me unpack.

2.1 - Generative models as effective control structures: structural similarity and action-guidance

Consider the following usage of the word "model" in the PP literature. According to Anil Seth:

> "The body of a fish can be considered to be an *implicit model* of the fluid dynamics and other affordances of its watery environment." (Seth 2015: 6, emphasis added).

similarly, according to Karl Friston:

> " [...] an agent does not have a model of its world—*it is a model*." (Friston 2013: 32, emphasis added)

These two passages point to the *thinnest* notion of model present in the PP literature: *models-as-control structures*. Recall that, on the view active inference offers, motor *control* depends crucially on how brains and bodies are wired (Ch.1: §4). Indeed, being active inference a variant of the so-called equilibrium point hypothesis of motor control (Friston 2011), motor control heavily depends on bodily features, such as the body's synergies and passive dynamics (Feldman 2009; Friston and Parr 2019). Moreover, bodily features such as type and location of sensors and actuators typically mirror (in a sense clarified below) to a non-trivial level the relevant features of an agent's niche, thereby providing an *implicit* model of the niche.[68]

[68] See (Pfeiffer and Bongard 2007: Ch. 3 to 5) for a systematic exploration of these themes devoid of the model-based talk. See also (Linson *et al*. 2018) for a PP exploration of these themes.

Similarly, since the brain controls the barrage of input it receives, it will be an *implicit model* of the generative process generating these inputs. But *why* should these control structures be labelled as (implicit) models, rather than, say, as *control structures*?

The answer is the following: in order for a control structure to be effective, it must be *homomorphic* to the structure controlled (Conant and Ashby 1970; Seth 2015). Thus, effective controllers are *action guiding structures* that are *structurally similar* to their controlleé.

An example to clarify. Suppose I use a thermostat to regulate the temperature of a room, so as to keep it to a specific value *x*. When the temperature drops below *x*, the bi-metallic strip of the thermostat must be *straight*, so as to close the circuit that powers the heating system. But when the temperature rises and exceeds the value *x*, the bi-metallic strip must be curved, opening the circuit and depowering the heating system. Now, in the present case, there is a controlled system (the room's temperature) which can be in any of two states: above or below *x*. There is also a controller, namely the bi-metallic strip, that can be in two states: it is either curved or straight. There is a one-to-one mapping between the states of the two: when the temperature is below *x* the strip is straight, and then the temperature is above *x* the strip is curved. Furthermore, such a mapping *preserves* the relations holding among the states: the *hotter* the temperature, the *bigger* the curvature. Thus, a structure-preserving mapping holds between the controller and the *controlleé*, and the two are structurally similar in the sense clarified in (Ch. 2: § 3.2; see also Ch. 5:§§ 2.1.1 - 2.1.2).

Notice also that their structural similarity *guides* the thermostat's "actions" to control the temperature. This idea should be *minimally* unpacked[69] in terms of counterfactual statements of the following form holding true: "weren't the controller structurally similar to the *controlleé*, then its actions would non-accidentally cause it to fail to aptly control the *controlleé* in a range of circumstances" (see Gładziejewski and Miłkowski 2017: 341-348). For instance, weren't the

[69] A stronger unpacking in terms of *exploitability* will be offered in the next chapter (Ch. 4:§ 1.2).

bi-metallic strip curved when the room's temperature exceeds values *x*, then the room's temperature would continue rising, and so the bi-metallic strip would fail in keeping the room's temperature at value *x*.

The example is doubtlessly simplistic[70], but suffices to show in which sense effective controllers are models in a *non-trivial* sense: they are models in a non-trivial sense because they structurally resemble the controlled plant, and that structural similarity is what *makes* them effective controllers. Hence, models-as-control structures thus bear the following two features:

> **Structural similarity**: An effective control structure V of a system T is structurally similar to T
>
> **Action guidance**: The success of a control structure V in controlling T non-accidentally depends on V being structurally similar to T in an appropriate way (see Conant and Ashby 1970)

These are the first two functional features of (generative) models, which spell-out the functional profile of models-as-controllers.

2.2 - From control structures to structural representations

Prima facie, control structures, such as thermostats, infrared receptors of garage doors, switches, and the like bear no content. But content is an essential feature of representations (Ch. 2: §2.2). Moreover, as seen in (Ch 2: § 2.3), it is typically assumed that representational vehicles are *decouplable* from their targets. But surely many control structures are not decouplable from the plant they control. A thermostat cannot control the temperature of a room unless it is somehow coupled with the room. So, the notion of models-as-controllers is not really a *representational* notion, and models-as-controllers are not *really* representations. But it is possible that some models-as-controllers are *also* decouplable and contentful. Hence, by imposing further constraint on the notion of models-as-controllers one might single out a subset

[70] See (Baltieri, Buckley and Bruineberg 2020) for a less simplistic example, in which a Watt Governor is treated as a generative model of the steam engine.

of models-as-controllers which are also models-as-structural-representations. This is exactly what the structural-representationalist interpretation of PP does (e.g. Seth 2015; Gładziejewski 2016; Williams 2017; 2018a).

Consider first, decouplability. One intuitive way to define it could be as follows:

> **Decouplability***: A control structure V is decouplable from T when it can successfully control T even in absence of any signal coming from T (e.g. Haugeland 1991: 62)[71]

Are generative models, as PP conceives of them, *decouplable* from their targets? If the answer were positive, then generative models would appear to be models in a *more representational* sense of the term - they would be more than *just* Models-as-controllers.

The answer is positive: generative models are decouplable from their targets. To see why, consider, first, active inference. As previously exposed (Ch. 1: §4) forward models are needed because the signals from the sensory periphery reach the brain *after a delay*, and so they are not available to the control system when needed. But this means that our brain can (and does) control our body *in absence of the reafferent signals coming from the body*. And, according to PP, our brain "just is" a forward model. Indeed, on the account active inference offers, motor control is essentially *proactive* in nature: in order to act an agent *first* generates a non actual stream of predicted sensory inputs, and *then* cancels out the error relative to the non-actual prediction through movement. But clearly this entails the "endogenous generation" of sensory inputs to which the system is *not yet coupled*; namely the ones that *will eventually* be brought about through movement (cf Gładziejewski 2016)

Importantly, as Grush (1997; 2003; 2004) repeatedly voiced[72], the fact that generative

[71] As I will clarify in (Ch. 4: § 1.3) the structural-representationalit interpretation of PP actually uses a slightly stronger (i.e. more restrictive) notion of decouplability; namely, it defines decouplability as the *absence of causal contact* between V and T. But not all forms of causal contact are signals: my shirt is in causal contact with my skin, but it is not *conveying a message* to my skin. This is why decouplability appears here with an asterisk.

[72] A historical note. Grush *did not* make that observation while dealing with PP. He was dealing with forward models as *special purpose* generative models as seen in (Ch. 1). Regardless, forward models are generative models in the relevant sense, hence the observation can be painlessly transposed in PP

models can predict not-yet-received-inputs naturally suggests that such models can (at least in principle) function *in absence of* environmental stimulation. This naturally suggests that generative models can be used *offline*, to engage in predictions regarding *counterfactual* scenarios. But surely counterfactual scenarios, being counterfactual and thus non-existent, send no signal to the system (See Gładziejewski 2016; Williams 2018a).

This should come as no surprise. In fact, the wake-sleep algorithm examined in (Ch. 1: § 2.2) leverages precisely that insight. Recall: according to the "wake-sleep" training schedule, the generative model is used to train the recognition model (i.e. the internal model mapping input patterns onto labels). It does so by spreading an *endogenously generated* pattern of activation from the output layer (containing the labels) to the input layer, so as to "tell" the recognition model where to map patterns of that kind. The important thing to notice here is that such a learning procedure makes explicit the fact that generative models can routinely work offline, without the guidance of any incoming stimulation. More sophisticated generative models-based learning procedures can almost "throw away the world", and let the agent learn inside an entirely "hallucinated dream" (see Ha and Schmidhuber 2018a). Strikingly, the body of knowledge and skills learned inside the "hallucinated dream" can then be transferred to guide *online* action in the actual environment with a high degree of success, testifying the power (and usefulness) of such off-line, generative models based learning procedures.

There is a last functional feature that, according to the structural-representationalist reading of PP, singles out models-as-structural-representations from models-as-controllers. Unlike simple control structures, models-as-structural-representations allow for *representational error detection*:

> **Representational error detection**: A control structure V can detect representational errors if:
>
> (a) V is equipped with a feedback channel that, by monitoring T, can detect failures in control, *or*
>
> (b) V is equipped with a feedback channel that, by monitoring T, allows a

> comparator to detect discrepancies between the expected and actual states of T (Gładziejewski 2015b: 80-81)

Surely generative models, as PP conceives of them, satisfy this condition: after all point (b), by design, simply *describes* how prediction error is computed (Gładziejewski 2015b: 81). Recall, for instance, the learning algorithm employed by Bongard, Zyklov and Lipson (2006) to enable a simple robotic agent to infer its own bodily morphology. First, the robot moves randomly, to generate some sensory data. It then uses the data thus acquired to "build" a series of bodily models, which are then consulted so as to choose the course of action upon which the predictions of all models disagree the most. The action is then executed, and the models which predicted its sensory consequences the worst are eliminated. Why? Well, because they evidently are *incorrect* models, since they are unable to account for the relevant incoming input. In this way, active inference allows agents to detect the *representational* problems of their models. As Gładziejewski (2016: 580) puts it: "the size of prediction error signifies for the system whether (or to what extent) it got things wrong representationally". Notice that, once again, such a form of error detection is *essential* to the functioning of generative models, at least as PP conceives of them. In fact, as seen in the previous chapter, PP conceives both perception and action as processes of error minimization. Weren't generative models able to generate system detectable error, the entire theoretical apparatus of PP would crumble to the ground.

Tacking stocks: generative models, as PP conceives of them, are a specific sub-type of models-as control structures: they are not just *controllers* bearing a *structural similarity* to their targets. They are *decouplable* effective controllers able to *detect their representational error*.

Do these four functional features specify a *representational* functional profile? The answer the PP literature provides is overwhelmingly positive (e.g. Gładziejewski 2016; Williams 2017; Wiese 2018), as these four functional features identify the core functional profile of paradigmatic *public* representations, such as cartographic maps. Maps are structurally similar

to their targets: a map of a city preserves the *pattern of spatial relations* holding between various points of the city. Such a structural similarity guides the actions of the system using the map (normally, a person): we can use maps to choose which road to take, for instance. In doing so, we exploit the structural similarity holding between the map and the terrain: if we see that the map displays A above B, and we wish to reach A from B, we head north. Maps are obviously decouplable for their targets: my map of Madrid need not be in any causal contact with Madrid in order to function. Maps also afford us the detection of their representational error through action. For instance if, by using a map, we reliably get lost, then we deem the map *inaccurate*, and seek for a different way to navigate the territory.

3 - Going towards content: control structures modeling the environment

As presented above, generative models seem special (decouplable and error-detecting) control structures. But control structures need only to be structurally similar to the *controlled plant* (cf. Chemero 2009: 60-65; Kelin 2018).[73] This seems to stand in the way of their representational status: standardly, representations in cognitive science are thought to represent the external world. Control and representational status appear here to clash. How, then, can generative models *represent* the external world, rather than merely *controlling* the motor plant?

The structural-representationalist view suggests that, by repeatedly controlling the agent's actions, and by improving this control, generative models *end up* modeling the salient regularities of the environment. Hohwy makes the point vividly:

> "Imagine being charged with plugging holes in a large, old, and leaking dam. [...] The occurrence, frequency, and nature of the leaks all depend on the water pressure on the other side [...] but you do not know anything about that. Your job is just to minimize overall leakage.[...] After a while you begin noticing patterns in the leaks [...] such knowledge of leakage patterns will allow you to be better at anticipating where leaks will be and plug up in advance. [...] Eventually you will have very efficient patterns

[73] Technically speaking, this is a simplification: what generative models model and control through active inference is the generative process yielding their data: a controlled plant actively coupled to a world (or niche in the world). Kirchhoff and Kiverstein (2019: 57-59) make a similar remark.

> of leak plugging, and the structure of the mechanical contraption will then carry information about the causal structure of the cause impinging on the other side of the dam [...]. The crucial bit, however, is that in achieving this successful representation of the causal structure of the world beyond the dam, you didn't have to try to represent it. All you had to do was plug leaks and be guided in this job by the amount of unanticipated leaks. Similarly, all that is needed to represent the world for the human brain is hierarchical prediction error minimization" (Hohwy 2013: 62-63).

In the next section, I will say something more about this form of "carrying information", sketching how the structural-representationalist view articulates the relevant structural similarity holding between the generative model and the environment. Here, I will instead sketch why, according to the structural-representationalist view, effective control does not stand in the way of representation.

Insofar as I can see, there are two broad, variously articulated, reasons as for why effective control can be thought to stand in the way of representation.

The first depends on (broadly speaking) evolutionary considerations. If brains have been selected by natural selection to control agent-environment interactions, they should be geared towards *survival and reproduction*, rather than *truth*. They should prioritize *effectiveness* over *accuracy*. In short, natural selection does not care about representational properties, so we shouldn't expect them to be selected. Brain representation, if at all present, should be "narcissistic", incorporating any survival enhancing distortion (Churchland 1987; Cummins 1996; Aikins 1996). And this, clearly, stands in the way of a generative model-environment structural similarity.

The second depends on the observation that many action-salient properties are not objective properties of the environment. If generative models aim at successfully controlling an agent's action, they must be sensitive to properties such as the *dangerousness* of a predator, the *attractiveness* of a potential mate, the *safety-ness* of a burrow, and so forth. But there is no physical structure "out there" that is objectively attractive, dangerous, or safe. This seems to shatter the generative model-environment structural similarity (e.g. Anderson 2017; Dołega

2017: 15-16).

The structural-representationalist interpretation of PP deals with these worries as follows (e.g. Williams 2017; 2018a).

To start, it points out that, in a sense, generative models do not capture the objective structure of the environment. Generative models are always models *of some specific data*, which, in the case of agents, are (at least partially) determined by the agent's sensors and transducers. These capture the inputs produced by middle-sized objects (e.g. chairs) and properties (e.g. a surface's texture). Commanding variables corresponding to such middle-sized objects and properties is typically *enough* to explain and predict one's sensory states (Gładziejewski 2021). To account for why I'm now sensing a strong, hot and bitter sensation in my mouth it seems sufficient for my generative model to command the hidden variable "coffee" - there is no need for it to venture into the depths of our fundamental physics. As Hohwy puts it:

> "It is a mistake to think that just because the brain only does inference, it must build up its internal model like it was following a sober physics textbook. As long as prediction error is minimized on average and over the long run, it doesn't matter which model is doing it." (Hohwy 2016: 20).

Thus, generative models do not model the objective physical structure of the world, only its "middle-sized", *partially agent relative*, rendition. But the variables describing such a middle sized rendition (e.g. chair) capture patterns that are *really and objectively* present in the data, even if they partially depend on the agent's physical makeup. So, agent-relativity does not, *by itself*, rule out objectivity: agent-relative properties need not be *illusory*.

Secondly, the structural representationalist reading of PP points out that generative models model the agents' bodies - active inference would be impossible otherwise. Bodies are rich sources of sensory signals, and so are among the things a generative model models. This is important to notice in order to face the second objection. For, although dangerousness, attractiveness and the like are not properties of the environment, there really are *objective*

bodily responses to these properties trigger. We react to dangerousness by releasing adrenaline, building muscle tension and increasing our heart rate. And we react to safety-ness by doing the opposite. Disgustingness makes us coil backwards, contracting thoracic muscles to limit air intake. These responses create a perfectly objective stream of multimodal (extero-, proprio- and viscero- ceptive) inputs that are part of the data a generative model tries to predict. And a good way to predict that stream is by including an appropriate hidden variable in the generative model, which captures the causes of such responses. Disgustingness, dangerousness and safety-ness look exactly like those variables (Williams 2017; 2018a; Clark 2018), that usefully predict our *objective* bodily responses by tracking their (most likely disjunctive, and surely agent-relative) causes.

What, then, about the "narcissistic" and biased nature of generative models? Does it stand in the way of representation? The answer the structural-representationalist reading of PP provides is negative, because, *at least when it comes to models*, there just is no un-narcissistic and unbiased representation. All models are idealized, selective, and to a degree distortive, not just action-guiding ones (Williams and Colling 2017; Gładziejewski and Miłkowski 2017). In fact, scientific models are partial, idealized, and distortive too (Giere 2004), but this does not prevent (at least some of) them from being paradigmatic cases of *truth-aiming and objective* models. The Mercator projection famously distorts the size of land masses far from the equator. Yet, this does not prevent maps using the Mercator projection from bearing a *partial homomorphism* to their targets (in this case, the Earth surface). And partial homomorphisms are all that is needed in order for an objective structural similarity to be present.[74]

Now, the similarity between generative models and scientific ones should not be exaggerated. The fact that truth-aiming and objective scientific models are partial and distortive

[74] It might also be worth noticing that, from an historical point of view, Clark's (2013b: 103-105) paradigmatic example of an action oriented representation *just is* a structural representation; namely, Matarics's (1991) "spatial map": a map coding for landmarks in terms of *perceptuomotor* signals and roughly mimicking the spatial map found in the rats' hippocampus. See also (Tani and Nolfi 1999) for a *predictive* spatial map of that sort.

(while still being structurally similar to their targets) is a hefty observation in favor of the claim that the "narcissistic" nature of generative models does not *stand in the way* of their representational status. Yet, unlike scientific models, generative models are *primarily* controllers. As such, they are not *aimed at* truth, and do not follow truth aiming policies such as the ones scientific models are subjected to (Bruineberg, Kiverstein and Rietveld 2018; Williams 2018c). This is because generative models need not *always* fit themselves to the data: they can also fit the data to themselves through active inference. The same is not true for scientific models: ideally, if the data collected speak against a scientific model, the model is discarded in favor of a better one.[75] This is an important part of what makes scientific models *truth-aimed*, and so it must be admitted that generative models are not truth-aimed in this specific sense (Clark 2015a, b; Williams 2017).

Yet, albeit generative models are not truth-aimed, they are, and *must often* be, "truth-stumbling-upon". In order to *successfully* control an agent's interaction with the environment, generative models *must*, to some degree, get things right. A generative model that simply avoids modelling an animal's predators will *not* help the animal to interact with its environment successfully. Although not truth-aimed, generative models must, to an extent, be accurate and truth-sensitive.

4 - Content: the relevant generative model-target structural similarity

So, the fact that generative models are first and foremost controllers does not stand in the way of their having *also* a representational status. And, in fact, as said above, they are typically considered to be structural representations. Hence, they should be structurally similar to their representational targets. How should this structural similarity be conceived? What are the

[75] This picture is simplified in a number of important respects (e.g. it makes no mention of auxiliary assumptions), but a full rendering of model testing and model choice in science is beyond the scope of this chapter.

contents of generative models? I address these two questions in turn

4.1 - The model-target structural similarity

Gładziejewski (2016), conceives the relevant generative model-target structural similarity along the following lines. Generative models can be thought of as complex graphs of the sort briefly seen in (Ch. 1,§ 2.1, **figure 2**). Hence, they can be conceived of as sets of nodes (or variables) connected by edges, which stand for probabilistic relation holding among variables. On the view Gładziejewski (2016: 572-573) offers, such a graphical model is structurally similar to the environment in virtue of the following mapping relation.

First, each node (or variable) encodes the *likelihood*[76] of the corresponding cause generating any given observation (or pattern in the input data). In his own words:

> "Worldly causes are thus represented in terms of the likelihoods of producing different sensory patterns in the system." (Gładziejewski 2016: 572; references omitted)

Secondly, the way in which the values of the variables (or nodes) changes over time "mirrors" the way in which environmental causes interact with each other; hence each *probabilistic dependence relation* among variables can be mapped onto a corresponding *causal relation* holding among worldly causes (*ibidem*: 573). Lastly, the prior probability[77] of each node in the graph should correspond to the prior probability of a given environmental cause.

Kiefer (2017: 14; 2020), proposes a different mapping rule. Inspired by the computational model presented in (Hinton and Sejnowski 1983), he proposes that each processing unit (i.e. node in the model's graph) corresponds to a proposition describing an environmental state of affairs, and the probability of the unit being "on" corresponds to the probability of the corresponding environmental states of affairs to be the case. Connections between processing

[76] Where the likelihood is the probability of an observation, conditioned over a cause (which is supposed to be correct).
[77] Roughly, the probability of a cause prior to any observation.

units (and their numerical weights) are instead mapped onto the inferential relations holding among propositions. Thus, for instance, if the proposition describing t_a (e.g. the glass is empty) strongly justifies the proposition describing t_b (Mary finished her martini), then there will be a large positive connection running from v_a to v_b. This proposal has been further fleshed out in (Kiefer and Hohwy 2018). On the view Kiefer and Hohwy now offer, the patterns of connections among units mimic inductive inferential transitions or "material inferences"; that is, the sort of inferences that allow one to infer "It is raining" from "The street is wet" (see Kiefer and Hohwy 2018: 2393). In this way, the pattern of connections among the units (i.e. nodes in the graph, and therefore variables) mimic the gross patterns of causal relations connecting the worldly states of affairs.

Although Gładziejewski and Kiefer propose two different model-environment structural similarities, their accounts agree on two fundamental points.

First, they both accept that the representational vehicle (the machinery instantiating the model) is *constituted by* discrete nodes representing variables standing in various relations of probabilistic dependence.

Secondly, they both accept that if two environmental states of affairs t_a and t_b causally interact in some way *R'*, then the corresponding nodes of the model v_a and v_b bear a relation of probabilistic dependence *R*. Hence, on both accounts, the *topology* of the model bears at least a second-order structural resemblance to the causal structure of the environment (see Williams 2018a: 106).

Generative models can also be conceived (in a simplified way) as a deterministic function of (nested) environmental causes plus estimated noise (see Ch 1: § 2.1):

- $c_2 = f_3(c_3) + \omega_3$
  $c_1 = f_2(c_2) + \omega_2$
  $s = f_1(c_1) + \omega_1$

If we conceive generative models in this way, then:

> "Everything describable by such sets of equations is part of the content of the brain's generative model: the equations define relations between parts of a system and thus provide a structuralist description." (Wiese 2018: 216)

Notice that these equations roughly describe the *computational* profile of each layer of the model[78]; that is, the predictions it *outputs* given the *inputs* it receives. Hence, adopting this perspective clarifies that the structure of the environment is "mirrored" in the *computational* structure of the model.

Now, as seen in (Ch. 2: §3.4), the fact that a device computes a function determines the *mathematical contents* of that device.[79] Hence, according to the structural-representationalist reading of PP, mathematical contents determine (at least partially) representational contents. Wiese (2017; 2018) is pleasingly explicit on this matter.[80] He writes:

> "[...] lest ascriptions of cognitive content become arbitrary, they must at least be constrained by mathematical contents entailed by computational models. [...] The more complex the computational model, the more the mathematical contents constrain the set of possible ascriptions of cognitive contents (which are compatible with the computational description). In principle, a computational model could be so complex and specific as to allow only a very limited set of cognitive content ascriptions." (Wiese 2017: 724)

even more explicitly:

> "[...] if hierarchical models in PP are structural representations, this means that contents carried by the representational vehicles are (at least partly) determined by their structure. [...] As we will see, this structure is

[78] To be precise, the topological/graphical structural similarity indicated by Gładziejewski and Kiefer describes the computational profile of the model too, though in a way that would strike many of us as "intuitively less computational". In fact, the topology of a graph can readily be turned into a series of equations detailing the relations of conditional (in)dependence holding among the variables represented by the nodes.

[79] Notice that in the case at hand, the vehicles of mathematical contents are *vehicle constituents* of the structural representation (the generative model). An equation such as $s = f(c) + \omega$ captures the functioning *of the entire model*; that is, the entire vehicle V representing T. But if such an equation correctly describes V, then it follows that V has at least some constituents v_a and v_b the state of which ranges in a way such that it captures the range of values variables c and ω can assume. And, at least if V is an accurate representation, the target constituents t_a and t_b upon which v_a and v_b map must be able to occupy a similar range of states - otherwise the entire target *would not be* describable by something of the form: $s = f(c) + \omega$ and V and T would not be structurally similar (*ex hypothesis*).

[80] Notice that, by so arguing, Wiese significantly departs from Egan's view on content as described in the previous chapter.

> determined by mathematical contents." (Wiese 2017: 726)

in his view, some computational PP description might also be so stringent so as to *completely determine* the relevant representational content the generative model represents; for instance:

> "Computational descriptions in PP models not only specify mathematical contents, they also specify at least some cognitive contents. For instance, according to the theory of active inference developed by Friston and colleagues, action is not brought about by motor commands, but by predictions of the (perceptual) changes that will be brought about by the respective actions. Motor commands and perceptual changes are not mathematical contents, they are cognitive contents. [...] Apart from that, optimizing precision estimates is typically regarded (by proponents of PP) as the computational mechanism underpinning the allocation of attention. This also entails a cognitive interpretation of mathematical contents." (Wiese 2017: 733; references omitted)

Wiese's view seems to be implicitly shared by most (if not all) the proponents of the structural-representationalist reading of PP, because it is *entailed* by their explicitly formulated claims.[81]

Thus, for instance, both Gładziejewski (2016) and Kiefer (2017; Kiefer and Hohwy 2018; 2019) define the relevant structural similarity holding between generative models and their representational target in mathematical (and, more specifically, probabilistic) terms. As seen above, Gładziejewski defines the relevant structural similarity in terms of *priors and likelihoods*. But priors and likelihoods are *functions* outputting values (mathematical contents) ranging between 0 and 1. Similarly, Kiefer holds that the probability associated with an hypothesis being true is represented by the probability of a unit being in the "on" state. But such a probability is the result of a complex *mathematical function* computed by the network. It thus seems that in both Gładziejewski and Kiefer's case the relevant structural similarity which determines the relevant representational content of the generative model is at least

[81] To be clear, I'm here assuming (for the sake of clarity and ease of exposition) that all proponents of the structural-representationalist reading of PP are committed to mathematical contents in the way Wiese is. Whilst this assumption might be strictly speaking false, there's a sense in which the commitment to mathematical contents is not *necessary* to articulate this point (as I will briefly discuss in Ch. 6: § 2.3). So, the assumption is fairly innocuous, and the point I'm articulating here applies whether proponents of the structural-representationalist reading of PP are committed to mathematical content or not.

partially built upon the mathematical functions the network computes. As a consequence, the mathematical contents involved in the computations of those functions end up at least partially determining the representational content of the generative model.

Clark (2015a: 2) seems to endorse an even more extreme position. In his own words:

> "To naturalize intentionality, then, "all" we need do is display the mechanisms by which such ongoing viability-preserving engagements are enabled, and make intelligible that such mechanisms can deliver the rich and varied grip upon the world that we humans enjoy. This, of course, is exactly what PP sets out to achieve"

Williams (2017: 164) espouses a similar position:

> "The core thesis of predictive processing is that brains install and deploy a generative model of environmental causes in the service of homeostasis. If we can explain how cortical networks come to embody these pragmatic structural models, and how such models can be exploited in cognitive functioning, we will have ''naturalized'' intentionality in the only way that could be important to the representational status of the framework"

But the explanation PP provides of how brains "install and deploy" generative models is a *computational* explanation, which will mention mathematical contents. It thus seems that both Clark and Williams support, at least implicitly, the idea that mathematical contents account for the intentionality (hence, the representational content) of generative models.

Notice also that this idea is *entailed* by the claim that the relevant content of a generative model is determined by the structural similarity holding between a model's topology and the causal structure of the environment. This is because the topology of the model is a *formal* property of the model that can be easily converted into a set of mathematical contents. For instance, if a graphical model displays no connection among two nodes, the corresponding variables will be conditionally independent, meaning that the value of the probability computed for the first node *X* is not affected by the value of the probability computed for the second note *Y*. This idea is so dominant in the structural representationalist reading of PP that Kiefer and Hohwy (2019: 400-401) take the relevant degree of structural similarity to be assessable

directly through graph-comparison techniques.

4.2 - The contents of generative models

What sort of contents are grounded by such a structural similarity? The question is ambiguous between two readings. It can be interpreted as a question concerning the *metaphysical status* of the contents: are they propositions, modes of presentations or something else entirely? But it can also be interpreted as a question concerning *what sort of things get represented*. I tackle both questions in order.

The question concerning the metaphysical status of contents has a crisp answer in the PP literature: contents are possible worlds (Kiefer and Hohy 2018; 2019). The idea comes directly from the connectionist tradition of generative modelling:

> "A mental state is the state of a hypothetical world in which a high-level internal representation would constitute veridical perception." (Hinton 2005: 1774)[82]

This shouldn't be surprising: after all, it seems natural to say that (metaphysically speaking) the contents of generative models are possible worlds, given that their content is rooted in the structural similarity they bear to the (causal structure of) the world.

The question concerning what gets represented by generative models has a less clear cut answer. Worlds (both possible and actual) will surely be represented. But this is too vague an answer to be informative. Mathematical objects will be represented too, since mathematical contents are the building blocks of the relevant structural similarity. But, again, this kind of answer is not really informative: presumably, questions like "what is the content of…?" inquire about representational contents, as opposed to mathematical ones.

In reality, what gets represented by generative models depends on how the relevant structural similarity the model bears with its target is spelled out. Hence, it varies as the relevant

[82] See also Ha and Schmidhuber (2018b) for a similar position.

similarity varies. Kiefer (2017; 2020), for instance, opts for a transparent code, according to which the relevant structural similarity holds between a system encoding network of material inferences and the gross causal makeup of the environment. So, in his view, representational contents seem for the most part what ordinary propositions express: facts such as "it is raining" or "the street is wet" (see above).

Others suggest instead something quite different: that the generative models might represent in a way that simply has *no translation* in terms of propositional contents. After all, the generative model is supposed to *simultaneously* capture causal regularities spread across multiple distinct spatiotemporal scales, ranging from milliseconds to years, together with the state-dependent noise expected in the input it receives. It is doubtful that such a content has a natural, non-arbitrary translation in terms of propositional contents. As Clark colorfully puts it, generative models:

> "make it even harder (perhaps impossible) adequately to capture the contents or the cognitive roles of many key inner states and processes using the terms and vocabulary of ordinary daily speech. That vocabulary is "designed" for communication, and (perhaps) for various forms of cognitive self-stimulation. The probabilistic generative model, by contrast, is designed to engage the world in rolling, uncertainty-modulated, cycles of perception and action." (Clark 2015a: 2, references omitted)

This, I believe, shouldn't come as a surprise. It is well known that the contents of "neuromorph" computational systems are opaque and hard-to-pin-down (e.g. McCloskey 1991) even in relatively simple systems.

At present, then, the question "what sort of things do generative models represent" seems only to be answerable only in reference to the specific model-target structural similarity one is committed to. Different conceptions of that structural similarity yield radically different contents. Given that such a structural similarity is partially constituted by the model's computational functioning, and that, at present, there is no "canonical" PP scheme (e.g. Ciria

et al. 2021), I think that it is wise, at present, to leave the question unanswered.

5 - Other representational posits (predictions and prediction errors)

Thus far, I have briefly sketched how the structural representationalist account of PP conceives generative models. But what about the *other* representational posits PP seems to postulate, such as predictions and prediction errors?

Consider prediction errors. They are bottom-up signals "telling" the system what "has been missing" from the original predictions. As such, they seem to be indicators of sort: an error signal *indicates* that something other than what has been predicted is the case. And, in fact, the bottom-up flow of error signals can be conceived as a "filtered" version of the standard bottom-up flow of sensory evidence, which is normally taken to *indicate* environmental features:

> "Prediction error signals are [...] not radically different to sensory information itself. This is unsurprising, since mathematically (as Karl Friston has pointed out) sensory information and prediction error are informationally identical, except that the latter are centred on the predictions. [...] The forward flow of prediction error thus constitutes a forward flow of sensory information relative to specific predictions." (Clark 2015c: 5)[83]

Now, according to the received knowledge, the feed-forward flow of information is carried out by receptors; and indeed, the *receptor* notion of representation seems to fit both top-down predictions and bottom-up errors pretty well (Hohwy 2013: Ch. 8; Orlandi 2014; Downey 2018). As seen in the previous chapter, predictions just are a "downward" flow of cortical activity aimed at re-constructing the expected input; whereas error is an "upward" flow of activity aimed at ensuring that the "downward" flow marches in step with the incoming sensory stimulation. In this way, the relevant states of the cortical hierarchy will come to *carry information* about the external causes of the sensory input; and indeed learning to predict

[83] See (Cao 2020) for sustained discussion of this point.

accurately the incoming input is guaranteed to increase the *mutual information* between internal and external states of a PP system (Hohwy 2013: Ch. 2). It thus seems possible to conceive the internal states of the brain as receptors of specific environmental contingencies.

The structural-representationalist view of PP, however, does not endorse this claim. According to the structural-representationalist view of PP, predictions and prediction errors are representations, *but they are not receptors*. And they *cannot* be receptors because, according to the structural-representationalist view of PP, receptors are not representations (e.g. Williams and Colling 2017: 1949).

There are two reasons as for why the structural-representationalist view of PP holds that receptors are not representations, and both have been inspired by Ramsey's (2003; 2007) attack on the receptor notion of representation.

The first has to do with the *content* of receptors, or better the content-grounding relation to which they are typically associated. As seen in (Ch. 2: §3.1), receptors are typically associated with teleo-informational accounts. These accounts typically considered a special case of "covariance" or "tracking" theories of content (e.g. Cummins 1991; Egan 2019); that is, theories of content that suggest that the relevant content grounding relation is a relation of *regular covariance* holding between the representational vehicle and its target. What singles out teleo-informational accounts from other tracking theories is the fact that, in the case of teleo-informational accounts, the relevant tracking relation is spelled out formally, using information-theoretic terms. Hence it need not be a *brutishly causal* relation (e.g. Fodor 1987).[84] Yet it is still a *covariance* relation in the relevant sense: it is only *because* the states

[84] I think I owe this observation to Manolo Martinez (personal communication). Sending and receiving information *need not* involve any form of mechanical energy transfer (or billiard-ball causality). Consider: A and B decide that if A calls before 5 pm, then the meeting is cancelled; otherwise, the meeting will be held as planned. At 5:01, B has received no phone calls from A. So now B is certain that the meeting will be held as planned: B's uncertainty has been reduced. But if B's uncertainty has been reduced, then the sender (in this case, A) *has conveyed a message to B* in the relevant sense. But no causal chain connects A to B in the vignette presented here. Hence, theleo-informational (or just informational) theories of content need not imply any form of causal contact between representational vehicle and target (Dretske 1981:26-39 makes essentially the same point while dealing with what he calls *ghost channels*).

of V and T reliably covary within a range of states that observing V occupying state v_a makes one more confident that T is in state t_a.

Yet covariance, even if purposive, surely is insufficient to constitute content (Hutto and Myin 2013). For one thing, covariance is a symmetric relation: if V covaries with T, then T covaries with V. But contents are not "janus faced": the activation of a V1 neuron might be about edges, but edges are not about anything. Covariance is also a transitive relation: if V covaries with T over some range of states, and T covaries with T* over some range of states, then V covaries with T* over some range of states (Cummins 2010). Perhaps the covariance relation between V and T* is *less regular* than the one between V and T, but a less regular covariance still is covariance. Lastly, covariance massively overgeneralizes: there are just too many things covarying with other things. Adding a teleological component does not significantly ameliorate the dialectical situation, for a number of non-representational "things" are supposed to covary with a number of other "things". For instance, the position of the firing pin of a gun is *supposed to* covary with the position of the trigger, but surely firing pins of guns are not representation of triggers (Ramsey 2003; 2007; Orlandi 2014).

The other reason as for why the structural representationalist reading of PP holds that receptors are not representations is that the receptor notion *over-reduces* representations. Receptors end up "explaining away" the relevant notion of representation because they end up functioning as *mere causal mediators* (Orlandi 2014; Downey 2018) relevant for the triggering of appropriate behavioral responses.

To see why, compare the following two scenarios (Ramsey 2007: 195-203). In both scenarios, a simple robot has to traverse an "S"-shaped track without physically bumping into the walls of the track. In the first scenario, the robot is equipped with an "S"-shaped groove in which a rudder fits. As the robot moves forward the rudder advances in the groove, steering the robot's wheels in a way that depends on its position on the groove. So, if the "S" shapes of

the groove and the track are structurally similar, the robot will traverse the track without touching the track's walls. In such a case, the robot's behavioral success is creditable to the internal groove; that is, the internal structure that guides the robot's behavior. Importantly, we would naturally describe such a structure as a sort of internal map that the robot "consults" in order to achieve its behavioral success. And the reason as for why we would naturally describe the "S"-shaped groove as an inner representation is because its fits the functional profile of models sketched above: the "S"-shaped groove is an internal structure that is structurally similar to the robot's behavioral target, guides the robot's actions directed to that target, can function while decoupled from the target (it would guide the robot's behavior even if no "S"-shaped track is causally affecting it) and could afford the detection of representational error (see Gładziejewski 2015b: 78-82).

Consider now the second scenario. Instead of relying on an internally "S" shaped groove, the robot's behavior depends on two rods protruding from its front side in opposite directions (left and right). If some pressure is applied to them, the rods slide inside the robot's body, moving the groove. Thus, when the *left* rod slides in, the groove is pushed to the right (and *vice-versa*). In this way, the robot can navigate the "S" shaped track without ever getting stuck into the track walls. Notice that, ins such a case, the rods effectively function as *indicators of proximity*: their state (i.e. the degree to which they are slit in the robot's body) reliably covaries with the distance between robot and wall, and they have, by explicit design, the function of indicating that distance. Yet, it seems natural to treat those rodes just as behavioral triggers: a complex causal structure that is driven by environmental contingencies and that triggers the appropriate response from the robot. As Ramsey puts it:

> "When explaining how the mindless car A [the robot, *n.a.*] makes it way though the curve, the account that seems most natural (and fulfills our explanatory goals) is one that treats the causal relay between the plunged rod and the turned wheels as just that – a causal relay that brings about an altered wheel alignment whenever the vehicle gets close to a wall. In fact,

> [...] we can explain this process as one of brute causal interaction between the wall and the wheels. [...] There are certainly more mediating links between the car's proximity to a wall and the turning of the wheels away from the wall. But – and this is the key point – there is no natural or intuitive sense in which one of the linking elements is playing the role of representing." (Ramsey 2007: 196-197).

In the case at hand, a simple causal story is sufficient to account for the robot's behavior: the fact that rode-wall contact *causes* the rode to slid in, which *causes* the robot to turn in the opposite direction is sufficient for us to understand how the robot manages to navigate the "S"-curve. There just seems to be no need to invoke representational (or otherwise semantically charged) notions. Rodes play a purely causal role.

For these reasons, the structural-representationalist view does not take receptors to be representations. Now, given that intuitively predictions and prediction errors *seem* to be receptors, it is reasonable to expect the structural-representationalist view of PP to conclude that predictions and prediction errors are not representations. If that were correct, predictions and prediction errors would then be just causal patterns of activity in the brain, whose occurrence triggers appropriate behavioral responses.

Yet this is not what the structural-representationalist interpretation of PP claims. In fact, the structural representationalist interpretation of PP takes prediction and prediction errors to be *constituents* of the overall generative model (e.g. Kiefer and Hohwy 2018: 2394-2395). This should not be surprising: recall, just to give one example, that, according to Gładziejewski (2016), the relevant generative model-target structural similarity is partially built upon *likelihoods*. But likelihoods are predictions, and prediction errors just are inverse likelihood functions (the "tell" how much the given hypothesis mispredicts - or fails to account for - the incoming data). Prediction and prediction error, thus, *are* representations according to the structura-representationalist reading of PP. They are representations because they are vehicle constituents of the overall generative model. And, as seen in (Ch. 2: §3.2) the vehicle

constituents of a structural representation are representational vehicles, which function as representations by functioning as elements in a map or model, and whose content is the relevant target constituent they are mapped onto by the structural similarity.

Importantly, the structural representationalist reading of PP contends that it is only by looking at neuronal responses (that is, predictions and prediction errors) through these lenses that the patterns of covariation between cortical activity and the environment *make functional sense*. By looking at predictions and prediction errors as constituents of a structural representation, one does not just see the brain's responses to environmental stimuli and some mysterious self-generated cortical activity. Rather, one sees a complex causal mechanism that "attunes" the behavior of an organism to the relevant environmental contingencies, by curating and maintaining a fine-grained statistical model of the environment:

> "[...] what is the *function* of such anticipatory dynamics? [...] *how* are they *achieved*? It is in answering *these* questions that the representationalist interpretation of predictive processing is required: effectively anticipating the incoming signal is necessary for the organism's ability to intervene upon the environment to maintain homeostasis, and it is made possible by the exploitation of an internal *model* of the signal *source*. Without this representationalist interpretation, the brain's ability to so successfully ''predict1'' [i.e. make its state coviary with] its incoming sensory inputs is both *unmotivated* and *unexplained*." (Williams 2017: 161).

Part II: Generative models as instantiations of sensorimotor mastery

Chapter four - Are generative models structural representations?[85]

1 - Gładziejewski's account of structural representations

The last chapter presented the structural-representationalist view of PP in broad strokes. Here, I focus on (Gładziejewski 2016), for this argument is widely supposed to have established, once and for all, that generative models are structural representations.[86]

The argument builds upon Gładziejewski's (2015b) account of structural representations, according to which

> **Structural Representation:** In a system S, a vehicle V is a structural representation of a target T if, and only if:[87]
>
> (a) V is structurally similar to T; &
>
> (b) V guides S's actions regarding T; &
>
> (c) V can satisfy (b) when decoupled from T; &
>
> (d) S can detect the representational error V generates.

Each point needs to be briefly clarified.

1.1 - Point (a): structural similarity

In (Ch. 2: §3.2) I clarified that the relevant notion of structural similarity is second-order structural resemblance, which is defined as follows:

> **Second-order structural resemblance (rewritten):** V is structurally similar to

[85] This chapter is based on (and expands upon) Facchin, M. (2021b). Are generative models structural representations?, *Minds and Machines*, *31*, 277-303.

[86] This is somehow and understatement: Gładziejewski's argument in favor of the structural representational status of generative models is the *sole* argument to that effect, which has informed *each and every* other subsequent representational analysis of PP, see (Wise 2017, 2018; Dołega 2107; Williams 2017; 2018a, 2018b, Pezzulo *et al.* 2017; Kiefer and Hohwy 2018, 2019; Sachs 2018; Vásquez 2019; Hohwy 2020a).

[87] Here, I present (a) to (d) as necessary and sufficient conditions even if, *strictly speaking*, Gładziejewski (2015b) presents (a) to (d) just as *necessary* conditions. However, it is clear that in his (2016) Gładziejewski takes them also to be *sufficient* conditions, whose satisfaction is *sufficient* to ensure the metaphysical status of a structural representation to an item. Weren't that the case, the argument presented in Gładziejewski (2016) would have little sense: weren't (a) to (d) sufficient conditions, showing that they are satisfied by generative models *would not be sufficient* to show generative models are structural representations. Notice, however, that *strictly speaking*, all I need for my argument to work is *only* that (a) and (b) are necessary conditions.

> T if and only if:
>
> (i) there's a one-to-one mapping from at least some vehicle constituents (v_xs) onto at least some target constituents (t_xs); &
>
> (ii) there is a one-to-one mapping from at least a relation *R* holding among the vehicle constituents onto at least a relation *R'* holding among the target constituents; &
>
> (iii) For all the vehicle constituents satisfying (i), $v_aRv_b \rightarrow t_aR't_b$ (i.e. the same *pattern* of relations hold in V and T)

Gładziejewski (2015b; 2016) accepts this definition.[88] I now rehearse two features of second-order structural resemblance that will be central in the following.

The first is *semantic unambiguity*. If V represents T because it bears an exploitable second-order structural resemblance to it, then, for each vehicle constituent of V (or relation among vehicle constituents of V) it is in principle possible to determine the corresponding target constituent (or relation among target constituents) of T. This is because of (i) and (ii): if a vehicle constituent v_a participates in the second-order structural resemblance, then there is one, *and only one*, target constituent t_a[89] to which it corresponds. The same goes for the relations: if a relation v_aRv_b participates in the resemblance, then there is one, and only one, relation $t_aR't_b$ to which it corresponds. Hence, it is always possible to say what each "bit" of the structural representations means in an unambiguous way.

Secondly, although a second-order structural resemblance can hold among any two systems, in the case of structural representations the relevant (content-determining) structural similarity must hold between *the representational vehicle* and its target. Structural representations are *concrete particulars* that represent by resembling their targets. This is also why the contents of structural representations are said to be *intrinsic* to their material structure. This is because they depend on the relevant (i.e. structural similarity-constituting) physical properties of the vehicle

[88] He typically refers to the definition of second-order structural resemblance given in (O'Brien and Opie 2004: 11) to which my modified definition is identical (apart from the notation).
[89] Recall that, to simplify the notation, I assume that the mapping in (i) is "subscript preserving".

(O'Brien and Opie 2001; O'Brien 2015; Kiefer and Hohwy 2018; Williams and Colling 2017).[90]

1.2 - Point (b): action guidance, or exploitability

Point (b) requires V to guide S's behavior directed at T. Thus, it requires the relevant structural similarity V bears to T to be *exploitable* by S (Gładziejewski and Miłkowski 2017). As far as I know, there is only one canonical definition of exploitability, which is provided by Shea (2014; 2018: 120). Shea (2018: 120) defines exploitability as the conjunction of two requirements:

> **Exploitable structural similarity:** The structural similarity holding between V and T is exploitable by S if, and only if:
>
> (iv) The relevant relation or relations *R* are relations S's downstream computational processing is systematically sensitive to; &
>
> (v) The target constituents t_x and the relation or relations *R'* defined over them are of significance to S[91]

Condition (iv) imposes that the structural similarity-constituting relation (or relations) holding among the constituents of V are, in a sense, *computed upon* by S.[92] Changes in these relations must thus affect the behavior of S in some systematic way. Notice that condition (v) further reinforces the idea that the relevant structural similarity must hold between a *vehicle* and a target. In fact, computation in cognitive systems is a causal affair. Hence, it seems that for a relation to systematically orient a system's computational processing, either the relation itself, or at least the relevant constituents among which it holds, must possess some relevant

---

[90] Johnny Lee (2018) goes so far that he *identifies* contents with the similarity-constituting properties of the vehicle.

[91] I adapted the notation used in Shea's definition for the sake of orthographic consistency.

[92] There is a potential problem here, at least insofar Shea's definition of exploitability does not mention consumers. This is because Gładziejewski's (2015b) account does include consumers, and it is intuitive to think that the relevant structural similarity must be exploitable for the vehicle's consumer. This would make the definition of exploitability a bit more restrictive. However, Gładziejewski's (2016) application of his account of structural representations to PP does not mention consumers, and the alternative definition of exploitability sketched in (Gładziejewski and Miłkowski 2017) does not mention them either. It thus seems fair to suppose Gładziejewski has (more or less implicitly) eliminated consumers from his account. Note, however, that albeit "naked" structural similarity is a two place relation, exploitable structural similarity is a three place relation. In fact, a vehicle (1st relatum) bears an exploitable structural similarity to a target (2nd relatum) for some system (3rd relatum).

causal power. And, in cognitive systems, causal powers pertain to *vehicles* (e.g. Egan 2012; 2020).

Condition (v) mentions *significance*. Significance is here relative to the way the system produces the output it is supposed to produce. Hence, a complete unpacking of (v) would require me to introduce Shea's complex account of functions (Shea 2018: Ch. 3). But, given that condition (v) will play no role in the overall argumentative structure of this chapter, I postpone the exhibition of that account to (Ch. 6: § 2.2).

In Ch.2 (§ 3.2) I mentioned that exploitability solves the problems of reflexivity and symmetry similarity-based accounts of content suffer from (Goodman 1969). It also (partially) solves the problem of content determination associated with these accounts. The reasons supporting these claims should now be clear enough. Conditions (iv) and (v) "chunk down" the number of targets a vehicle can be *exploitably* structurally similar to, thereby contributing in determining a vehicle's content (see Nirshberg and Shapiro 2020). Condition (v) solves the problem of symmetry, reflexivity and transitivity. V surely is structurally similar to itself, and it is surely correct to infer that, if Vs structurally similar to T, then T is structurally similar to V. However, V is not *of significance* to S. Hence, V is not *exploitably* structurally similar to itself, and T bears no *exploitable* structural similarity to V.

1.3 - Point (c): Decouplability

According to (c), V is required to be able to satisfy (b) even when decoupled from T. Hence, the relevant notion of decouplability needs to be clarified. Decouplability is notoriously hard to define (Chemero 2009; Orlandi 2014: 122-134; 2020). Luckily, Gładziejewski (2015b: 76-77) provides his own crisp definition. In his view, decouplability comes in two degrees: weak and strong.

> **Decouplability:** A vehicle V is decouplable from T if, and only if, V is weakly decouplable or strongly decouplable from T

> **Weak decouplability:** V is *weakly decouplable* from T if, and only if, V can perform its action-guiding duties even when:
>
> (vi) No causal connection holds between V and T; &
>
> (vii) No causal connection holds between T and V's consumer within S
>
> **Strong decouplability**: V is *strongly decouplable* from T if, and only if, V can perform its action guiding duties when:
>
> (viii) No causal connection holds between T and S

Notice here decouplability is spelled out in terms of *causal connections* (or lack thereof) between V and T. This reinforces the idea that the relevant structural similarity is required to hold between a vehicle and a target; as, among all the components of a representation, only the *vehicle* and the target can be in any causal contact.

There are few other things to notice. First, weak decouplability is a form of decouplability, and so it is sufficient to satisfy (c). Secondly, (c) is satisfied when V and T are decouplable, that is, when they *can* be decoupled. Hence V and T need not be always decoupled for (c) to be satisfied.

As a technical aside, notice that (vi) mentions *consumers*. As suggested above (fn 96), Gładziejewski's application of his account of structural representations to PP does not mention consumers, and it really seems Gładziejewski (2016) has implicitly eliminated consumers from his account of structural representations. Hence, condition (vii) will be largely ignored in the following.

1.4 - Point (d): Error detection.

Everyone agrees on the fact that if V represents T, then V can also *misrepresent* T. Point (d) can be seen as an extreme version of that simple truth (Bickhard 1999; Miłkowski 2013). In this view, a good theory of representations need not just account for the possibility of misrepresentation; it must also provide a way for representational systems to detect their misrepresentations.

In public representations error can be readily detected by comparing the vehicle to its target. If I have a picture of *x*, I can check whether the picture correctly depicts *x* by comparing the two. But a system can access a representational target only *through* its internal representations of it. Hence it has no *independent* access to its representational targets, and it cannot compare them to its inner representations of them. So how can the error of cognitive representations be detected?

Gładziejewski (2015b: 80) suggests such a detection is possible if the relevant system possesses a sub-component detecting *pragmatic* failures. It is standardly assumed that correct representations non-accidentally lead to successful actions, and that unsuccessful actions are non-accidentally due to incorrect representations. Thus, *pragmatic* successes and failures can function as reliable indicators of the accuracy of a system's representations.

More in detail, Gładziejewski suggests there are two ways in which pragmatic successes and failures can indicate the semantic status of representations. The first, and most obvious, way in which representational error can be detected is through a feedback signal, which indicates whether pragmatic success has been attained. The second, less obvious, strategy is to use a "predict and compare" strategy (Gładziejewski 2015b: 81). The idea is as follows: the representation generates a "mock" signal of how actions should unfold were its content correct. The "mock" signal is then contrasted with the signal delivered by actually unfolding actions. The mismatch between "mock" and real signal can then be used to assess by a system to monitor the semantic status of its own inner representations.

Notice that, in both cases, an *additional monitoring mechanism* is needed, either in the form of an appropriate mechanism "reading" the feedback signal or in the form of a *comparator*, computing the mismatch between expected and actually achieved outcomes.

This concludes the presentation of points (a) to (d).

1.5 - The scope of Gładziejewski's account

Before moving forward, it is important to clarify the scope of Gładziejewski's account of structural representations. Following Chemero (2009: 67-68), it is possible to distinguish between an *epistemic* representationalist claim and a *metaphysical* representationalist claim. Bluntly put, the epistemic representationalist claim is the claim that our *best explanations* of cognition need to posit representations. The metaphysical representationalist claim is instead the claim that cognitive systems contain components that *really* are representations. The two claims can in principle come apart. A fictionalist about representations, for instance, endorses the epistemic claim and denies the metaphysical one (Sprevak 2013).[93] Gładziejewski's account of structural representations aims at vindicating both claims (Gładziejewski 2015b: 70).[94] Thus, his account of structural representations succeeds just in case the relevant representational posits of PP (i.e. generative models) satisfy features (a) to (d) *and* these are the relevant sort of structures identified as representation by our best explanatory practices.

2 - Generative models as structural representations: Gładziejewski's argument

Gładziejewski (2016) argues his account of the functional profile of structural representations is tailored to fit generative models. His argument is as follows.

(P1) Items satisfying conditions (a) to (d) in conjunction are structural representations.

(P2) Generative models satisfy conditions (a) to (d) in conjunction

(C) Generative models are structural representations

The argument has a straightforward structure, and its force hinges almost entirely on (P2). I now examine each step of Gładziejewski's argument for (P2).

[93] See also (Downey 2018) for a fictionalist interpretation of PP.

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

2.1 - Point (a): Gładziejewski's argument for structural similarity

Gładziejewski's (2016: 571-576) argument is as follows. Generative models can be formally treated as graphs (see Ch. 1, § 2.1), and in particular as Bayesian nets or Directed Acyclic Graphs (DAGs). DAGs are made up of a finite number of nodes connected by edges. In a DAG, each (labelled) node corresponds to one, and only one, environmental variable. Edges can be formally treated as relations (see Leitgeb 2020), and they can be mapped onto some relevant relation in the target domain. In the case of DAGs, the relevant relations in the target domain are typically causal relations. In a graph, two nodes are connected if, and only if, the corresponding relation holds among the corresponding variable in the target domain (Danks 2014: 39-41).[95] Hence, since the graph's topology mirrors the causal relations among the elements in the target domain, graphs bear at least a second-order structural resemblance (really, and homomorphism) to their targets (Ch. 3: § 4.1).

Gładziejewski (2016) also proposes a *specific* structure preserving mapping holding between the generative model and its representational target. As seen in the previous chapter, he claims that each of the graph encodes the *likelihood* of encountering a given pattern of sensory stimulation, that relations between nodes mimic the dynamical relations among different environmental causes and that each node also encodes some expectation about the prior probability of encountering a given pattern of sensory stimulation.

This specific structure preserving mapping, however, is unconvincing. As Wiese (2017) rightfully notices, not every PP model encodes likelihoods (cfr. Buckley *et al*. 2017). Moreover, it is hard to understand *how* prior probabilities can partake in a structure preserving mapping with a represented target, given that, in Bayesian statistics, prior probabilities are

[95] A bit more formally: let each labeled node in a graph V be a vehicle constituent v_x. and let each variable describing a target T be a target constituent t_x. Let the labeling be "subscript preserving" (i.e. node v_a maps one-to-one onto t_a). Let *R* denote the relation "being connected to" holding among the nodes, and let *R'* denote the causal relations holding among the target variable. Now, in a graph, $v_x R v_y \Leftrightarrow t_x R' t_y$; which is clearly enough for a second-order structural resemblance to hold between V and T.

subjective degrees of belief (Feldman 2016; Bolstad and Curran 2017). Yet, the fact that Gładziejewski's proposed mapping is unconvincing should not distract us from the more general, more widespread and more defensible claim; namely, that a structure preserving mapping holds between the topology of the graph and its represented target. And, as said above, this claim seems sufficient to vindicate point (a). In the following I will discuss only this claim, ignoring the (questionable) structure-preserving mapping Gładziejewski proposes.

2.2 - Point (b): Gładziejewski's argument for action guidance/exploitability

The argument Gładziejewski (2016: 575-576) provides to claim generative models satisfy (b) is basically the following: generative models can engage in active inference. Hence, they can guide the actions of the system.

More in detail, Gładziejewski's reasoning seems to be the following. If generative models really are graphs, the success of perceptual inference non-accidentally depends on the structural similarity holding between the graph and the target. The more graph and target are structurally similar, the more the model is able to infer the correct cause of the incoming sensory barrage. As a cause (or hypothesis) is selected, it can then be tested in active inference. But, again, a model's success in bringing about the selected sensory states through active inference non-accidentally depends on the structural similarity holding between model and target. Only if a tight structural similarity holds between the two a given course of action will deliver the predicted sensory inputs.

To simplify, consider the following toy example. I've inferred that the current cause of my visual inputs is a glass of vodka. Now, my generative model commends some vodka-related expectations. However, the model incorrectly predicts some vodka-related sensory states (for instance, it predicts that the ingestion of vodka will cause the same states that would be caused by ingesting water). As these states are actively inferred (i.e. as I drink vodka) my prediction

error will *not* be minimized, as the expected states will not be encountered (drinking vodka and drinking water do not bring about the same sensory states). In such a case, the failure in error minimization is due to a mismatch between the generative model and the relevant target; that is, the fact that the "vodka" node is connected to the wrong set of sensory observations. Were the relevant connections different, active inference would have been successful.

As the simple example above shows, PP systems are sensitive, in their functioning, to the relevant relations holding among the nodes. This seems a vindication of (iv). But what about (v)? To claim that generative models satisfy (v), Gładziejewski points to the fact that prediction error minimization is a tool for homeostasis (e.g. Seth 2015). Hence, we should expect generative models not just to be organism-relative, but also busy modeling what matters for an organism's survival - which, *prima facie*, is significant for the system.

2.3 - Point (c): Gładziejewski's argument for decouplability

To claim generative models satisfy (c), Gładziejewski (2016: 576-577) puts forth a number of different considerations.

First, he argues that generative models can be decoupled because they provide an *endogenous* and *future oriented*[96] source of control. Generative models are *endogenous* sources of control as active inference need not be a reaction to external stimulation: it might also be the way is which an agent controls, from the top-down, its sensory states (e.g. Linson, *et al.* 2018; Ramstead, Kirchhoff and Friston 2019). Moreover, the sensory states that active inference tries to bring about are not yet present: they have to be brought about through movement. But surely generative models cannot be in causal contact with something not present. Hence, at least weak decouplability follows.

[96] It is important to notice, however, that the future-oriented nature of predictions in PP systems should not be overstated. In fact many predictions are predictions of current sensory states (Bubic, von Cramon and Schubotz 2010). The relevant sense of prediction at play in PP is the statistical sense of prediction, which is not synonymous with "foreseeing".

Moreover, Gładziejewski notices that *early* sensory and motor cortices are active not just when a subject is perceiving or moving, but also when the subject is *imagining* to perceive and/or to move (e.g. Miller *et al.* 2010; Albers *et al.* 2013). But we surely *are* strongly decoupled from imaginary targets. So, generative models can function even when strongly decoupled from their targets.

Generative models can also represent *counterfactually*. For instance, they might represent what sort of sensory states would be encountered were a given policy pursued (e.g. Seth 2014; Friston *et al*. 2012b; FitzGerald, Dolan and Friston 2014). And, again, *counterfactual scenarios* are not-yet-actual scenarios, with which a generative model cannot be in any form of causal contact.

Lastly, Gładziejewski notes that PP has been used to account for multiple aspects of REM dreaming (e.g. Hobson and Friston 2012). And we surely are strongly decoupled from *dreamed* objects.

Thus, Gładziejewski concludes that we have a number of strong empirical reasons to accept that generative models can function when decoupled from their targets.

2.4 - Point (d): Gładziejewski's argument for error detection

Gładziejewski's argument (2016: 577-579) for (d) is straightforward: "predict-and-compare" strategies to detect representational error capture exactly how prediction error is computed (mainly, by subtracting expected and received sensory inputs). In this way, prediction error can function as a reliable indicator of the generative model's semantic standing (see Ch. 3: 2.2).

This concludes the presentation of Gładziejewski's argument. I will attack it in the next paragraph.

3 - A critique of Gładziejewski's argument

I claim that the argument above fails to establish the status of generative models as structural representations. This is because of two major flaws in Gładziejewski's argument for (a). The two flaws are the following: first, the argument does not establish that generative models satisfy condition (a). Secondly, were the argument sufficient to show generative models satisfy (a), then (b) would not obtain. I discuss these two flaws in turn.

3.1 - Gładziejewski's argument for (a) fails

Gładziejewski's argument for (a) is based on graph-theoretic consideration (§ 2.1). If I understood it correctly, it boils down to the following:

(P1) Generative models are graphs

(P2) Graphs are structurally similar to their targets

(C) Generative models are structurally similar to their targets

But this line of reasoning does not vindicate the claim that (a) obtains. If one is a realist about representations (as Gładziejewski is, see Gładziejewski 2015b; 2016; Gładziejewski and Miłkowski 2017), then one is committed to the claim that representations are *concrete particulars* encoding content (e.g. Shea 2018: 25-43). Hence, structural representations are *concrete particulars* (i.e. representational vehicles) carrying content in virtue of the relevant exploitable structural similarity holding between them and their representational targets. They are *representational vehicles* that do the representing by structurally resembling. The relevant structural similarity must thus hold between a representational vehicle (a concrete particular) and a represented target. This is why the content of structural representations is supposed to be *intrinsic* to their material constitution (e.g. O'Brien and Opie 2001; Williams and Colling 2017; Lee 2018). Their content is intrinsic as it is inscribed in the physical form of the representational vehicle (the concrete particular that does the representing).

But Bayesian nets, and graphs in general, are *mathematical objects* (e.g. Danks 2014: 40; Leitgeb 2020). They are defined as a finite *set* of *nodes* connected by a finite *set* of *edges* (Koski and Noble 2009: 41) and sets, nodes and edges are mathematical objects. Mathematical objects might or might not be particulars (it is irrelevant for the purpose of the argument), but definitely are not *concrete*. So they cannot be representational vehicles, given that representational vehicles are *concrete* particulars. Hence, the structural similarity Gładziejewski points to cannot be used to vindicate point (a). It just isn't what point (a) requires.[97] In even simpler terms: *there literally are no graphs in the brain*.

3.2 - If Gładziejewski's argument for (a) were successful, then (b) would not obtain

Further, were Gładziejewski's argument for (a) to succeed, then (b) would fail to obtain. Recall that (b) requires the relevant structural similarity to be exploitable. *Which* structural similarity? Obviously, the one satisfying (a). But it is very doubtful that the relevant structural similarity Gładziejewski leverages to satisfy (a) could be exploitable, because of condition (iv) on exploitability.

Condition (iv) requires the system relying on the putative vehicle to be sensitive, in its downstream computational operation, to the relations holding among the vehicle constituents. But, as noticed above, computational operations are, in the relevant sense discussed here, a physical affair. Hence, either the relations themselves or the vehicle constituents upon which they are defined must have some appropriate causal power, so as to systematically influence a system's downstream computational operations.

However, neither nodes nor edges have causal powers, as they are mathematical objects; and mathematical objects typically lack causal powers. Hence, the structural similarity graphs

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

bear with their targets is not exploitable, as it fails condition (iv) on exploitability. Therefore, if Gładziejewski's argument for (a) were successful, then (b) would not obtain.

3.3 - A diagnosis

What went wrong in Gładziejewski's argument for (a)? The problem seems to be the following: structural representations are defined in terms of their vehicle properties. Hence they should be identified at the level of the *physical machinery* doing the computation (what Marr would call the implementation level). Structural representations are bits of information processing systems *literally* resembling their target. But graphs, wherever they sit in the explanatory hierarchy of cognitive science, surely do not sit at the level of the *physical machinery* doing the computing (Danks 2014, pp. 13-37; 218-221). Gładziejewski's argument seems to be pitched at the wrong level of the relevant explanatory hierarchy.

I believe that the problem lies in (P1). It is ambiguous between generative models as *mathematical objects* (joint probability distributions represented by graphical models) and the *physical machinery* implementing them; that is, the representational vehicle. If we consider generative models mathematical objects, then (P1) is true. But then the argument, albeit valid and sound, does simply not concern representational vehicles, leaving the claim that the vehicles instantiating generative models are structurally similar to their targets unsupported. Hence, under this reading, the argument would simply leave (a) unsubstantiated. To substantiate (a), the term "generative model" in (P1) should be read as "the physical implementation of a generative model". Whilst under this reading the argument is surely valid, it ceases to be *sound*. For, under such a reading, (P1) would simply be false: no physical implementation *literally is* a graph.[98] Hence (a) would again be left unsubstantiated. And,

[98] Let me add, for the sake of clarity, that physical objects (such as appropriately traced inkmarks) can *represent* graphs. But surely representing is not being: a picture of me represents me without *being* me.

clearly, we cannot read "generative model" in (C) as referring to physical implementations *while* reading "generative model" in (P1) as referring to mathematical objects, for that would be a *quaternio terminorum*. Hence the argument would not be valid.

4 - Alternative arguments for (a)

Gładziejewski's argument is not the only possible argument to claim that (a) obtains. Here, I consider some alternative arguments to the same effect, which might be used to establish the claim that generative models are structural representations.

4.1 - Alternative argument #1: Graphs, physical machinery, and transitivity

One could try to apply graph theoretic notions at the implementation level to vindicate (a): after all, system-level neuroscience *routinely* applies graph theoretic notions to the study of biological brains (cf Sporns 2010). So, graphs *can* be used to describe the functional and structural properties of the neural machinery, which plausibly hosts the relevant vehicles of generative models. This suggests an alternative way to vindicate (a) by transitivity.

The basic idea is this. Consider a graph detailing a generative process (i.e. the process generating the sensory data the generative model is trying to account for). That graph is, *qua graph*, structurally similar to its target. But that graph (or, an approximation of it) should also be somehow encoded in the brain by some well-defined set of neural regions, which are candidate vehicles of the relevant generative model. Hence, if that graph can be mapped in a structure-preserving way onto a candidate vehicle, and the graph is structurally similar to the generative process these neuronal region purportedly represent, then the candidate vehicle is structurally similar to its target, since structural similarity is a *transitive* relation.[99] Importantly,

[99] Notice that the argument tries to vindicate point (a), which *only* requires a structural similarity. And structural similarities *are* transitive; only *exploitable* structural similarities are not transitive. But exploitability is *not* required to vindicate (a) - it is only required to vindicate (b). So an argument for (a) can *rightfully* leverage the fact that structural similarities are transitive. An appeal to transitivity does not undermine the argument.

some (Kiefer 2017; Kiefer and Hohwy 2018; 2019[100]) defend the claim that generative models are structural representations in a somewhat similar way.

This argument is fairly attractive, and it nicely integrates with the existing *scientific* literature on PP, at least insofar some generative models, rendered as Bayesian nets, have been mapped onto cortical structures (e.g. Bastos *et al.* 2012; Friston, Parr and de Vries 2017). Isn't this *sufficient* to show that at least these generative models are structurally similar to their targets?

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

But what, then, is the purpose of the graphs in (Bastos *et al.* 2012; Friston, Parr and de Vries 2017)? The answer, if I understand the literature correctly, is the following: these graphs are, in a sense, *didactic tools*, aimed at showing, with a fair degree of approximation, that the cortical machinery is arranged in a way such that it can easily perform the inferential processes PP revolves around (see Bastos *et al*. 2012, p. 703; Friston, Parr and de Vries 2017, p. 393). In fact, it seems to me that, within the PP literature, graphical models are often deployed to capture the *message passing* within the brain; that is, how inference is performed (see, for instance, de Vries and Friston 2017; Friston *et al*. 2017b, c; Donnarumma *et al.* 2017; Matsumoto and Tani 2020).[101] I believe that this is an important point to notice for two distinct reasons.

[100] To anticipate something that will emerge later on in the discussion, I actually believe that Kiefer and Hohwy's essays *do not* defend (or end up not defending) the claim that generative models are structural representations; they *only* defend the claim that generative models are structurally similar to their targets.

[101] Notice that the scope of my claim is restricted to PP and the usage of graphical models in the PP literature. I

First, if these graphical models are intended to be models of the relevant *message passing*, it seems more natural to suppose they will map onto the cortical machinery *performing* the inferences (i.e. the entire computational system), rather than on the representational vehicles manipulated in inferential processes (i.e. the single computational/representational vehicles).[102] Secondly, and relatedly, if those graphical models are accurately characterized as portraying the *inferential* message passing in the brain, it seems to me that they presuppose the presence of some relevant representational vehicle, as inferences are *defined over* representations. These representations might (but, as far as I can see, need not) be structural representations. However, as these graphical models seem to *presuppose* the presence of representations, it seems to me that they cannot be invoked to *justify* one's representationalist claim, on pain of circularity.

Importantly, I do not believe that the considerations offered above rule out in any way the concrete possibility of using graphical models to justify (a). As far as I can see, one might still resort to a graphical model to argue that at least some representational vehicles in the brain are structurally similar to their targets using the argument by transitivity sketched above. However, to do so, one would need a graphical model depicting some specific worldly target. And, to the best of my knowledge of the PP literature, no such graphical model has yet been proposed.

4.2 - Alternative argument #2: Artificial Neural Networks, weights and structural similarity I

Artificial neural networks might provide a different way to leverage graph theoretic notions to vindicate (a). As formal objects, artificial neural networks *are* graphs. But they are also somewhat plausible sketches of the *physical machinery* implementing or realizing some given computational process of interest (see Haykin 2009: 1-18; Rogers and McClelland 2014).

make no claim on how graphical models are used in the rest of cognitive neuroscience (and related disciplines).

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

Moreover, at least some artificial neural networks encoding generative models (such as Helmholtz machines) are Bayesian graphs (Dayan and Hinton 1996). Therefore, even if these artificial neural networks cannot *prove* that generative models in the brain are structural representations, they can show that generative models can be structural representations, and thus provide circumstantial evidence in favor of (a) obtaining. If our plausible sketches of the physical machinery encoding generative models are graphs (or at least graph-like), then we have a solid reason to believe the *real* physical machinery encoding generative models is graph-like. And given that graphs are structurally similar to their targets, we have a solid reason to believe in the obtaining of (a).

However, I think such a belief would be misplaced. Indeed, it seems to me that a closer consideration of artificial neural networks provides a reason to believe that (a) *does not* obtain.

To see why, consider first that artificial neural networks are often said to encode generative models in their weighted connections (e.g. Dayan and Hinton 1996; Hinton 2014; Spratling 2016: 3).[103] But weighted connections (or, more precisely, weight matrices) are typically considered to be *superposed* representations (Ch. 2, §3.3). And superposed representations are *not* structurally similar to their targets. As a consequence, if considering artificial neural networks provides evidence regarding the status of (a), the evidence they provide is *not* in favor of (a) obtaining.

Recall the notion of a *superposed* representation. A representation R is said to be a superposed representation of two targets T and T* when R encodes information about T and T* using the *same* set of physical resources. When applied to weight matrices, the idea is that weight matrices superpositionally represent their targets when each individual weight is assigned a value such that the network can exhibit the functionality needed to operate on all its targets (Clark 1993: 17-19, see Van Gelder 1991, 1992 for further discussion). For instance, a

[103] The claim, however, is not *entirely* correct, as I will soon clarify in the main text (§4.4)

single net can be first trained to recognize (or generate) instances of T. If the network is then trained so as to recognize (or generate) both instances of T and T*, then the weights of the net will encode information about *both* representational targets, and the weight matrix will be a superposed representation of both.

However, in weight matrices: "Each memory trace is distributed over many different connections, and each connection participates in many different memory traces" (McClelland and Rumelhart 1986: 176). So, it seems each individual weight maps onto *many* different representational targets (or aspects thereof). But if this is the case, then either condition (i) or (ii)[104] on structural similarity is blatantly violated, since they require a one-to-one mapping. As a further proof of their violation, recall that the obtaining of (i) to (iii) in conjunction *entails* semantic unambiguity. That is, if (i) to (iii) jointly obtain, it is always *in principle possible* to tell which bit of the represented target each vehicle constituent corresponds to. However, in superposed representations: "It is impossible to point to a particular place where the memory of a particular item is stored" (Rumelhart and McClelland 1986: 70). Superposed representations are thus not semantically unambiguous. Therefore, *at least one* condition among (i) and (iii) is not met. As a consequence, they do not support the claim that (a) obtains.

4.3 - Alternative argument #3: Artificial Neural Networks, weights and structural similarity II

The strength of the previous argument hinges on the fact that no discernible structural similarity seems to hold among *individual weights* and represented targets. But what if individual weights are not the right unit of analysis? A discernible weight matrix-target structural similarity might emerge at a different level of analysis. If I understand it correctly, a

[104] Or both. The formulation in terms of "either (i) or (ii)" is due to the fact that it seems to me that one might interpret weighted connections both as vehicle constituents or as relations holding among vehicle constituents (that is, nodes).

(fairly obscure) paper by O'Brien and Opie (2006) aims to show precisely that.

Their argument hinges on the notion of *fans-in*, defined as *the set of connections* providing inputs to a node.[105] According to O'Brien and Opie, the weight matrix of a connectionist system and the system's target domain are structurally similar roughly in the following way: each fan-in v_x maps onto an element t_x of the target domain, and there are two relations *R* (defined over fans-in) and *R'* (defined over the elements of the target domain) such that if $t_x R' t_y$ then $v_x R v_y$.[106]

To justify this claim, O'Brien and Opie (2006) trained a number of simple (three-layer, fully connected, feedforward) artificial neural networks to classify color hues. They then compared color hues and fans-in, showing that the mean spectrum of each color (rendered as a diagram showing wavelength on the *x* axis and amplitude on the *y* axis) is similar to one, and only one, fan-in in the network (rendered as a diagram showing input index on the *x* axis and weight value on the *y* axis). On this grounds, they concluded that, in trained connectionist systems, weighted connections, analyzed in terms of fans-in, bear a structural similarity to the target domain a system has been trained to operate upon.

Yet their conclusion does not seem to follow from their data. Their data only show that each fan-in v_x of the network maps on an element t_x of the domain the net has been trained to operate upon. But this does neither entail nor show that there is a pattern of *relations* among the elements of the target domain which is systematically mirrored by the relations holding among fans-in. Indeed, as far as I can see, O'Brien and Opie (2006) mention *no* such pattern of relations. Hence, the claim that weight matrices, analyzed in terms of fans-in, are structurally similar to the network's task domain is *not* justified by the empirical evidence O'Brien and Opie show.

Moreover, even if O'Brien and Opie (2006) had found such a pattern of relations, it would

[105] Notice that, thusly defined, fans-in have nothing to do with ordinarily understood fans-in (i.e. the number of input that a logic gate can handle).

[106] Notice that, in order for O'Brien and Opie's claim to work, fans-in *need* to be interpreted as vehicle constituents.

still be doubtful whether their claim generalizes to *all* weight matrices of *all* networks. Surely, we cannot generalize *by induction*: all O'Brien and Opie "provided" is just *one*, fairly specific, case. Some argument seems needed to claim that such a sweeping generalization holds - or, better, that it *would* hold, were fans-in structurally similar to the network's task domain. Yet, no such argument is provided.

In fact, it seems to me that there are compelling reasons to believe that such a claim would *hardly* generalize. To see why, consider the fact that not *every* layer in an artificial neural network does the same kind of computational job. There are convolutional layers, whose job is to identify some feature in the input pattern *irrespectively* of the feature's position in the incoming input (Foster 2019: 46-51). There are simple recurrent layers, whose job is that of providing a "short-term memory" to the system (Elman 1991). There are dropout layers, whose job is that of avoiding overfitting (Forster 2019: 54-56). As far as I can see, there is just no reason to believe that the fans-in of these layers will correspond to *anything* in the target domain the net operates upon.

For these reasons, it seems correct to conclude that O'Brien and Opie's (2006) argument provides no evidence in favor of (a).

4.4 - Alternative argument #4: Artificial Neural Networks and weightless structural similarity

Perhaps the two preceding arguments have over-emphasized the importance of connections in artificial neural networks. Generative models are not encoded in connections alone; they are *jointly* encoded by connections *and activity vectors* (e.g. Buckley *et al* 2017: 57). Moreover, the definition of structural similarity relevant to the obtaining of (a) quantifies only over *some*. Thus, noticing that connections do not participate in any one-to-one mapping does not, in and by itself, provide a compelling argument to the effect that (a) does not obtain: connections

might simply be excluded from the vehicle constituents (or relations) participating in the structural similarity. Perhaps the relevant vehicles encoding generative models are activity patterns, and the relevant structural similarity should be sought between activity patterns and a target domain. Alternatively, the generative model might be encoded by *both* connections and activity patterns, but only activity patterns bear a structural similarity to the represented target. After all, the definition of structural similarity quantifies only over "some".

For the sake of clarity, it is now important to stress that generative models are *jointly* encoded by activity vectors and weighted connections as the alternative argument suggests (Buckley *et al.* 2017). So the two arguments provided above *did* overemphasize the importance of weighted connections in the encoding of generative models. To my excuse, I'd like to stress that the PP literature encourages this excess of emphasis on emphasis on weighted connections:

> "We allowed the network to learn a hierarchical internal model of its natural image inputs by maximizing the posterior probability of generating the observed data. *The internal model is encoded in a distributed manner within the synapses* of the model at each level". (Rao and Ballard 1999: 80, emphasis added)

> "The representation at any given level attempts to predict the representation at the level below; at the lowest level this amounts to a prediction of the raw sensory input. *It is the backward connections, therefore, that instantiate the generative model*." (Shipp 2016: 3, emphasis added)

> "The generative model, which in theories such as hierarchical predictive coding is hypothesized to be *implemented in top-down cortical connections*, specifies the *Umwelt* of the organism the kinds of things and situations it believes in independently of the current sensory data [...]" (Kiefer 2020: 2, emphasis added)

This might be enough of an *excuse*, but excuses are not arguments. So, does factoring in activity patterns (alongside weighted connections) afford a way to vindicate (a)? I don't think so.

First, simply factoring activity patterns in (presumably, as vehicle constituents) does *not* change the fact that weighted connections do *not* map one-to-one onto target constituents or relations holding among target constituents (§4.2). So, considering activity patterns in the

relevant structural similarity (as vehicle constituents) is not enough. It is also necessary to *exclude* weighted connections: it must be claimed they do not participate in the relevant structural similarity.

Can some relevant network-target structural similarity be defined without taking into account weighted connections? The answer is surely positive. There is nothing particularly new in this claim: Paul Churchland's state-space semantic is the most obvious example of a network-target structural similarity that does not involve connections (see Churchland 1989; 2012). In his view, the entire *activation space* of the hidden layers of a network structurally resembles the target domain upon which the network has been trained to operate. And I'm ready to concede that a similar structural similarity can be found by considering artificial neural networks encoding generative models.[107]

Isn't this *just conceding* that (a) obtains? No, it is not. For activation spaces (the first relevant *relatum* of the structural similarity) are not vehicles, because they are not concrete particulars. They are *abstract mathematical spaces* that are used to account for the systematic behavior of artificial neural networks. So, they fail to vindicate (a) for the same reasons Gładziejewski's argument fails to vindicate (a). The same reasoning applies also to other accounts of network-target structural similarity that do *not* factor in connections (e.g. Grush 2008; Garzón and Rodríguez 2009). All these accounts show that *the abstract mathematical space* that describes the activity of the network is structurally similar to the target. But these *abstract* mathematical spaces are not *concrete* particulars, so they are not vehicles. The vehicles are the individual patterns of activity tokened within the network. And the argument here considered does neither entail nor show that these *individual patterns* are structurally similar to their representational targets.[108] Indeed, there are good reasons to hold that individual patterns of activation *are not*

[107] I'm doing so for the sake of discussion. I'm actually now persuaded that finding such a structural similarity in generative models will be hard, and that, even if present, it would not substantiate a structural representationalist reading of generative models. See (Ch. 6: § 5.2) for discussion of this point.

[108] To be clear: I'm not denying that it is possible to use the structural similarity holding between the activation

(in general) structurally similar to their targets. This is because patterns of activation *are superposed representations too*.[109] Even leaving aside the activity vectors resulting from the combination of other activity vectors (e.g. Smolensky 1990), all individual activity vectors represent, to some degree, multiple possible targets (e.g. McClelland and Rumelhart 1986, vol 2: 390-341).[110] So, if this is correct, and individual vectors are superposed representations, they will not be structurally similar to their targets, for reasons that, *mutatis mutandis*, are identical to the one provided in (§4.1).

Moreover, I honestly doubt that it is possible to *rightfully* exclude weighted connections from the relevant structural similarity. To see why, consider the following: if a vehicle represents in virtue of the structural similarity it bears to a target, then the more the vehicle and the target are structurally similar, the more the representation will be accurate. The accuracy of a structural representation non-accidentally increases when (and, at least *prima facie*, only when) the elements of the vehicle that participate in the relevant structural similarity are rearranged in a way such that their newfound arrangement increases the extent to which the vehicle is structurally similar to the target (see Gładziejewski and Miłkowski 2017).

If this is correct, then there seems to be a solid reason to deny that we can exclude connections from the relevant network-target structural similarity, for modifications of weighted connections made in accordance to the relevant learning algorithm *do improve* the representational accuracy of connectionist systems. Thus, if these systems represent by means

---

space of a network and its target domain to ground the content of each individual vector of activation. But this would *not* entail that each vector is a structural representation of its target. To see why, notice that in such a case the entire"vehicle" V is the entire activation space, of which individual vectors are "vehicle constituents". But, in general, the vehicle constituents of a structural representation need not be structurally similar to the target constituents they map onto; hence they need not be structural representations (although they *may*). Hence, in order to substantiate the claim that activation vectors are structural representations, one would need to show that each individual vector is structurally similar to the target it represents. And that is not what the argument shows, nor something that can be shown by looking very carefully at activation spaces.

[109] Thanks to Erik Myin for having reminded me of this.

[110] To be precise: all activity vectors are superposed, *unless* the network uses a localist (1 node = 1 feature) coding scheme. But such networks were ancient even in the '80s (cf. Clark 1989), and are not plausible models at the implementation level. So, I won't consider them here.

of structural similarity, it seems that weighted connections *must* be counted among the elements participating in the similarity. Surely, the relevant definition of structural similarity provided when unpacking condition (a) quantifies only over some, but that "some" seems to include weighted connections. However, if the argument offered above (§4.2) is correct, weighted connections are not structurally similar to their targets. In sum: if artificial neural networks deploys structural representations, connections must be involved. Yet, their involvement seems to prevent the obtaining of (a).

4.5 - Alternative argument #5: Artificial Neural networks, weights and structural similarity III

It is now tempting to wonder whether the relevant definition of structural similarity could be relaxed, allowing connections to participate in the structural similarity in spite of the lack of any intelligible one-to-one mapping holding between them and the elements of the target domains. In fact, some definitions of structural similarity do not seem to require such one-to-one mappings. Kiefer and Hohwy (2019: 400), for instance, define structural similarity as follows:

> "The notion of structural representation is of course only as clear as the relevant notion of structural similarity. Gładziejewski and Miłkowski adopt the definition offered by O'Brien and Opie (2004: 11), which may be paraphrased as follows: suppose that a system S consists of a set of elements E and a set of relations R defined on those elements. We may say system S_1 is structurally similar to S_2 just in case there is a mapping from members of E_1 to those of E_2 and a mapping from R_1 to R_2 that together preserve the relational structure among the elements of S_1 for "at least some" elements and relations in E_1 and R_1." (Kiefer and Hohwy 2019: 400).

Notice that this "paraphrasis" of O'Brien and Opie's definition *does not* require the relevant mappings (from vehicle constituents onto target constituents and from *R* to *R'*) to be *one-to-one*. Hence Kiefer and Hohwy's notion of structural similarity is significantly less demanding than O'Brien and Opie's. Similarly, Shea (2018: 117) argues that structural similarities can

allow for many-to-one mappings.

I do agree with Shea: many-to-one mappings are fine. So, I would allow the relevant definition of structural similarity to be relaxed so as to include many-to-one mappings. What I would not allow (for I believe allowing it would be a mistake) are *one-to-many* mappings. And I believe that the philosophers interested in defending structural representations are better off not allowing them either, for allowing them makes the content of structural representations indeterminate. More precisely, allowing one-to-many mappings makes the content of structural representations disjunctive.[111]

To see why this is the case, let us consider the simplest possible structural representation V. Its vehicle is constituted by two vehicle constituents v_a and v_b, in a relation *R*. Suppose that v_a maps on a target constituent t_a, and that *R* maps on a relation *R'* holding among target constituents. Suppose further v_b maps onto many (for the sake of simplicity, two) target constituents t_b and t_c.

Now, given the mapping sketched above[112], V is accurate when $t_a R' t_b$ is the case. But it is also accurate when $t_a R' t_c$ is the case: after all, v_b *maps also onto* t_c, and so $v_a R v_b$ maps in the desired way also onto $t_a R' t_c$. So, there is also a structural similarity holding between V and $t_a R' t_c$, given the relevant mapping, and, as a consequence, V is accurate also when $t_a R' t_c$ is the case.

But this means that V is inaccurate when, and only when, *both* $t_a R' t_b$ and $t_a R' t_c$ are not the case, and these are the conditions of satisfaction of a vehicle representing ($t_a R' t_b$ or $t_a R' t_c$); hence the content of V is disjunctive. Yet, as seen in (Ch.2: § 2.2), a theory of content must deliver non-disjunctive contents. And, in the case at hand, disjunctive contents are a result of one-to-many mappings. So, it seems to me that, in order for a structural-resemblance based

[111] This issue is further discussed in (Ch.6: §4).

[112] Importantly, I'm assuming that both mappings will be exploitable and thus that they both contribute to determine the content of V.

theory of content to be successful, it must exclude *one-to-many* mappings. Now, the issue with weights in connectionists systems is that they seem to map onto many: each weight encodes information about many targets (see Clark 1993: 13-17; Van Gelder 1991: 42-47; Ramsey, Stich and Garon 1991: 215-217 for early renditions of this point). So, it seems to me that allowing weighted connections to participate in the relevant structural similarity is bound to generate a problem with content determinacy. But, as argued above, there are compelling reasons not to exclude weighted connections from the relevant structural similarity

This, I believe, creates a nasty dilemma for the philosopher willing to resort to artificial neural networks to defend the claim that generative models are structural representations.

4.6 - Alternative argument #6: Wiese's defense of structural similarity

Wanja Wiese (2018: 215-217) puts forth an argument in favor of (a), which might supplement Gładziejewski's original argument.

The argument is straightforward. Wiese notices that the deterministic equations such as:

- $c_2 = f_3(c_3) + \omega_3$
  $c_1 = f_2(c_2) + \omega_2$
  $s = f_1(c_1) + \omega_1$

are often used to describe *both* neural dynamics instantiating a generative model and the causal dynamics of the environment (cfr Ch 1: §2.1; Ch. 3: §4.1). But of course this implies that both the neural dynamic instantiating the generative model and the causal dynamics of the environment satisfy the same mathematical description. And that, at least *prima facie*, entails that the two systems are structurally similar. In fact, the same set of arguments can be mapped onto *both* systems; in a way such that the relevant mathematical relation holding among arguments are preserved in both systems. Isn't this sufficient to establish that a relevant structural similarity is present?

A negative answer seems warranted for several reasons. Firstly, Wiese does not specify how

the mathematical contents of the vehicle constituents are determined. In (Ch. 2: §3.4) I have suggested that mathematical contents seem to be determined by computational implementation: the inner vehicles of a system carry a determinate mathematical contents *because* the system computes a specific function; that is, implements a specific computation. But, plausibly, different accounts of implementation will assign different mathematical contents. Moreover, some accounts might assign *indeterminate* mathematical contents, whereas others might assign no mathematical contents at all (e.g. Piccinini 2015: 137-138; Fresco, Copeland and Wolf 2021; Facchin *submitted*). So, what Wiese needs seems to be an account of computational implementation, which, as far as I can see, is nowhere to be found in his articles. This is troubling, since Wiese (2017; 2018) explicitly claims that the relevant pattern of relations preserved on both sides of the structural similarity is the mathematical structure obtained by ascribing mathematical contents to vehicle constituents (see Ch. 3, § 4.1). But, to put it bluntly, if we do not know *how* mathematical contents are assigned to the vehicle constituents of the structural representation, we simply *cannot know* whether the vehicle of the structural representation is *really* structurally similar to its target.[113]

Secondly, Wiese's argument for (a), even if successful, would fail to vindicate *epistemic* representationalism; at least when it comes to generative models.[114] This depends on the interplay between the vehicle of the structural representation and its vehicle constituents, which are vehicles of input-output representations carrying the relevant mathematical contents (see Ch 3: §4.1, *fn.* 80). Equations of the form:

- $c_2 = f_3(c_3) + \omega_3$
  $c_1 = f_2(c_2) + \omega_2$

[113] Wiese seems to have recently endorsed the mechanistic account of computational implementation (see Wiese and Friston 2021). Endorsing it, however, does not solve the problem I raised here, for the mechanistic account of implementation does not assign determinate mathematical contents to vehicles. Indeed, they seem to allow for multiple assignments of mathematical contents, leaving it indeterminate at best (Piccinini 2015: 137-138; Dewhurst 2018; Fresco, Copeland and Wolf 2021; Facchin *submitted*).

[114] Notice, however, that Wiese's account does vindicate epistemic representationalism in respect to the constituents of generative models, which, in his account, are *at least* construed as input-output representations.

$$s = f_1(c_1) + \omega_1$$

describe the computational functioning of the *entire* generative model. If the entire generative model is a vehicle V, and the description these equations provide is correct, we can conclude that certain parts or vehicle constituents of V carry specific mathematical contents; hence, that they are input-output representations. Hence, according to Wiese, the structural similarity holding between the generative model V and its target T is (at least partially) constituted by the way in which input-output representations are tokened within V. Notice that this fits nicely with Wiese's claim that these equations describe the *dynamics* of the nervous system; for, if we assume the nervous system computes, its activity, and thus, its dynamics, must at least in part consist in the tokening of input-output representations.

However, and this is the key point, notice that, on the side of V, the relevant relations holding among the relevant vehicle constituents (i.e. the input-output representations) are *computational state transition*; that is, the rules according to which input-output representations are tokened within a computational system. So, on Wiese's view, V is the *entire computational system*; that is, a system in which the tokening of representational vehicles takes place. But, in general, computational systems are distinct from the representations tokened within them and they are not considered to be representations (or representational vehicles) in their own right.[115] Compare: it seems literally false to say that *my computer* is a representation of anything, although it seems literally true to say that a physical state within my computer (e.g. a physical state of a register) is a representation. Compare further: it seems false to say that brains are representations of edges, whereas it seems correct to say that specific activation patterns in the early visual cortex are representations of edges.

More in general, it is descriptively accurate to say that in the actual practice of the mind

[115] The point is ambiguous between computational systems as mathematical objects (e.g. a finite state machine defined as a quintuple) and their physical implementations. Whilst real and important, this distinction makes no difference when it comes to the point I'm trying to articulate: mathematical objects are not vehicles, and their physical implementations are typically not considered to be vehicles, as explained in the main text.

sciences computational systems are considered to be distinct from the (input-output) representations tokened within them. Consider the following representative citations:

> "Granting these limitations, we may nonetheless be able to catch a glimpse of what representations might look like within the parallel style architecture of the brain by taking a look inside a connectionist network. *The place to look is in the dynamics of the system*; that is, in the patterns of activity generated by the system of interconnected units." (Churchland and Senjowski 1992: 358; emphasis added).

and, coming from the PP literature:

> "In general, 'representation' in machine learning (and in particular in connectionist approaches) refers to *an internal state of a system that carries information*, as it does throughout most of cognitive science." (Kiefer and Hohwy 2018: 2396; emphasis added).

In both these citations, the authors make a sharp distinction between the (input-output) representations tokened within computational systems and the systems within which (input-output) representations are tokened. Notice further that in *neither* citation the system in which the (input-output) representations are tokened is considered to be either a representation or a representational vehicle in its own right.

Now, to vindicate *epistemic* representationalism, one must provide an account of representations which identifies as representations the kind of things cognitive science ordinarily refers to using the term "representations". *These* are the relevant explanatory posits of cognitive science (or, at least, the relevant explanatory posits cognitive scientists call "representations"). But the relevant structural similarity Wiese points to does not hold between *those posits* and their alleged targets. Rather, it holds between the *entire system* (in which the tokening of those posits takes place) and the environment surrounding it. Hence, Wiese's proposal does not vindicate epistemic representationalism: the structural representations it delivers (if any), do not vindicate the explanatory posits of cognitive science.

Lastly, is it correct to say that the dynamics of the generative model and of its representational target can be described by the same set of equations? Many commentators

(e.g. Baltieri 2019: 34-35; Raja *et al*. *forthcoming*: 40-41) have noticed that this is an *assumption* that is often made in computational modelling. That is, it is common to *assume* that the probability distributions in the model match, in the relevant sense, the probability distributions in the environment. But if this is correct, then it seems that Wiese's argument in favor of (a) is subtly circular, as the empirical studies Wiese invokes to substantiate his argument *assume* (for modelling sake) the presence of the structural similarity Wiese's argument tries to vindicate.

4.7 - Alternative argument #7: "whole brain" representations?

In all the arguments considered above, it was assumed that the relevant structural similarity holding between the vehicles of generative models and their targets should be the upshot of adopting some specifical theoretical perspective on generative models. But what if the relevant structural similarity could be found just by *looking* at the relevant vehicle? Consider, for instance, The following citations:

> "Hierarchical models enable empirical Bayesian learning of prior densities and provide a plausible model for sensory inputs. Single-level models [...] depend on prior constraints for unique inference and do not call upon a hierarchical cortical organisation. On the other hand, if the causal structure of generative processes is hierarchical, *this will be reflected, literally, by the hierarchical architectures trying to minimise prediction error*, not just at the level of sensory input but at all levels" (Friston 2003: 1343; emphasis added)

> "[...] every aspect of our brain can be predicted from our environment. [...] A nice example is the anatomical division into what and where pathways in the visual cortex. Could this have been predicted from the free-energy principle? Yes – *if the anatomical structure of the brain recapitulates the causal structure in the environment, then one would expect independent causes to be encoded in functionally segregated neuronal structures*" (Friston 2013: 133; emphasis added)

Since the definition of structural similarity quantifies only over *some*, these quotes by Friston are *sufficient* to vindicate the obtaining of (a): if Friston is right, there is a structure-preserving mapping from *some* cerebral regions onto *some* environmental targets. Furthermore,

examples of the sort the quote highlights seem fairly easy to multiply. It might be pointed out, for instance, that the anatomical segregation of visual and auditory cortices reflects the fact that visual and sensory input can have different worldly causes. So there *is*, I submit, a relevant brain-world structural similarity. Therefore, if *the whole brain* is the generative model (a claim that is not uncommon in the PP literature, e.g. Bastos *et al.* 2012: 702), then condition (a) is met.

However, it seems to me that such a vindication of (a) is, at best, a Pyrrhic victory.

To begin with, Friston's claim that *if* the structure of the generative process is hierarchical, *then* the generative model must be hierarchically organized in a similar manner is disputable. In fact, "shallow" (three layer) networks can in principle approximate *all* functions computed by deep (hierarchically structured) ones, because three-layer networks can approximate *every* computable function (Hornik 1991). This does not mean that shallow networks compute *better* (or even as well as) deep ones - in fact they don't (Lin, Tegmark and Rolnick 2017). But it means that Friston's conditional is false: hierarchically structured generative processes do not *entail* the presence of hierarchically structured generative models "mirroring" the structure of the generative process. Thus, Friston's arguments lacks the force Friston's wording suggests: a model capturing a hierarchically deep generative process *may*, but *need not*, be itself hierarchically structured. Hence, the relevant structural similarity Friston is seemingly pointing to *may*, but *need not*, be present.

Now, as a matter of fact cortical networks are hierarchically structured, and so the kind of structural similarity Friston is pointing to seems to be present. Yet, notice that such a structural similarity seems to hold between *the entire brain* and the environment. And, typically, the brain is assumed to be a computational system rather than a vehicle tokened in a larger computational system. Thus, whilst the structural similarity Friston points towards can surely vindicate (a), it cannot vindicate epistemic representationalism for the same reasons seen in the previous

subsection. The structural similarity presented above seems to enable us to vindicate *only* metaphysical representationalism about the whole brain (i.e. the claim that the whole brain *really* is a "big" representation). Given that Gładziejewski's account of structural representations aims at vindicating both metaphysical and epistemic representationalism, this way of vindicating (a) seems to lead to a partial failure of his account.

One might contend this verdict is premature because vehicle constituents and their relations of structural representations are representational vehicles in their own right (e.g. Shea 2018: 118; Ramsey 2007: 79, footnote 3). Thus, claiming that the brain as a whole is a structural representation might in principle justify the claim that the relevant elements of the structural similarity (i.e. patterns of activation) are representations too, thus vindicating epistemic representationalism. I believe that the problem with this line of reasoning is the following: the brain-world structural similarity Friston envisages is *not* defined over patterns of activation in the brain. Rather, it is defined over the anatomical structure of the brain. The relevant elements in the structural similarity are not patterns of activation. Hence, this way of vindicating (a) entirely fails to vindicate the epistemic representationalist claim.[116]

Secondly, a complaint about content. What would such a "whole brain" structural representation represent? If I understand Friston correctly, the brain is supposed to recapitulate the causal structure *of the world* or *the environment*. Thus, the relevant structural similarity holds in between the anatomical structure of the brain and the causal structure of the world/environment. But a structural representation represents the target whose structure is mirrored in the structure of the vehicle, and here such a target is *the world/environment* (see Wiese 2018: 219; Williams 2018a: 154-155). This is not the kind of content naturalistic theories of content are supposed to deliver, for *the world/environment* is not the kind of content invoked

[116] Thus notice that, in this regard, the proposal under examinations scores worse than Wiese's one, which was able to vindicate epistemic representationalism about patterns of neural activation, treating them as input-output representations.

in the scientific explanations of our cognitive capacity, nor the kind of content relevant to our personal-level mental states. This isn't a knockdown objection, but it surely shows that the argument has some very undesirable consequences.

Lastly, and, I believe, most importantly, this way of vindicating (a) seems to prevent (c) from obtaining. If the entire brain is a single gigantic representation representing the world/environment, it is very hard to see how decouplability might be met. There is always *some* sort of causal contact between brains and world/environment. Since point (c) spells out decouplability in terms of causal contact, this way of vindicating (a) seems to prevent the obtaining of (c).[117]

This problem is even more worrisome than it *prima facie* appears. For, arguably, the representational target of a generative model is not "the world", in the sense of the extra-organismic environment. Rather, as seen in (Ch. 3 (§3), the target of a generative model is the generative process, which includes the animal's body and bodily responses. And surely well functioning brains are never decoupled from those. And even if there are extreme cases in which one's brain is decoupled from one's body (e.g. if one suffers from complete locked-in syndrome, or brain-in-a-vat scenarios), one is *never* decoupled from the generative process. For "the generative process" simply denotes the process generating one's sensory signals. And there is no decoupling from that - not even in cases of complete sensory deprivation: the absence of a sensory signal is a sensory signal in its own right.

At this juncture, one might be tempted to purge the relevant account of structural representations from point (c), thereby vindicating the claim that generative models are structural representations of the relevant generative process. As far as I can see, this is a

[117] One might worry that the arguments presented in the two previous indents hinge upon an extremely uncharitable interpretation of Friston, as he sometimes more cautiously claims that the structure of the generative model "mirrors" the structure of the *generative process* (rather than the one of the world/environment). Whilst entirely correct, I fail to see how such a reading is more charitable. For one thing, it entirely prevents the obtaining of (c), as described in the main text.

legitimate move. However, it seems quite an *ad hoc* move. There are good independent reasons to hold that representations are necessarily decouplable from their targets (see Grush 1997; Webb 2006; Pezzulo 2008: Orlandi 2014: 120-134; 2020). Moreover, abandoning (c) would likely make Gładziejewski's account of structural representations far too liberal, as Gładziejewski himself acknowledges (Gładziejewski 2016: 571).

4.8 - Alternative argument #8: Making "whole brain" representations work?

Perhaps there is a way to make "whole brain" representations work. Thus, consider Kiefer and Hohwy's (2018; 2019) defense of generative models as structural representations.[118]

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

Kiefer and Hohwy's account of "whole brain" structural representations seems to me a significant improvement from the previously scrutinized one. For one thing, given that in this view the relevant elements of the structural representation are patterns of activation, and given that the elements of a structural representations can be counted as representations in their own right, Kiefer and Hohwy's proposal seems better poised to substantiate the epistemic

[118] To be fair, Kiefer and Hohwy do not *explicitly* set out to defend "whole brain" representations. However, it seems to me that their account entails that the whole brain is a structural representation, at least insofar they take the entire causal network *instantiated by the brain* to be the relevant structural representation.

representationalist claim.

However, it seems to me that relying on Kiefer and Hohwy's proposal to vindicate (a) has serious drawbacks.

To start, the problem with (c) is not solved by Kiefer and Hohwy's account.[119] If the brain is a complex causal network mirroring the causal structure of the world, it is correct to say that the relevant structural representation (i.e. the brain) represents the world. I simply do not see how one could sever the constant brain-world (or brain-generative process) causal contact so as to vindicate (c).[120]

Secondly, Kiefer and Hohwy's account rises a puzzle about the inferential status of brain processes. If causal relations holding among patterns of activation are the relations holding among vehicle constituents that "mirror" the relations on the other side of the structural similarity, it follows that they *are part of the vehicle*. But if this is the case, then it seems to me that these causal relations cannot be inferential processes, for inferential processes seem to be distinct from the representational vehicles upon which they operate. So, it seems that if Kiefer and Hohwy's (2018; 2019) account of structural similarity is accepted, causal interactions among neural activity patterns cannot be rightfully called inferences. And this seems a problem, given that the inferentialist reading of predictive processing tends to go hand in hand with the claim that generative models are structural representations (e.g. Kiefer 2017; Gładziejewski 2017; Hohwy 2018).[121]

Lastly, a wholesale acceptance of Kiefer an Hohwy's (2018; 2019) account might,

---

[119] Notice, importantly, that Kiefer and Hohwy consider decouplability a necessary feature of representations, see (Kiefer and Hohwy 2019: 400)

[120] Of course, individual patterns of activation can be decoupled from the target they represent in virtue of the overall brain-world structural similarity. However, point (c) is defined over the entire structural representation, not its individual vehicle constituents.

[121] Notice that this problem is closely related to a problem emerged in §4.6; namely the fact that, on Wiese's account, computational state transitions are taken to be the relevant relations holding among vehicle constituents. In both cases, what happens is that a set of relations typically defined over representations (computational transitions defined over input-output representations or inferential relations) ends up being treated as the set of relations holding among the vehicle constituents of the generative model that "mirrors" the relational structure of the model's target.

paradoxically, force one to abandon the claim that generative models are structural representations. The point is subtle but important. According to Kiefer and Hohwy:

> "The contents of parts of a structural representation are (at least in the case of causal generative models of an environment) in effect determined by their internal functional roles." (Kiefer and Hohwy 2018: 2393; see also Kiefer and Hohwy 2019: 402; Kiefer 2020, endnote 17)

But this is not how the parts (i.e. vehicle constituents) of a structural representation acquire their contents. The content of a structural representation is determined by the relevant structural similarity it bears to a target; and the content of its vehicle constituents is determined by the way in which they participate in the relevant structural similarity; that is, by the way in which they map onto a corresponding element of the target (e.g. Cummins 1996: 96). The relevant relation determining the contents of the elements of a structural representation is the structural similarity holding between the vehicle and the target; not the relations holding among vehicle constituents. Surely, since structural similarity is *structural* it must, in some relevant sense, be *sensitive* to these relations. But this does not entail that they *determine* the content of the vehicle constituents of a structural representation.

Another, perhaps more perspicuous, way to flesh out the same point is this: were the content of the vehicle constituents determined by the relations holding among them, then a vehicle constituent v_a would represent a target constituent t_a *whether* V is structurally similar to T or not. If the content of vehicle constituents is determined by the relevant relations holding among them, it follows that their content is *not* determined by the structural similarity holding between V and T (if any), for the relations holding among vehicle constituents (and hence their contents) would be the exactly the same even in cases in which *no* structural similarity between V and T holds.[122] **Figure 4** exposes the point in a pictorial format.

[122] Moreover, even if V and T are structurally similar, there is no principled reason as for why the contents that vehicle constituents would bear, were these content determined by their relations, should match the content they would bear, were their content determined by the relevant structural similarity.

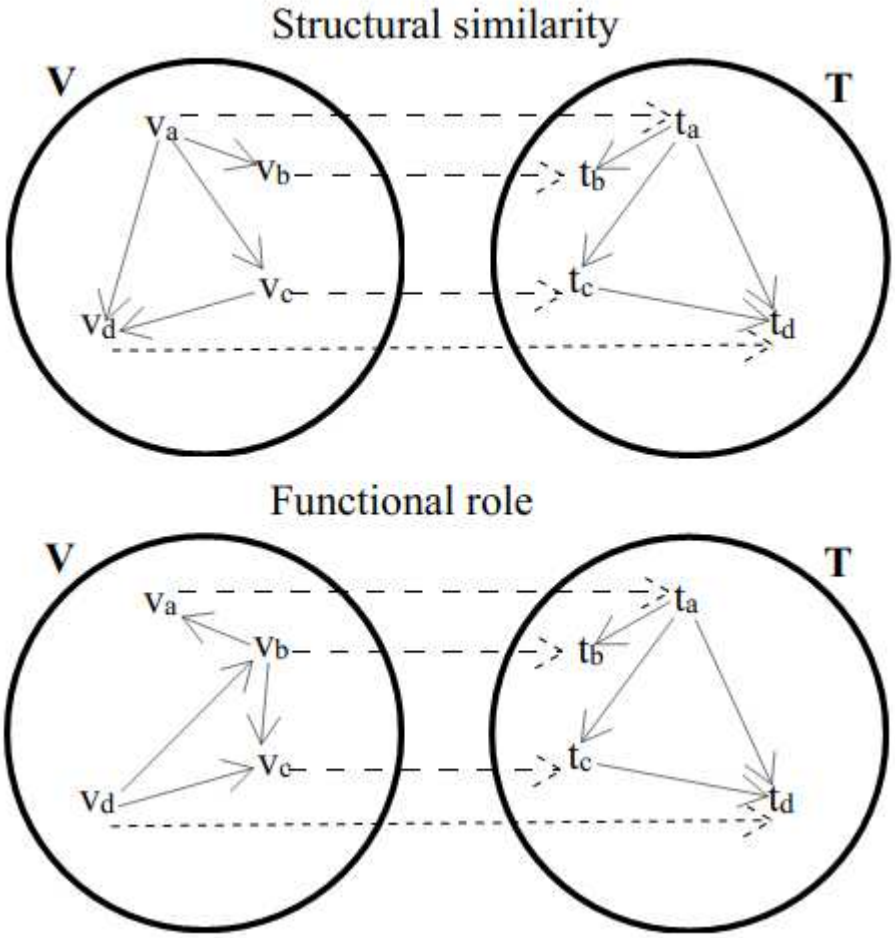

Figure 4. According to a structural similarity theory of content a vehicle constituent v_a represents a target constituent t_a *because* V is structurally similar to T. Conversely, according to functional role semantics, v_a represents t_a *because* it bears some specific relations with over vehicle constituents, whether V is structurally similar to T or not. The graphical conventions are the same as in figure 3 (drawing by the author).

To put the point bluntly, what I'm trying to point out is this: Kiefer and Hohwy espouse a form of functional role semantics. But functional role semantics and structural similarity have no essential connections, *pace* Kiefer and Hohwy. It thus seems to me that a wholesale adoption of Kiefer and Hohwy's proposal ends up undermining the broader structural-representationalist claim. Kiefer and Hohwy might provide a way to vindicate (a); but a wholesale acceptance of their proposal seems to make such a vindication redundant. If one adheres to functional role semantics, one is *not in need* of a structural similarity.

This is not to deny that Kiefer and Hohwy (2018: 2393; 2019: 402) stress that the relevant

(i.e. content conferring) functional relations among vehicle constituents mirror, in the relevant sense, the causal structure of the world: on their view, functional role semantics *entails* a relevant structural similarity. But this surely does not allow us to count Kiefer and Hohwy as defenders of structural representations. For, on their account, the relevant structural similarity does not determine the content of the vehicle, nor, strictly speaking, holds between the vehicle and the target. In fact, on the view functional-role semantics offers, what I have thus far called "vehicle constituents" are simply *vehicles*, whose content is determined by their mutual relations. And no such vehicle is (or needs to be) structurally similar to anything. Hence, when it comes to vindicate (a) the problem with Kiefer and Hohwy's account is the same problem examined in §4.4 above: the structural similarity does not hold between a vehicle and a target, but rather between the entire set of representational vehicles tokened within a system and the set of targets the system can represent.[123]

4.9 - Alternative argument #9: The "whatever" argument

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

[123] There is, however, an important difference between Kiefer and Hohwy's proposal and the one examined in §4.4. Whereas in §4.4 the structural similarity between the set of representational vehicles and the set of represented targets determined the content of each vehicle, in Kiefer and Hohwy's case the content of each vehicle is *not* determined by that structural similarity, but by the relations holding among vehicles. Thus, in their view, the structural similarity, even if present, is purely "epiphenomenal", as it does not contribute to content determination.

5 - Tacking stocks (and pointing forward)

I have examined the argument Gładziejewski offers to claim that generative models are structural representations. I have argued that the argument is faulty, because it fails to establish that the vehicles instantiating generative models bear an exploitable structural similarity to their targets. I have also considered some alternative arguments to the same effect, and argued that none of them suffices to establish the desired conclusion. Hence, at present, the claim that generative models are structural representations is unjustified.

The previous discussion might have left a halo of confusion concerning the physical structures implementing generative models: are these entire brains/neural networks, or neuronal responses, or sets of connections about processing units, or something else entirely? The correct answer is provided in §4.5 above: they are patterns of activation and sets of weighted connections between units. In Ch. 6, I will carefully analyze one such structure, claiming that, thusly conceived, generative models are non-representational structures instantiating an agent's sensorimotor mastery.

But before doing so, I must face a more pressing problem. The bulk of the argument I have offered here concerned the fact that, at present, there is no convincing reason to claim that the physical structures encoding generative models are structurally similar to their targets. But *what if it were*? Would the structural-representationalist reading of predictive processing be vindicated? I think that the correct answer is negative, because structural representations do not meet the Job Description Challenge. I will defend this claim in the next chapter.

Chapter Five - Structural representations do not meet the Job Description Challenge[124]

1 - The Job Description Challenge

In the last chapter, I've argued that, at present, we have no reason to believe that the vehicles of generative models are structurally similar to their targets. But what if such a structural similarity were to be found? Wouldn't that turn the tide in favour of a structural representationalist reading of PP?

Here, I argue the answer to this question is negative, for structural representations, as Gładziejewski (2015b, 2016) defines them, do not meet the "Job Description Challenge" (Ramsey 2007). For this reason, conditions (a) to (d) do not spell out a *representational* functional profile, and items satisfying them do *not* function *as representational vehicles* within the systems in which they operate.

As I understand it, the Job Description Challenge (Ramsey 2003; 2007; 2016) begins with the following premise: representations belong to *two distinct kinds* at once.

To start, representations belong to an *intentional kind*, as they necessarily encode, carry, or "have" *content*. Representations are items[125] (vehicles) that *are about* other items (targets). If I say that something is a representation, it makes perfect sense to ask what its target is, and how the target is represented (see Ch. 2: §2.2).

Secondly, representations belong to a *functional kind* (Ch. 2: §2.3). The point is well accepted in the relevant literature[126], yet I know of no "official" argument substantiating this claim. As far as I can see, the claim rests only on two informal, yet persuasive, observations.

One is that the items belonging to the class of representations are *unruly disjunctive*, as the

---

[124] This chapter is based on, and expands upon, Facchin, M. (2021c). Structural representations do not meet the Job Description Challenge, *Synthese*, https://doi.org/10.1007/s11229-021-03032-8

[125] Here, "item" is used broadly, to designate at once objects, events and states of affairs.

[126] On representations forming a functional kind, see (Peirce 1931-58; Millikan 1984; 2020; Haugeland 1991; Ramsey 2007; 2016; 2020; Lee 2018: 2).

class includes pictures, sentences, engraving, utterances, maps, scale models and (according to many) neural states. Moreover, it is a class whose borders can be constantly expanded, as we might in principle *use any object* to represent something else (e.g. we can stipulate that my glasses represent *x*, so when you see me wearing glasses you know *x* is the case). It seems clear that it is very unlikely that all these items will end up having some *common property*[127], around which the class of representations is built. It thus seems more likely that what ties all these items together in a single class is a common *function*. Functional kinds can in fact be realized in many different ways: a quick survey on Wikipedia, for instance, reveals there are no less than six different types of pumps. Nevertheless, all these pumps are (*bona fide*) pumps because of *what they do*; namely, displacing fluids (liquids or gasses) by exerting some mechanical action. That is, all pumps are pumps because of the *function* they perform, in spite of their superficial differences. Hence, if "representation" really denotes a functional kind, it can easily be understood why the various different types of representations all qualify as members of a single kind (Ramsey 2007: 7-14)

The other is that representations can be, and often are, characterized *functionally*, as "stand-ins" for absent targets.[128] As cognitive science conceives of them, representations are theoretical posits that are needed to explain how intelligent systems can organize their behavior in respect to something that is not actually present. But, as Orlandi (2020) has recently noticed, this means that representations all have a specific function; namely, that of allowing a system to "coordinate with the absent". Strikingly, this is roughly the same role *public* representations are supposed to play: namely, that of providing us information about targets out of our

[127] Intentionality or "aboutness" might be considered as the property tying all representations together in a single kind. However, things are not so straightforward. On the one hand, intentionality or aboutness might *not* be a single property (for example, original and derived intentionality might be two different properties). On the other hand, there might be forms of "aboutness" or intentionality that have nothing to do with representations - or so some phenomenologists and enactivists claim (e.g. Hutto and Myin 2017; Rietveld, Denys and van Westen 2018). See also (Schlicht and Starzak 2021) for a fair critical discussion of these proposals.

[128] See, for instance, (Haugeland 1991; Grush 1997; Clark 1997; Clark and Grush 1999; Pezzulo 2008; Orlandi 2020)

immediate reach, so as to allow us to coordinate with them *in spite* of their absence.[129] And this, again, suggests that representations all have a similar functional profile.

Now, if representations really belong to both kinds, and cognitive scientists really posit *representations*, it follows that the relevant posits must satisfy two demands in conjunction. First, they must have content. Secondly, they must have a representational functional profile. Notice that both demands spell out a *necessary* condition: something belongs to an intentional (or functional) kind *only if* it possesses content (or the relevant functional profile). Notice further that, albeit these two requirements need to be satisfied in conjunction, they are conceptually independent (Ramsey 2016) and are satisfied in very distinct ways.

Naturalistic theories of content (e.g. Millikan 1984; Fodor 1990) are typically invoked as a principled way to satisfy the first requirement, as they aim at accounting for the content or aboutness of representations in non-semantic and non-intentional terms. These theories are not my focus here.

My focus will be squarely on the *second* requirement of the Job Description Challenge. Sadly, however, it is a bit unclear how the second requirement should be satisfied. As far as I can see, there is no *analogon* of naturalistic theories of content when it comes to satisfy the second requirement; hence we lack general theories of representational functioning.[130] In fact, aside from fairly quick remarks on "standing-in" (see references given above) no well-defined functional characterization of reprepresentation has been offered (see Millikan 2020; Egan 2020 for this complaint).

Here, "compare-to-prototype" arguments, as Gładziejewski (2016) dubbed them, come into play.[131] This is how Gładziejewski understands them:

---

[129] See, for instance, (Bechtel 2008: 159-161; Godfrey-Smith 2009).

[130] Perhaps (Orlandi 2020) could be read as proposing one, but her account still deals with aspects of representations that at least prima facie have little to do with their functional profile, such as being explanatory posits of psychological sciences.

[131] It should be noted, however, that such a similarity, albeit *sufficient* to meet the challenge, is not *necessary* to meet it. In fact, Ramsey seems to allow that certain posits actually qualify as genuinely representational mostly

> "[...] one starts out by pointing out a type of structure that can be pretheoretically categorized as a representation in an uncontroversial way. In particular, one concentrates on the functions served by the structure in question—on what it does for its users that makes it a representation. This is our representational prototype. Subsequently, one concentrates on a particular concept of representation used in cognitive science and verifies whether structures that fall under this concept have a functional profile that matches, to a non-trivial degree, the functional profile of the pretheoretical prototype. In other words, one asks whether a given type of representation posited by cognitive scientists plays a functional role that is similar enough to the role played by the prototype that the former can be naturally regarded as (truly, non-trivially, genuinely, etc.) representational in nature. If it does play this role, then the Job Description Challenge is successfully met." (Gładziejewski 2016: 565)

The idea behind this procedure is intuitive and straightforward: in fact, "compare-to-prototype" arguments simply are arguments by analogy. If some putative representation posited by a cognitive theory (or family thereof) functions in a way that is sufficiently similar to some paradigmatic public representation, such as a map, then we have *at least* a pretheoretical and intuitive understanding of how that posit functions *as* a representation within the cognitive system; namely, it functions as a representation *by functioning as a map*. Notice, however, that the very same procedure can deny that a given posit actually qualifies as a representation. If, for instance, an alleged representational posit has a functional profile nontrivially similar to that of a battery, we clearly cannot say that it functions as a representation *by functioning as a battery*. Indeed, if the structure under scrutiny is *correctly* characterized as a battery, describing it as a representation (e.g. by saying it represents how much longer a process can still run) is explanatorily redundant, and might put at risk future research (Freeman and Skarda 1990; Webb 2006). In the case at hand, future research would be hindered because considering that item as a representation leads us to wonder *how content is encoded* rather than *how energy is stored*.

To further clarify the matter, I will now consider two "prototypical" compare-to-prototype

---

because of their explanatory role within a theory. Arguments by analogy, however, are by far the most popular way to confront the challenge, and therefore they will be the focus of the present treatment.

arguments. The first concerns receptors, and yields a negative result. The second concerns structural representations, and allegedly yields a positive result.

1.1 - Compare-to-prototype: receptors and firing pins

The first case I intend to examine is that of receptors, which provide the (almost) universally accepted case of a representational posit *failing* the Job Description Challenge.[132] Painted with a broad brush, the idea behind the receptor notion of representation is fairly simple: if an internal state V of some system reliably co-occur with some distal event T, then V is a representation of T.

Receptors are often further elucidated referring to Dretske's (1981; 1988) account of representation (Ramsey 2003; 2007; Morgan 2014; Nirshberg and Shapiro 2020). At the core of Dretske's account of representation lies the notion of indication. As defined in (Ch. 2: §3.1):

> **Indication**: For all the states of V and T in a relevant range of states, v_a indicates t_a if, and only if, $P(t_a|v_a) > P(t_a)$; that is, the occurrence of v_a increases the odds of t_a being the case[133]

Recall that, to determine content in an appropriate way, this notion of indication[134] needs to be conjoined with a teleological component. To represent T, V need not only indicate T, it also must be "supposed to" indicate T, where the "supposed to" part gets unpacked by saying that V is supposed to indicate T just in case V has been recruited within some system in virtue of the fact that it indicates T (according to revised definition of indication). The recruitment procedure might vary: Dretske (1988) extensively relies on reinforcement learning, but natural selection and intentional design are typically held to be sufficient recruitment procedures too (e.g. Neander 2017; Shea 2018, Ch. 3).

---

[132] See (Ramsey 2003; 2007; Orlandi 2014; Anderson and Chemero 2019; Williams and Colling 2017; Downey 2018). For a tight defense of their representational status, see (Artiga 2021).

[133] See (Dretske 1988; Rupert 2018).

[134] As clarified in (Ch. 2: §3.1, fn. 46), not all informational accounts of content need to incorporate a teleological component. However, accounts that do *not* incorporate it face formidable challenges, see (Artiga and Sebastián 2018; Roche and Sober 2019).

Several structures[135] qualify as receptors according to this picture. Single neurons, for instance, are often said to represent whichever distal variable (object or state of affairs) triggers their suprathreshold firing the most (e.g. Levittin, *et al*. 1959; Hubel and Wiesel 1962, 1968). In this view, their increased firing rate indicates the presence of some specific target in the animal's visual field (see Eliasmith 2005 for an updated discussion). In a similar spirit, the nodes in the hidden layers of connectionist architectures are often said to represent the input patterns with which their activity correlates the most. Furthermore, each individual node is said to represent the microfeature driving the node's activity the most (e.g. Goschke and Koppelberg 1991).

It seems obvious that these structures can leverage Dretske's (1988) account of content to satisfy the demand for content. Yet, they seem unable to satisfy the *functional* demand. Indication is surely not sufficient for representation (the sea level indicates the position of the moon, but surely the sea does not represent the moon[136]). Having the function of indicating does not seem sufficient either. In fact, all sorts of things are recruited within systems in virtue of their indicator properties, without thereby becoming representations of what they indicate. Bi-metallic strips of thermostats and photosensitive cells of optical smoke detectors all have the *function* (by purposeful design) of indicating some distal target; yet they are not, *prima facie*, representations. In fact, within these mechanisms, both receptors act just like reliable causal mediators, allowing the system to robustly produce a certain output (for instance, turning off a furnace) when a given environmental condition obtains. The same holds, for instance, for

[135] A reviewer of the journal *Synthese* noticed that taking entire structures as representations is a deviation from Dretske's framework. In Dretske's view, it is not correct to say that, for instance, a barometer represents the pressure. Rather, we should say that the barometer being in state *s* represents the fact that the pressure is *n* Pascals. However, this loose usage is not just prominent in the literature (e.g. Morgan 2014: 231-232; Williams and Colling 2017: 1947), it also strikes me as entirely unproblematic. To continue with the previous example, the claim that a barometer represents the pressure is entirely intelligible and easily unpacked by saying that the barometer represents the pressure of a given environment by occupying, at any moment, the state that indicates the pressure at that moment.

[136] In order to justify this claim, it is sufficient to notice that the level of the sea cannot misrepresent the position of the moon. But something can count as a representation only if it can misrepresent in at least some cases.

the firing pin of a gun. The state of the firing pin indicates the position of the trigger: if the firing pin is in contact with the bullet, then the trigger has (typically) been pulled. Hence P(trigger pulled|firing pin in contact with the bullet) > P(trigger pulled). Moreover, firing pins are included in guns *because* of this relation: it is the fact that their position indicates whether the trigger has been pulled that enables us to control when to shoot. But surely guns are not representational systems. Thus, when it comes to the functional profile of receptors, they behave as mere causal mediators (such as firing pins); and, for this reason, they shouldn't be considered representations. Indeed, many believe that considering receptors as representations has nasty consequences.

Panrepresentationalism is the first. Considering receptors as representations entails that they satisfy both the demand for content requirement and the functional requirement imposed by the Job Description Challenge. But then it is almost impossible to deny bi-metallic strips (or firing pins) also satisfy them. Given the shared functional profile, if receptors satisfy the functional demand, then bi-metallic strips (and the like) satisfy it too. And we can apply Dretske's (1988) account of content to allow them to satisfy the content demand. After all, they have, by design, the *function* of indicating something within the systems deploying them. Thus, accepting that receptors are representations entails panrepresentationalism: the (clearly mistaken) view that whichever entity reliably coordinates with environmental contingencies is *representing* these contingencies.[137] But any account of representations entailing

[137] Notice here that panrepresentationalism is a problem only because I'm assuming that the content at play here is *original*. There is, I believe, no problem of panrepresentationalism related to *non-original* (or derived) content, for each and every thing can, in principle, be assigned some derived content. We could surely stipulate, for instance, that a mug represents Napoleon, or that a pair of shoes represents Castor and Pollux. This seems also the reason why semioticians (who are interested in representations with both original and derived content) have no problem in saying, for instance, that a cigarette butt found on a crime scene represents the fact that the murder is a smoker, or that finding my fingerprints on a surface signals the fact that I touched that surface. In all these cases, the relevant signs (or representations) are tied to their targets only by a loose causal connection. However, this does not generate any problem with panrepresentationalism because their content is derived, as it depends on the interpretation of some clever detective (or some other interpreter). Notice further that the distinction between mental and public representations is *orthogonal* to the distinction between original and derived content according to at least some naturalistic accounts of content. For instance, according to Millikan's teleosemantics, bee dances have original content, even if they are not mental representations (see Millikan 1984; see also Lyre 2016; Vold

panrepresentationalism is surely metaphysically flawed, as it fails to establish a substantial distinction between representational and non-representational states (Ramsey 2003; 2007: 125-127).

The empirical adequacy of the relevant notion of representation is also under threat. If philosophical theories of cognitive representation aim at capturing the notion of representation cognitive science deploys, they must provide a notion of representation which is distinctively psychological or cognitive. But a notion of representation that applies to thermostats or firing pins seems to lack any distinctively psychological or cognitive connotation (Orlandi 2014: 107-110; Ramsey 2017, see also Webb 2006).

Accepting that receptors are representations also reduces the *explanatory power* of the notion of representation invoked. Since treating bi-metallic strips (and the like) as representations add nothing to our non-semantic comprehension of these devices, the notion of representation appears to be merely a semantic *gloss* glued to an ultimately non-semantic understanding. This explanatorily inert notion of representation is at odds with the representationalism of cognitive science – at least as long as we regard it as a *substantial* empirical hypothesis (cf Ramsey 2017).

Many found that these problems are collectively sufficient to reject the receptor notion of representation (e.g. Ramsey 2003; 2007; Orlandi 2014; Downey 2018). And even when the notion is not *explicitly* rejected, more than a shadow of doubt is cast over its explanatory potential (e.g. Williams and Colling 2017: 1949). Importantly, as the essay by Williams and Colling nicely testifies, structural representations are often taken to be substantially immune from these problems, as they *do* meet the Job Description Challenge. Or so Gładziejewski (2015b; 2016) apparently showed.

---

and Schlimm 2020). There might even be mental representations whose content is not original (see Clark 2010 for a possible case). Hence the problem of panrepresentationalism cannot be avoided just by stipulating that public representations have only derived content. One has to *argue* for that claim, and doing so forces one to confront prominent accounts of content, such as Millikan's.

1.2 - Compare-to-prototype: structural representations and maps

Recall how Gładziejewski (2015b; 2016) defined structural representations:

> **Structural Representation:** In a system S, a vehicle V is a structural representation of a target T if, and only if:
>
> (a) V is structurally similar to T; &
>
> (b) V guides S's actions regarding T; &
>
> (c) V can satisfy (b) even when decoupled from T; &
>
> (d) S can detect the representational error V generates.

On Gładziejewski's (2015b; 2016) view, features (a)-(d) are the functional features paradigmatically associated with a class of items we pre-theoretically recognize as representations; namely, cartographic maps.

Very little elaboration seems needed. To start, recall the relevant notion of structural similarity:

> **Second-order structural resemblance (rewritten):** V is structurally similar to T if and only if:
>
> (i) there's a one-to-one mapping from at least some vehicle constituents (v_xs) onto at least some target constituents (t_xs); &
>
> (ii) there is a one-to-one mapping from at least a relation *R* holding among the vehicle constituents onto at least a relation *R'* holding among the target constituents; &
>
> (iii) For all the vehicle constituents satisfying (i), $v_a R v_b \rightarrow t_a R' t_b$ (i.e. the same *pattern* of relations hold in V and T)

It is intuitively clear that cartographic maps satisfy (i) to (iii). Consider a map of Italy. Each point on the map (i.e. each vehicle constituent) maps onto an Italian city (i.e. a target constituent). And the relevant spatial relation holding among vehicle constituents map onto the relevant spatial relations holding among cities, in a way such that the relevant *pattern* of relation is preserved on both sides of the mapping. So, for instance, if the map shows a point v_a been *left of* v_b, then the city t_a upon which v_a maps, is *east of* the city t_b, upon which v_b maps.

Secondly, we do *exploit* the relevant structural similarity holding among maps and the

terrains they represent. In particular, we are systematically sensitive to the relations holding among the vehicle-constituents of the map, for instance when we scan the map to find the *shortest path* from v_a to v_b so as to decide which road to take. And we generally use maps to interact with targets that are *significant* to us, given our current tasks and purposes (e.g. arriving on time at a conference venue, or finding our hotel in a foreign city).

Cartographic maps can also be surely used in a totally decoupled fashion. I can rely on a map of Tokyo to plan my trip to Japan while still in Europe; that is, when no causal contact ties me (or the map) to Japan.

Lastly, we can detect the representational error of maps. For instance, if by relying on a map I reliably get lost, it is likely that I will deem the map inaccurate, and buy another one.

It thus appears that structural representations, as Gładziejewski (2015b; 2016) defines them, "fit" the prototype offered by cartographic maps. For this reason, it seems that structural representations successfully meet the Job Description Challenge, and, unlike receptors, really qualify as representations.

In the next paragraph, I will argue that this is not the case. More in detail, I will argue that at least *some* receptors satisfy points (a) to (d) too. Hence, if, as many agree, receptors paradigmatically *fail* the Job Description Challenge, it should be concluded that structural representations, as Gładziejewski conceives of them, fail it too.

2 - Structural representations fail the Job Description Challenge

In this section, I will argue that structural representations, as Gładziejewski (2015b; 2016) conceives of them, actually fail the Job Description Challenge. To do so, I offer a two stepped argument. First, I will show that *at least some* receptors satisfy, as a matter of fact, points (a) to (d). Secondly, I will show, by means of a "compare-to-prototype" argument, that some *non-representational* structures (such as capacitors) satisfy (a) to (d). I will therefore conclude that

structural representations, as Gładziejewski defines them, do not meet the Job Description Challenge.

2.1 -At least some receptors satisfy (a) to (d) in conjunction

I wish to substantiate two claims, with a different scope. The first is a *universal* claim: I will argue that *all* receptors satisfy (a) and (b). More precisely, I will claim that if a candidate receptor V does *not* satisfy (a) and (b), then V *cannot be* a receptor, because it does not indicate its target. Having done so, I will argue that, as a matter of fact, *some* receptors can also satisfy (c) and (d). I will, however, discuss each point (a) - (d) separately.

2.1.1 - All receptors satisfy (a)

All receptors satisfy (a). This is not a new claim, and it is well-attested in the literature on the argument (Morgan 2014; Nirshberg and Shapiro 2020). In fact, Gładziejewski and Miłkowski (2017) have conceded the point, if only as a matter of contingent empirical fact. I believe instead that all receptors satisfy (a) *as a matter of conceptual necessity*, but the claim is better presented when discussing (b), so I postpone its discussion to (§ 2.1.2).

As for now, let me illustrate *why* every receptor satisfies (a). Consider a paradigmatic receptor such as the bimetallic strip of a thermostat. It surely indicates the temperature: finding the strip occupying a given state raises the probability that the temperature in the room is in the corresponding state. Moreover, the strip has the function of indicating the temperature. In fact, bi-metallic strips are included in thermostats (by human design) precisely because of their properties as indicators.

It is fairly easy to show that such a receptor is structurally similar to the environmental temperature (its target). Let the various states of the strip be defined as elements v_x belonging to a set V, and let the range of temperatures indicated by the strip be defined as elements t_x

belonging to a set T. By definition, V and T have the same cardinality. Moreover, since each element of V indicates one and only one element of T, the one-to-one mapping from v_xs onto t_xs required by (i) obtains.

Let now two relations be defined, one (*longer than*) over the elements of V, and one (*hotter than*) over the elements of T. Notice that these relations are not *arbitrarily* defined: in fact, they are essential to the functioning of the thermostat. Importantly, these two relations can be easily mapped onto one another as (ii) requires.

Notice now that both relations impose a strict total order among the elements over which they are defined: for each arbitrary pair of elements (v_a,v_b) ordered by *longer than*, there exists a pair (t_a,t_b) ordered by *hotter than* such that v_a maps onto t_a and v_b maps onto t_b. Hence, V and T have the same internal mathematical structure and *non gratuitously* map onto each other, thereby satisfying (iii).[138] Notice that this is just an abstract description of the way the bi-metallic strip works: it gets longer as the temperature rises. Hence, the relation of indication making the bi-metallic strip a receptor of the environmental temperature is *per se sufficient* for a structural similarity to obtain between the two.

This point easily generalizes. Given any arbitrary receptor, its states will always map one-to-one onto the states of the environment they indicate, providing the mapping in (i). The states of the receptor and the states of the environment will also always bear some receptor specific reciprocal relations, providing what (ii) requires.[139] Lastly, each arbitrary pair (or other polyadicity) of receptor states in a given relation will map one-to-one onto the corresponding states of the environment in the corresponding relation, satisfying (iii). This is just how receptors work. Thus, (a) obtains for all receptors.

---

[138] I owe the phrasing of this point to my colleague Silvia Bianchi.

[139] Some examples in service of intuitive clarity: the hair in a hair hygrometer gets *longer* as the humidity *rises*; the floating unit of a fuel gauge gets *lower* as the tank gets *emptier*; the return signal of a proximity sensor is *faster* as the target gets *closer*, and so on.

2.1.2 - All receptors satisfy (b)

Receptors can surely guide a system's behavior. A number of automated (or semi-automated) systems deploys them, from automatic faucets to simple, "purely reactive" robots (e.g. Nolfi 2002; Pfeifer and Bongard 2007).[140] But that does not mean that any receptor target structural similarity is *exploited*; or so, at least, Gładziejewski and Miłkowski (2017) argue.

Their argument is roughly as follows: consider again the bi-metallic strip of the thermostat. Let it be sensitive to three environmental temperatures, ordered by *hotter than* in the triplet (t_a,t_b,t_c). Let v_a, v_b and v_c be the corresponding states of the bi-metallic strip. Suppose now that *longer than* orders these states in the triplet (v_b,v_c,v_a), which prevents the relevant strip-temperature structural similarity from obtaining. Yet the strip can still successfully orchestrate the behavior of the thermostat, at least as long as it enters in each state when the environment is in the corresponding temperature (i.e. as long it correctly indicates) and each state leads the system to behave as it has been designed to behave. So, the relations among indicator states are *irrelevant* to the functioning of the system. As a consequence, the structural similarity is not exploited, as a structural similarity is exploited *only if* a system is sensitive to the relations holding among the relevant features of the vehicle, in our case the indicator states (Shea 2014; 2018 p. 120). Receptors might be structurally similar to their targets (and as a matter of contingent empirical fact they are). Yet, this similarity does nothing for the system and deserves to be called a mere epiphenomenon.

In reply, I claim that there exists *at least one* target-receptor structural similarity which every receptor *must* instantiate (as it is built upon the relevant indicator states) and that cannot be epiphenomenal in the sense just seen. Consider again the triplet (t_a,t_b,t_c), this time letting the three temperatures be ordered by their temporal relations (i.e. t_x *is followed after an amount of*

[140] Notice that this observation provides strong support in favor of the claim articulated in (§2.1.1), at least if effective control structures must be structurally similar to what they control, see (Ch. 2: §2.1).

time x by t_y). Again, let v_a, v_b and v_c be the corresponding states of the strip. Let them be ordered again in the triplet (v_a,v_c,v_b), this time by their temporal relations[141] (i.e. v_x *is followed after an amount of time x by* v_y). *Ex hypothesis*, the structural similarity is again absent. Yet, in this case, the system will malfunction. The reason is simple: if v_a *is followed after an amount of time x by* v_c and t_a *is followed after an amount of time x by* t_b, then the strip will occupy state v_c when the temperature is t_b. But the state of the strip indicating t_b is v_b, not v_c. Therefore, the receptor mis-indicates. As a consequence, the system will malfunction: its inner state will bring about the behavioral outcome appropriate to t_c instead of the one appropriate to t_b. Therefore the system is sensitive to (at least) the temporal relations holding among the elements of V; and the obtaining of such a time-dependent structural similarity between V and T determines the appropriate functioning of the system. Hence, at least this time-dependent structural similarity is not epiphenomenal. Notice that this structural similarity too obtains purely in virtue of indication, as indication is *time-dependent*: if V is a receptor of T, then V must occupy state v_a *when* T is in state t_a. In fact, each and every receptor *must* instantiate the kind of time-dependent structural similarity seen above, as an item failing to instantiate it cannot be a receptor. This can be shown by *reductio*.

Suppose V is a receptor of T. Suppose further no relation (not even temporal ones) can be found such that (i) to (iii) obtain in conjunction. *Ex hypothesis*, V and T are not structurally similar. But this entails that *when* the receptor is in a state v_a, the target can be in *any* arbitrary state t_x. To see why, consider the following scenario. Suppose that, at time *t*, the receptor is in a state v_a and the target is in a state t_a. Now, at time t^*, the receptor and the target change state: the receptor goes in state v_b and the target goes through a sequence of state changes $t_b...t_n$.[142] Suppose further that, at time t^{**}, the receptor returns in state v_a. Let us call *x* the amount of time

[141] Notice having the same kind of relations on both sides of the similarity is perfectly legitimate. Indeed, maps do represent spatial relations through spatial relations.

[142] This sequence might also include t_a. States can also repeat within the sequence. The point I'd like to make does not require these assumptions.

lapsed between *t* and t^{**}. It is thus correct to say that v_a was followed v_a after an amount of time *x*. Now, it is fairly easy to show that, *ex hypothesis*, at time t^{**} the target *must* be in any other arbitrary state t_x different from t_a. For, if it were in state t_a, it would be correct to say that t_a was followed by t_a after an amount of time *x*, which is enough to make the receptor and the target structurally similar.[143] But, by stipulation, V and T are not structurally similar. Thus, if a receptor and its target are not structurally similar, *when* the receptor is in a given state v_a, the target can be in any arbitrary state t_x.[144] But if when the receptor occupies state v_a the target can be in any state t_x, then the probability of finding the target in any individual state given that the receptor is in state v_a equals the probability of that state itself. Hence, it would be false that v_a indicates any state t_x of the target, as $P(t_x|v_a) = P(t_x)$. But since this line of reasoning holds for all the states of the receptor, it would then be false that V is a receptor of T. And this runs counter to the initial stipulations; namely, that V *is* a receptor of T.

In perhaps less convoluted terms, for any arbitrary receptor state v_a to indicate an arbitrary target state t_a it must be the case that, *when* the receptor occupies state v_a, it is more likely than otherwise that the target occupies state t_a. The same holds for all other receptor states $v_b...v_n$ and the corresponding target states $t_b...t_n$. As a consequence, if v_a is followed after an amount of time *x* by v_b, then it must be likely that t_a is followed after the same amount of time by t_b.

Notice that this line of reasoning is perfectly general, as it holds for all time-spans, receptor states and target states. Thus, it seems that the relevant time-dependent structural similarity holds purely in virtue of indication. Notice also the important corollary of this: every (action-guiding) receptor must, *qua* (action-guiding) receptor, exploit *at least* this time-dependent structural similarity with its target.

---

[143] To be sure, that would be a very *thin* structural similarity. Yet notice that the relevant definition of structural similarity Gładziejewski endorses quantifies only on "at least some", and it thus seems satisfied by what it is shown in my example. On the same issue, see also (Morgan 2014: 232).

[144] Notice that t_a is included, as it was (by stipulation) the state occupied by the target in the beginning of the example.

2.1.3 - Some receptors satisfy (c)

Notice, to begin, that the scope of the claim is now restricted to *some*. Thermostats, hygrometers and the like can indicate only in virtue of a constant causal contact holding between them and their target. A thermometer somehow shielded from the causal touch of the surrounding mean kinetic energy would simply stop indicating. I'm not denying this. I'm only denying that *all* receptors are thermometer-like in constantly needing the causal touch of their target to function.

As an example of a very simple receptor which does *not* constantly need the causal touch of its target to function and orchestrate the behavior of a system, consider the control system of a simple Braitenberg vehicle displaying a light-following behavior (Braitenberg 1984: 6-9, vehicle 2b). The control system of this robot is fairly rudimentary: it consists only in two laterally placed front-facing photoreceptors, each contralaterally connected to a motor by an excitatory link. When this simple agent faces a light source, two beams of light will impinge onto its photoreceptors, coupling the two. The receptors will thus excite the two motors, causing the robot to beeline towards the light source. But if the light source is located on one *side* of the vehicle, only one receptor will be coupled to it by a light beam. Thus only one wheel will turn, causing the robot to spin in place, re-orienting it towards the light source. Notice that albeit in this case only one receptor is coupled, the behavior is orchestrated by *both* receptors. Indeed, it is only *because* one receptor is not coupled to the light source that one wheel does not turn, allowing the robot to spin in place. This is a very minimal case in which a decoupled receptor is playing a key role in orchestrating the behavior of a system.

Now, as in the case above *one* receptor was still coupled to the light source, it might be objected that (c) is not actually satisfied. However, a minimal increase of complexity allows for a *weak* decoupling to doubtlessly obtain. A nice example is provided by DidaBots (Maris

and Schaad 1995; Maris and te Boekhorst 1996): simple robots tasked with clustering cubes in an arena. Their control architecture consists of four lateral proximity sensors connected to two lateral motors through both excitatory (ipsilateral) and inhibitory (contralateral) connections. Thus, when a receptor “sees" a cube, it speeds up the movements of the wheels on its side and slows down the speed of the wheels on the other side, causing the robot to turn away from the cube. Notice these robots are “blind” to the front, so if a robot and a cube are lined up, the robot will impact the cube, “picking it up” and pushing it along the way. When, while pushing a cube, the robot “sees" a cube on its side, it will turn away from it, “dropping” the cube it was pushing near the one it has sensed. This is how the robot cluster cubes. The important point to notice here is that the “picking up” and pushing of a cube is a behavior *governed by* decoupled receptors, as the robot can enact this behavioral routine only as long as *all* its sensors are not coupled to any cube. Were one of them coupled to a cube, the robot would immediately turn away from it, dropping the cube it was pushing as a result. So the “picking up a cube” behavioral routine is, in Gładziejewski's terminology, orchestrated by weakly[145] decoupled receptors.

It might be argued that the case presented above is *still* not sufficient to show that receptors satisfy (c). This is because representations (structural or otherwise) are supposed to provide the means for *endogenous and proactive* control (e.g. Gładziejewski and Miłkowski 2017; Pezzulo 2008; 2017). But receptors merely *react* to the presence of some environmentally delivered magnitude, or lack thereof. Hence, receptors do not satisfy condition (c), at least, not in the way in which genuine representations supposedly satisfy it.[146]

[145] Notice strong decouplability fails to obtain: the whole robot is coupled to the cube it's pushing.

[146] Notice, however, that this line of objection, pursued in (Gładziejewski and Miłkowski 2017), factually changes the relevant notion of decouplability mentioned when unpacking (c). In fact, Gładziejewski (2015b: 77) defines decouplability in purely causal terms; namely, as the lack of a causal connection between the representational vehicle (or the whole system) and the represented target. Hence, the original definition of “decouplability” at play in Gładziejewski’s account of structural representations has no *essential* tie to the idea that representations are a proactive locus of endogenous control. Artiga (2021) makes an analogous point.

However, this objection fails too, as some receptors *can* be the endogenous causes of proactive behaviors directed to targets from which the whole system is strongly decoupled. The recurrent artificial neural networks Harvey and colleagues "evolved" as control systems for robotic agents provides a nice example (Harvey *et al.* 1997). One such agent was tasked with visually tracking a moving target (Harvey, Husbands and Cliff 1994). Since the target was moving and the robot was not placed in front of it at every trial, there were significant spans of time in which no robot-target coupling obtained, and thus significant spans of time in which the two were *strongly* decoupled. In such cases, the robot self-initiated an exploratory behavior (namely, spinning in place to detect the target). This behavior was produced by a generator unit of the net (Husbands, Harvey and Cliff 1995): an artificial neuron able to "recycle" its output at time t as input at time t_{+1} through a self-recurrent connection. Since the network was noisy, generator units were able, by constantly feeding themselves back their noisy output, to generate significant activity within the net in absence of any environmental input. In the case at hand, the generator unit was a tactile receptor selected (by genetic algorithms) to trigger the "look around" behavioral routine in absence of any relevant external input. Notice the "look around" routine is caused by the *intrinsic* (noisy) dynamics of one receptor in the net. In other terms, the causal starting point of that behavior is within one of the net's receptors, not in the environmental input or lack thereof. Hence, a simple receptor was able both to coordinate a system's behavior regarding a strongly decoupled target and to do so by endogenously initiating the causal chain leading to the relevant behavior of the system. In conclusion, it appears that at least *some* receptors can satisfy (c), even according to this revised, and more demanding, notion of decouplability.

2.1.4 - Some receptors satisfy (d)

Some receptors can generate system detectable errors. As a nice example, consider the

control architecture for robotic agent Bovet (2007) engineered. The architecture consists in a series of homogeneously connected feedforward artificial neural networks, one for each sensory or motor modality of the robot.[147] Simplifying a bit[148], each net consists of three identical populations of simple neuron-like receptors. Two of these populations jointly form the input layer, and the other is the output layer. Each net works as follows.[149] The *current state population* receives input from the sensors of the modality controlled by the net, entering in the state corresponding to the incoming sensory barrage. The *desired state population* receives input from the nets of all other modalities, thus entering in the state the controlled modality *should* occupy, given the activity of the rest of the system. For instance, if the visual desired state population receives the signal that the robot is moving forward, it will enter in the state corresponding to an optic flow expansion, as moving forward typically correlates with optic flow expansion. Together, the current state population and the desired state population constitute the input layer. The output layer consists of the *desired state change* population, responding to the *difference* between the states of the two halves of the input layer, and spreading that difference to the rest of the system. So the receptors of the output layer respond to the *mismatch* between "desired" and received sensory input, which is a very simple form of *prediction error*.[150]

Notice these "error receptors" are as causally potent as any other receptor in the system. In fact, the activity of the motors is determined (through the motor desired state population) by the output layer of each modality, which spreads the *mismatch* between the two halves of the

---

[147] Notice these nets lack both self-recurrent connections and hidden units: the typical resources that are considered representational vehicles in connectionist systems (e.g. Shea 2007; Shagrir 2012). Their activity is thus interpretable in a straightforwardly non-representational manner (Ramsey 1997).

[148] The architecture will be the focus of the next chapter, so I will introduce the relevant details there, when they are actually needed.

[149] After the learning period, in which the net learns the robot's sensorimotor contingencies (see O'Regan & Noë 2001): the ways in which stimulation changes as a consequence of movement.

[150] Technically, the architecture behaves as if it were detecting the mismatch between the received inputs and the ones self-generated by a forward model (see Bovet 2007, pp. 79-106). This mismatch is ordinarily considered as prediction error in the predictive processing literature, and Gładziejewski (2015b; 2016) himself relies on this very same notion of error.

input layer. This means the motors are active *only if* there is at least one net spreading error. So error is what, causally speaking, drives the system around. Moreover, in a series of experiments (Bovet and Pfeifer 2005a; 2005b, see also Pfeifer and Bongard 2007, pp. 295-333) the robot learned to solve a simple working memory task (i.e. finding the reward at one end of a T-maze) by learning to trust a tactile-motor correlation (it learned to "expect" to turn in the direction of the active tactile receptor sensing the cue) over a visuomotor one. This shows that the robot can, implicitly, assess which error is important to minimize and which error is irrelevant.

Importantly, in these experiments, the receptors of the net satisfied (a) to (d) *jointly*. If the arguments provided thus far are sound, (a) and (b) must obtain, as they obtain for every receptor, and the net is just a series of receptors systematically connected. (d) obtains, as the system has a specialized set of receptors in the task of detecting the error between "expected" and actually occupied sensory states. Lastly, (c) obtains too, as, at the onset of each trial, the robot was *strongly decoupled* from the reward it had to find. Indeed, at the onset of each trial the robot and the reward are in different "arms" of the T-maze, and no causal chain connects the two. Moreover, the robot exploration of the maze was self generated, as it was due to an inbuilt discrepancy in the two halves of the input layer for the "reward" modality (i.e. battery level).

2.2 - Moving towards the second step

As the example provided above demonstrates, in appropriately complex systems, receptors satisfy (a) to (d) in conjunction. Given that receptors *paradigmatically* fail the Job Description Challenge, it seems that structural representations, as Gładziejewski defines them, fail it too.

Or do they? After all, one could simply object that all that I've shown is that there are receptors that *meet* the Job Description Challenge, namely the receptors that jointly meet (a) to

(d).[151] Perhaps one could say that receptors that do not satisfy (c) and (d) actually function merely as causal mediators, but those which *do* satisfy (c) and (d) are endowed with a genuine representational status. Or perhaps one could say that I've only shown that some structures that *prima facie* qualify as receptors actually are, upon closer scrutiny, structural representations and thus meet the Job Description Challenge. This would be in line with the conclusions of (Morgan 2014; Nishberg and Shapiro 2020).

Moreover, notice how, in moving from (a) to (d), the receptors I considered while articulating my claims were embedded in progressively more complex structures. While defending the claim that receptors satisfy (a) and (b), I considered very simple receptors, such as thermometers. Yet, to defend the claim that *some* receptors satisfy (c), I had to introduce full-blown artificial *agents*, such as Braitenberg vehicles and DidaBots. And to defend the claim that receptors *can* allow proactive behaviors and allow for error detection (point (d)), I had to introduce robots governed by neural networks that are way more sophisticated than thermometers and Braitenberg vehicles. Yet, agents governed by such networks could qualify as representation-users (at least, intuitively). This observation provides further support to the claim that receptors satisfying (a) to (d) actually are representations of some sort, perhaps even structural representation.

I wish to resist these conclusions. In the next block, I will put forth an argument by analogy to intuitively show that (a) to (d) do not spell out a representational functional profile, in the style of both Ramsey's (2003; 2007) original analysis or receptors and Gładziejewski's (2015b; 2016) treatment of structural representations. The argument will also show that receptors satisfying (a) to (d) are in no way specific to artificial agents, and thus that they can be found in physical structures that, *at least intuitively*, do not qualify as representation-users.

[151] I owe this observation, and its brilliant framing, to an anonymous reviewer of the journal *Synthese*.

2.3 - Compare-to-prototype: structural representations and capacitors

Consider an optical smoke detector: a simple device tasked with ringing an alarm when it detects a fire. Fires generate smoke, and, as smoke fills the air, it fills the inner chamber of the detector, refracting a beam of light on a photosensitive surface. This, in turn, closes a switch supplying electric power to an alarm. This is a simple, receptor-based, non-representational device.

Suppose one such device operates in an environment in which the typical combustion also generates *heavy smokes*: toxic fumes that tend *not* to rise even when heated, and that linger in the environment even *after* the fire has been put off. Suppose we want to enable the device to signal us their presence. It has to keep the alarm ringing when heavy smokes linger in the environment, putting it off when the heavy smokes have been dispersed by the ventilation system. This poses a challenge: heavy smokes tend (being *heavy*) to linger on the *floor*. But the optic smoke detectors are mounted on *ceilings*: "normal" smoke *rises* when heated. So the system, as it stands, is incapable of indicating the presence of heavy smokes, as they will not deflect the light beam. Indeed, the two are in no obvious causal contact.

Placing a *capacitor* between the switch and the alarm enables the system to indicate the presence of heavy smokes. When the system detects a fire, it closes the switch feeding energy to the alarm. If a capacitor is placed between the two, it will store some energy when the circuit is closed, slowly releasing it when the circuit opens (i.e. when the fire has been put off). So it will keep the alarm ringing when there is no fire but heavy smokes still linger.

Strikingly, the capacitor will *function as a receptor* for heavy smokes. This is because the amount of energy stored by the capacitor depends upon the time the circuit has been closed, which, in turn, depends on the time the fire has been raging. But so does the amount of heavy smokes. The longer the fire, the more the material combusted, and the more the material

combusted, the more the heavy smokes produced. Thus, due to a common cause[152], the states of the capacitor actually indicate the amount of heavy smokes present in the environment. Observing the capacitor having in store an amount of energy v_x rises the probability that a corresponding amount of heavy smokes t_x is present in the environment.

Notice also that the capacitor satisfies (a) to (c). If the arguments given above are correct, there is at least one non-epiphenomenal structural similarity holding between it and the heavy smokes, ensuring that (a) and (b) obtain. Namely, the chronologically ordered sequence of capacitor states (v_a,v_b,...v_n) must map onto the chronologically ordered amounts of heavy smokes (t_a,t_b,...t_n). Otherwise, the system malfunctions: it either shuts up the alarm too soon (failing to indicate the presence of heavy smokes) or too late (indicating the presence of non-existing heavy smokes). Moreover, the whole system is not in any causal contact with heavy smokes, so the capacitor is *strongly decoupled* from them. Indeed, this is the reason why the capacitor is needed.

A slight modification of the system enables the capacitor to satisfy (d). Suppose a second switch is placed after the capacitor, and let it be *closed* by default. Suppose further the first switch also feeds energy to a mechanical timer running a countdown. When the countdown reaches 0, the timer opens the second switch, putting off the alarm. Lastly, let the circuit supplying energy to the timer be controlled by a bi-metallic strip, whose expansion opens the circuit, stopping the countdown.[153] Collectively, these components will act as a *control mechanism* for the device. Their functioning principle is simple: if, in a set amount of time, no significant increase in temperature is detected (i.e. the bi-metallic strip does not expand), then there likely is *no fire*. So, the photosensitive cell misdetected a fire, leading the capacitor to "hallucinate" heavy smokes. The system corrects the error of its receptors opening the second

[152] Notice that this is just a "ghost channel" in the sense of Dretske (1981, pp. 38-39): a set of statistically salient dependency relations between the state of two systems that are not in causal contact.

[153] Notice that in thermostats bi-metallic strips are used as switches in the same way.

switch, putting the alarm off. However, if a high temperature is detected (i.e. the bi-metallic strip expands), then there likely is a fire. So, the photosensitive surface and the capacitor are working properly and the timer is stopped to keep the alarm ringing.

In this modified system, the capacitor satisfies (a) to (d), and thus, according to Gładziejewski, has the functional profile of a structural representation. But capacitors surely are *mere* causal mediators, and even in this (fairly complex) toy system the capacitor functions simply *as a battery* to keep the alarm ringing. It thus seems that bearing features (a) to (d) is not *sufficient* for an item to function as a representation. Hence (a) to (d) do not spell out a robustly representational functional profile. As a consequence, if structural representations are defined in terms of items bearing features (a) to (d), structural representations do not meet the Job Description Challenge. Indeed, it seems to me that the same sort of worries that motivated either the rejection of the receptor notion of representation (e.g. Ramsey 2007; Orlandi 2014) or a strong suspicion about its explanatory potential (Williams and Colling 2017) emerge again. If our *most demanding* account of structural representations identifies simple capacitors as representations, how could panrepresentationalism be avoided? How does such a notion of representation capture a distinctive psychological or cognitive phenomenon? Is the proposed notion of representation doing valuable explanatory work? Surely my toy system's functioning can be entirely *and transparently* understood without invoking representations. If these are reasons to reject, or be skeptical of, *receptors*, they will also be reasons to reject, or be skeptical of, structural representations. As Nishberg and Shapiro (2020, p.2) nicely put it, structural representations and receptors have a common fate.

3 - Facing some objections

Here, I defend my claim by a number of foreseeable objections.

3.1 - Receptors and exploitability: a counterexample

An anonymous reviewer of the journal *Synthese* greeted my argument to the effect that all receptors satisfy (b) with a counterexample and a challenge. In this sub-paragraph I deal with the counterexample. The challenge will be met in the next sub-paragraph.

The counterexample is as follows: consider litmus papers; that is, stripes of chemically treated paper that change color when immersed in chemical substances, thereby indicating the pH of the substance. Suppose I use one such device to measure the pH of a substance at time t. At t^*, I extract the paper from the substance, which I then dilute with water. The substance's pH has changed, but the color of the paper has not. Yet, it is still correct to treat the paper as a receptor representing the substance's pH. Isn't this a proof that the time-dependent structural similarity discussed above does not hold universally for receptors?

I concede that the litmus paper at t^* is still indicating. In fact, I would add that it is mis-indicating[154], as its color does *not* match the substance's pH. Notice however, that such a mis-indication occurs *at time* t^*, and only because the litmus paper has not changed color as the relevant time-dependent structural similarity prescribes. As long as misindication occurs, the time-dependent structural similarity is broken. But suppose now that, at a further time t^{**}, the litmus paper is put in contact again with the substance. It *would* change color, and it *would* correctly indicate the substance pH. Let x be the amount of time lapsed between t and t^{**}. The substance pH at t is thus followed, after an amount of time x, by the substance pH at t^{**}. But the same relation holds for the states of the litmus paper: color at t is followed, after an amount of time x, by color at t^{**}. Hence the time-dependent structural similarity is restored. Of course, the time-dependent structural similarity instantiated by the litmus paper has, in this example, proven insensitive to the change of state of the substance at t^*. But similarity is a *graded* notion;

[154] The anonymous reviewer also suggested that in such a case the litmus paper would count as a decoupled receptor of pH-in-the-past. I think I disagree. It seems to me that litmus papers have, by design, the function of indicating the pH of substances *in the present*.

and even uncontroversial cases of structural representation are manifestly not *perfectly* structurally similar to their targets (Cummins 1996; Williams and Colling 2017 p. 1947; Gładziejewski and Miłkowski 2017). A map perfectly (e.g. millimeter by millimeter) similar to the depicted terrain would be useless. Hence the example fails to pose any substantial challenge to my claim regarding (b).

The same line of reasoning, it seems to me, applies to all counterexamples with a similar structure. Local failures of indication (i.e. short spans of time in which a receptor fails to covary with its target) are as admissible as local failures of the relevant structural similarity. Hence, noticing that there are, as a matter of fact, short time spans in which a receptor fails to indicate a target does not, in and by itself, challenge the claim that indication is a special case of structural similarity.

Notice, further, that the failure in the relevant receptor-target structural similarity the example highlights seems to depend directly on the fact that receptor and target are *decoupled*. In fact, it seems entirely correct to say that if the litmus paper had not been extracted from the chemical substance at *t**, it *would* have changed color so as to match the substance pH. But if this is correct, then that case fails to constitute a compelling case against my argument, as I'm not committed to the claim that *all* receptors can function when decoupled from their targets; I'm only committed to the claim that *some* can. As noticed in (§2.1.3), if a thermometer were (somehow) shielded from the mean kinetic energy of the surrounding environment, *it would simply stop indicating*.

3.2 - Dealing with cases of apparent unexploited structural similarity: Shea's example

Now, the challenge. Shea (2018, p. 119) illustrates a non-exploited structural similarity with the following example: suppose that a pack of vervets has three kinds of predators p_1, p_2 and p_3. Suppose that the vervets have three types of alarm calls c_1, c_2 and c_3, one for each predator.

Suppose that p_1, is taller than p_2; which is in turn taller than p_3. Suppose further that the same ordering holds for the calls: c_1 has a higher pitch than c_2 which in turn has a higher pitch than c_3. The system of calls is thus structurally similar to the system of vervet's predators. However Shea argues that vervets are not sensitive to the relation "*higher pitch than*" holding between their calls. All vervets do, in Shea's view, is to respond separately to each individual call. Thus, Shea concludes that the structural similarity holding between vervet calls and predators is not exploited. Importantly, Nishberg and Shapiro (2020, p. 16) concede Shea the point that, *taken as an array*, the system of calls is not a SR of the heights of predators.[155] The reviewer asked whether the time-dependent structural similarity I'm discussing contradicts Shea's verdict, showing that an exploitable structural similarity holds between the system calls and the predators.

I believe that, in this regard, it is important not to conflate two distinct issues. The first is whether it is *necessary* that a structural similarity holds between an array of receptors and the ensemble of their targets. To this question, I, together with Nirshberg and Shapiro (and presumably Shea), answer negatively. The hygrometer measuring the humidity of room A is structurally similar to the humidity in room A, and the thermometer measuring the temperature in room B is structurally similar to the temperature in room B. However, the *thermometer plus hygrometer* system need not be (albeit it might[156]) structurally similar to anything. An array of structural representations need not be a structural representation on its own. Notice that the same thing holds for *uncontroversial* instances of structural representation too. I can place my map of Sydney *north of* my map of Rome without thereby generating a new structural representation that misrepresents the relative positions of Sydney and Rome.

---

[155] Albeit they hold that each call is structurally similar to one predator (see Nirshberg and Shapiro 2020, p.16).

[156] Morgan (2019:8-9) presents a case in which an array of receptors (each structurally resembling a target) is a structural representation in its own right. In his example, each receptor is a bucket of water, the volume of which is proportional to the distance a boat has travelled in one direction. The array of water volumes is thus structurally similar to the cartesian coordinates of the boat (i.e. the fact that bucket 1 is fuller than bucket 2 maps onto the fact that the value of the coordinate of the Y axis is higher than the one on the X axis).

The second issue regards whether that system of calls actually is structurally similar to something (and whether such a structural similarity is exploited). And I believe the time-dependent structural similarity I introduced actually allows for a positive answer to both questions. For the alarm calls to be effective, these must be tokened in a way such that the temporal ordering between calls matches the one holding between the apparences of predators. Thus, if the three predators appear in the temporally ordered sequence (p_1, p_2, p_3), the alarm calls need to be uttered in the corresponding temporally ordered sequence (c_1, c_2, c_3). Changes in this sequence result, at least *prima facie*, in dead vervets. Hence, the system relying on these calls to orchestrate its behavior (i.e. the pack of vervets) seems sensitive to at least these relations.[157]

3.3 -Do "compare-to-prototype" arguments sidestep the Job Description Challenge?

The same reviewer also raised a more general concern. The concern is that my treatment has simply *sidestepped* the Job Description Challenge. This concern articulates in two distinct worries. The first concerns the call to intuitions embedded in "compare-to-prototype" arguments. These are arguments by analogy, and thus rely heavily on our pretheoretical intuitions. But surely not all pretheoretical intuitions are correct. The second is that not enough care has been taken in discussing whether truth/accuracy conditions are causally relevant in accounting for a system's success. If they are, then the Job Description Challenge is met (the reviewer also pointed out that this is the argumentative strategy of Gładziejewski and Miłkowski 2017).

Let me begin by addressing the first worry. As things stand, it seems to me that calls to intuition are licensed as valid moves to address the Job Description Challenge (see Ramsey

[157] Notice also that, at least in this case, single calls afford the detection of representational error. It is in fact suggested that repeated mistokening of these calls might cause the "liar" vervet to be ignored by the pack (e.g. Cheney and Seyfarth 1985, p. 160).

2007: 10-11). Indeed, one of the significant aspects of the challenge is that of checking whether the term "representation", as it is used by cognitive scientists, is sufficiently "in touch" with its everyday usage. Moreover, arguments by analogy seem *sufficient* to face the challenge. This is the case, for instance, of Ramsey (2007: 83-89) and Gładziejewski (2015b; 2016). Hence, if these arguments by analogy are sufficient to face the Job Description Challenge, mine should be too. Surely, one can deny that these arguments are sufficient to face the challenge, perhaps because they rely too much on intuition.[158] However, determining the role intuitions should play in philosophical theorizing lies significantly outside the scope of the present chapter.

What, then, about the second worry? Is checking whether the truth or accuracy conditions of a posit are causally relevant to a system's success *sufficient* to determine whether the posit meets the Job Description Challenge? I doubt this is the case. To see why, consider the following two cases.

First, the firing pin of a gun. As highlighted above, it indicates the position of the trigger, and has (by design) the *function* of doing so (firing pins are included in guns precisely *because* their state indicates the state of the trigger). Under mild teleo-informational commitment, this is *sufficient* to yield accuracy conditions to the firing pin: the firing pin accurately represents the position of the trigger if, and only if, it occupies the position it *should* occupy, given the state of the trigger. It is now possible to follow Gładziejewski and Miłkowski (2017) and wonder whether intervening on the degree to which these accuracy conditions obtain causally influences the success of the gun. And this is surely the case. The less the position of the firing pin corresponds to the position of the trigger, the more *unreliable* the gun is. In fact, the less the positions of the trigger and the pin correspond, the more the gun will fire at random. So, the accuracy conditions of the firing pin are causally relevant to the successful functioning of

[158] Notice also that such a move would undermine the claim that SRs meet the Job Description Challenge. In fact, to the best of my knowledge, that claim has only been supported by means of arguments by analogy.

the gun, but I (and, I think, many others) would be hard pressed to conclude *on this sole basis* that guns are representational systems.

Consider now false, but *useful*, beliefs. The research on optimism bias, for instance: "Highlights the possibility that the mind has evolved learning mechanisms to mis-predict future occurrences, as in some cases they lead to better outcomes than do unbiased beliefs" (Sharot 2011: R495). It is also said that the lack of such an optimism bias negatively correlates with mental health (Taylor 1989; Sharot 2011). It thus seems that certain beliefs lead a system to its success *because* they are false or inaccurate. However, it is commonly assumed that only *correct* representations non-accidentally lead to a system's success (e.g. Shea 2018: 10). Thus, when checking whether the conditions of satisfaction of a posit lead to a system's success, one checks whether *correct* representations lead to successful behavior. But this is not the case for optimistically biased beliefs. So our verdict, in this case, should be negative: these beliefs do not meet the Job Description Challenge and thus are *not* representations. However, optimistically biased beliefs are *beliefs* (in the ordinary sense of the term), and thus surely qualify as representations.

One could perhaps argue that this is too fast, as optimistically biased beliefs serve a psychological function other than representing reality (e.g. a motivational function). Whilst sympathetic with this line of objection, I can help but notice that, even if true, it wouldn't substantially damage the conclusion I'm trying to establish. In fact, it would still be true that the "accuracy condition" of some clearly non-representational items (such as firing pins of guns) are still crucial to the success of the system in which they operate. Hence, checking whether the conditions of satisfaction of one posit are causally relevant in explaining a system's success is not *sufficient* (albeit it surely is necessary) to meet the Job Description Challenge.

3.4 - Do receptors fail the Job Description Challenge?

The argument I have presented here hinges on a crucial premise; namely, that receptors do not meet the Job Description Challenge, and thus that they are not representations. But what if they were? Clearly, arguing effectively that receptors are representations would invalidate my argument.

As far as I can see, there are, in the current literature, only two explicit defenses of the representational status of receptors. I discuss each in a separate subsection.

3.4.1 - Rupert's argument

The first is due to Robert Rupert (2018). I will not discuss Rupert's argument in detail, as in the present context, it suffers from a major problem that makes it *unable* to block my argument.

To put it bluntly, the problem is the following: if Rupert's account were correct, then predictive processing would be a non-representational theory of cognition *by definition*.

To see why this is the case, consider consider that according to Rupert's (2018) account something (receptors included) qualify as a representation if, and only if, it bears some content *and satisfies the following additional conditions* (Rupert 2018: 205):

> 1. It appears in an architecture which produces the distinctive *explananda* of cognitive science (i.e. intelligent behavior); &
>
> 2. Its contribution to the functioning of these architectures rests on its representational capacities; &
>
> 3. Its playing an explanatory role as a representation depends on the presence, within the architecture, of distinctively cognitive forms of processing

Clearly, the acceptance of these conditions is *sufficient* to invalidate Ramsey's (2003; 2007) arguments by analogy. Since firing pins of guns (and the like) do not satisfy (1) to (3), they are not even *putative* representations, and every analogy between them and genuine representations is fallacious (Rupert 2018: 213) So, Rupert's argument can (in principle) "rescue" receptors,

endowing them a representational status.

However, the philosopher interested in defending the claim that predictive processing is a representationalist theory of cognition simply *cannot* resort to Rupert's defense of receptors to counter the argument I have offered here. This is because of condition (3).

Now, I must confess that I'm not sure about what counts as a "distinctively cognitive form of processing", and Rupert is not very clear on the matter. However, he (Rupert 2018: 210) suggests that only forms of processing found only in cognitive architectures count as distinctively cognitive. And, if that is the case, then predictive processing is, *by definition*, a non representational theory of cognition. This is because its core processing algorithm (i.e. predictive coding) is *not* found only in cognitive architectures. To the contrary, predictive coding is a data compression strategy which is routinely deployed by non-cognitive architectures to perform non-cognitive tasks. And, in fact, predictive coding originated only as a data compression strategy which simplified the transmission of images (see Shi and Sun 2008: ch. 3; Spratling 2015; 2017). As such, predictive coding is not a "distinctively cognitive form of processing". Hence, if Rupert is correct, the (putative) representations involved in predictive processing would not satisfy condition 3, failing to qualify as representations as a consequence. Thus, if Rupert's (2018) proposal is on the right track, then predictive processing is *by definition* a non-representationalist theory of cognition. Hence, philosophers interested in defending the representational credential of predictive processing cannot resort to Rupert's defense of receptors to block my argument.

3.4.2 - Artiga's argument

Mark Artiga (2021) presents two arguments to the effect that receptors meet the Job Description Challenge. The first is that receptors are nothing over and above structural representations, and since structural representations meet the Job Description Challenge,

receptors meet it too (Artiga 2021:13).

To start, notice that one *cannot* use Artiga's argument as a way to reply to the claim I've been articulating here, for it would simply beg the question. This is because the argument assumes as a premise that structural representations meet the Job Description Challenge - that is, it assumes as a premise that my conclusion is wrong. And this clearly makes Artiga's first argument unavailable when it comes to refuting the argument I've articulated here: one cannot refute an argument by *assuming* that the argument is wrong.

Moreover, I think that Artiga's first argument has very little bite, at least, if "receptors" means "all receptors" and structural representations are assumed to have the functional profile Gładziejewski proposes. In fact, as explicitly noticed in (§ 2.1.3) and reminded in (§ 3.1), although *some* receptors satisfy (a) to (d), not *all* receptors do: some fail to satisfy (c) and (d). And although Artiga's first argument could be rescued by erasing point (c) and (d) from the functional profile of structural representations, doing so would plunge us knee-deep in panrepresentationalism, as (Gładziejewski 2016) acknowledges.

Artiga's (2021: 14-15) second argument defends the claim that receptors meet the Job Description Challenge because receptors are input-output representation, which meet the challenge. In short, the idea is that in order to understand the distally characterized behavior of a (computational) system deploying receptors to orchestrate its behaviors, we must see the system's receptors as representing the environmental states of affairs relevant to the system's behavior.

In spite of the argument's simplicity, it is important to notice at least three distinct things.

The *first* thing to notice is that the argument *can, but it is not guaranteed to*, avoid the charge of panrepresentationalism that ensues from treating receptors as representations. This is because input-output representations are tokened only in computational systems, and *prima facie*, not all systems are computational systems. The "*prima facie*" qualifier, however, is

important, because the number of systems counting as computational systems depends on one's theory of computational implementation. If one's theory of computational implementation licenses a form of pancomputationalism (i.e. the claim that all physical systems compute at least some functions), then the charge of panrepresentationalism is not avoided: since all system compute, all systems relying on receptors token input-output representations, guns and optical smoke detector included. Now, Artiga does not say which theory of computational implementation he endorses, and in this dissertation I've committed myself to understand "computation" as "generic computation" (see Ch. 2: §1), without espousing any particular theory of computational implementation. So, I do not press the point further, and assume that Artiga's account of computational implementation does not license pancomputationalism, keeping his account safe from panrepresentationalism.

The *second* thing to notice is that it is manifestly *false* that in order to understand how receptors enable distally characterized behaviors we *must* regard them as representations of the distal states of affairs that are relevant for said behavior. We are surely able to understand how a thermostat *keeps the temperature in a room constant* or how an optical smoke detector *alerts us of the presence of a fire* without having to deploy a representational lexicon. The thermostat keeps the temperature constant because when the temperature exceeds certain thresholds the bimetallic strip is too curved to keep the circuit feeding power to the furnace closed, thereby putting the furnace off. The bimetallic strip alerts us of the presence of fires because the smoke the fires generate scatters the beam in the detector's reflective chamber, thereby triggering the alarm. These are perfectly good and intelligible explanations of the distally characterized behaviors of systems deploying receptors that do not cast receptors as representations, and treat them as mere causal mediators. Notice that similar explanations can be offered for the behavior of systems that are far more complex than thermostats and optical smoke detectors. Indeed, in (§§ 2.1.3 and 2.1.4) I've briefly explained how robotic agents can rely on their receptors to

enact distally characterized behaviors and achieve distally characterized goals without using the word "representation" once (the next chapter will provide a further example). This is not to deny that we *can* regard receptors as representations of distally characterized states of affairs, for everything can be regarded in that way: the fact that the trigger of the gun has been pulled *can* be regarded as representing my homicidal desires, but it must not be regarded that way, and typically isn't.

The *third and last* thing to notice is how making the representational role of receptors piggyback on their role as input-output representations falls short of vindicating their status as representations *usually understood*. For, given the framework developed in (Ch. 2), regarding receptors as input output representations only licenses the claim that they carry mathematical contents. But mathematical contents are not representational contents usually understood. Unlike standard representational contents, they are not determined by a privileged naturalistic relation holding between a vehicle type and the target of the vehicle type. Hence, if the representational status of receptors were to depend exclusively on them being input-output representations, they would still fall short from being representations *in the usual sense* of the term (although it would of course be sufficient to make them representations in a fairly specialized sense, picked up by the label "input-output" representation).

One might reply that mathematical content and content "in the usual sense of the term" are not mutually exclusive: many vehicles of input-output representations (i.e. computational states) also carry regular representational content. Whilst this is true, appealing to such an observation seems to me to undermine Artiga's argument. The observation conceptually teases apart mathematical contents and regular representational contents. But Artiga wants to conclude that receptors are representations (in the standard sense) *because* they function as input-output representations. The more input-output representations and regular representations are teased apart, the weaker his argument gets.

In conclusion, Artiga's arguments fail to provide compelling reasons to regard receptors as representations. I thus conclude that receptors fail to meet the Job Description Challenge.

3.5 - Changing the definition of structural similarity does not help

It might be possible to defuse the conclusion of my argument by changing the relevant definition of structural similarity mentioned in (a). Perhaps second-order structural resemblance is *too* cheap, and structural similarities might be better understood in terms of isomorphism or homomorphism (see Swoyer 1991; Shea 2018). As these are more *restrictive* than second order structural similarity, leveraging them might prevent receptors from meeting (a) or (b) or both. But this is not the case. In every example I proposed when discussing (a) and (b) an *isomorphism* obtained. Each and every relation (v_x,v_y) among the features of the vehicle corresponded to only one relation (t_x,t_y) among the features of the target *and vice versa.* So appealing to isomorphisms does not challenge my conclusion. As isomorphisms are a special class of homomorphism, appealing to them will not alter my claim either.

Another way in which condition (a) could be strengthened so as to defuse my argument is by placing some restriction on the relevant class of structural similarities apt to satisfy (a), such that the restriction would exclude, in a principled way, structural similarities instantiated *through time*. This would not counter my claim that all receptors satisfy (a); in fact, there are receptor-target structural similarities that are purely "static", and are not instantiated through time. It would, however, counter my claim that all receptors satisfy (b), as the relevant structural similarity each and every receptor exploits is a structural similarity instantiated through time. Doing so would effectively refute my claim.

However, I see two problems with this line of argument.

First, what *independent* reasons support the claim that the only relevant structural similarities are not time-dependent? I know of no such reason. And in fact, it is typically

claimed that the relevant structural similarities holding between neural structures (or model thereof) and their targets are realized dynamically through time (e.g. Grush 2008; Garzón and Rodríguez 2009; Shagrir 2012; Morgan 2014; Shea 2018).

Secondly, and relatedly, such a stipulation would *hinder* the empirical adequacy of structural representations. If the relevant structural similarities holding between computational models and their targets are time-dependent, ruling out that time-dependent structural similarities are "real" or "genuine" structural similarities would imply that many paradigmatic cases of structural representations are actually *not structural representations* at all.

And, in fact, adding further restrictions to Gładziejewski's account does not strike me as a promising way to deflect my claim.

3.6 - Could adding a fifth condition rescue structural representations?

Another way to block my claim by making the relevant definition of structural representation more demanding is by adding a *fifth condition* to Gładziejewski's account. That might be sufficient to differentiate structural representations from receptors, blocking the argument here presented.

However, I believe that adding a fifth condition to Gładziejewski's account will do no good to the philosopher interested in defending the representational credentials of structural representations (and/or predictive processing). This is because of two distinct reasons.

First, there is no *obvious* candidate for the fifth condition. Neither I nor the audience of the conferences in which this essay has been presented managed to find a plausible candidate. This surely does not prove that a fifth condition does not exist. But it suggests that such a fifth condition is far from obvious and hard to find.

Secondly, Gładziejewski's account is already very demanding. Condition (c) is arguably not *necessary* for representations (e.g. Chemero 2009: 50-65; Miłkowski 2017) and the

dispensability of condition (d) has already been suggested (Lee 2018). Gładziejewski (2015b) himself acknowledges this, and takes the demandingness of his own account to be a virtue, as it shields his account from many trivializing counterexamples (see Miłkowski 2013: 160-161 for a brief, but insightful, case). But virtues, taken to the extreme, might easily turn into vices: in particular, it seems to me that adding a fifth condition to Gładziejewski‘s account would make it so demanding that few, if any, structures will satisfy the account. That is, adding a fifth condition to Gładziejewski‘s account *might* easily make it *too demanding* to be satisfied.

Of course, the observations above do in no way *rule out* the possibility that adding a fifth condition might be sufficient to defend Gładziejewski‘s account of structural representations from the argument I’ve offered here. It is likely that any candidate fifth condition should be individually evaluated for how it impacts and modifies Gładziejewski‘s original four conditions. But doing so clearly presupposes the presence of some candidate fifth condition, and, as far as I can see, no such candidate has been proposed (yet).

3.7 - Does adopting a minimalist attitude toward representations help?

If strengthening Gładziejewski’s account by adding further conditions to it is an unpromising way to deflect my argument, then perhaps doing *the opposite* might help the representationalists’ cause. Maybe, then, one could argue that Gładziejewski is *wrong* in taking his own account to be a very demanding one, and stress instead that it is pretty *undemanding*. This provides a way to at least “tame” my argument: in fact, if Gładziejewski’s account of structural representations is undemanding, then no doubt very simple entities, such as the capacitors of somewhat complex systems, will satisfy it. But that is no problem, for, since the account is undemanding, structural representations *can* (and should be expected to) turn out pretty minimal. So, if Gładziejewski’s account really is undemanding, my capacitor-based example fails to provide a compelling counterexample to the account. If an argument is

designed to over-generalize, the fact it over-generalizes is no objection to the argument.

However, when I added the control system, so as to enable the capacitor to meet condition (d), the *functioning* of the capacitor was not modified by the addition of the control circuit. The control circuit that I added in the final version of the system enabled the whole system to "figure out" the instances in which fires and heavy smokes where "hallucinated"[159], without thereby modifying the functioning of the capacitor. The capacitor itself functions as it functions in the version of the system that has no control circuit, and that is thus unable to meet condition (d). And, in Gładziejewski's own view, *that* way of functioning is not representational. So, it seems correct to say that the capacitor in my example *does not* function as a representation, not even if one takes Gładziejewski's account to provide a very minimal definition of structural representations. In other words, even if Gładziejewski's argument *were* designed to overgeneralize, it *would* be correct to say it overgeneralizes too much: it identifies as representations things with a functional profile *which is not representational even according to the account*.

Moreover, I must confess that it is not clear to me *on what grounds* one might hold that adding the control circuit would transform the capacitor into a structural representation. The addition of the control circuit does not modify the way in which the capacitor functions, nor its overall role within the system. If the way in which the capacitor functions when the control circuit is absent is non representational (and, on Gładziejewski's account, that is true), why, then, the addition of the control circuit, which *does not* modify the way in which the capacitor operates in the system, makes its functioning representational? Surely, we can *stipulate* that it does, but why should we? Gładziejewski (2015b, pp. 78-79) simply asks us to accept condition (d) without offering any substantial justification for it. And the reasons as for why Bickhard (1999; 2009) deems error detection *necessary* for genuine representations seem to be fairly

[159] Or, in more mundane terms, the cases in which the system malfunctions.

alien to the theoretical commitments of cognitive science. For instance, Bickhard greatly stresses the fact that, in order for some internal state to count as a representation, it must be a representation *for the organism* "consuming" it. But such a requirement is by no means necessary in the theoretical framework of cognitive science; indeed, many paradigmatic examples of representations (e.g. syntactic trees, Marr's 2 ½-D sketches) are not representations for the organism consuming them. And, in fact, cognitive scientists do not simply introspect them or somehow intuit their presence: they *posit* them as explanatory tools deemed necessary to account for the functioning of our cognitive system and the production of intelligent behavior.[160] Now, I do not wish to *simply* rule out (d) as a necessary condition. Perhaps it is. But if it is, then there must be a way to spell out why error detection is necessary. As far as I can see, this reason has not yet been spelled out.

3.8 - A reductio of the Job Description Challenge?

One might also object that my argument is a *reductio* of the Job Description Challenge.[161] The reasoning behind this objection seems to be as follows. Any successful naturalistic account of representation *should* cast representations (more precisely, their vehicles) as causal mediators, whose causal role is systematically related to their semantic properties.[162] Now, it is widely assumed that, in the case of structural representations, the relevant semantic

[160] To be fair to Bickhard, it is important to point out that the idea that genuine representations are representations for whole organisms is not the sole reason as for why he deems error-detection a necessary condition. The prospect of avoiding the problems of content indeterminacy seems to play an important role too. Yet, I do not see how requiring error detection helps in this regard: in order for the tokening of a representation to be counted as a representational error, the representational content must have already been determined: I do not see what *else* could justify considering that specific tokening as an error. Thus, it seems that content determination must logically preceed error detection: only *once* a determinate content certain tokens of a vehicle can be rightfully counted as misrepresentations. However, even if there were a way of making the possibility of error detection partially constitutive of the content of a representation (which I doubt), my main point would still be left unanswered: in the theoretical framework cognitive science offers genuine representations need not be representations for the entire organism.

[161] As an anonymous referee of the journal *Synthese* did.

[162] Notice that I do not actually dispute this claim. Above I have denied *only* the fact that the accuracy conditions of a posit are causally relevant to a system's success is *sufficient* for that posit to qualify as a representation. But this clearly does not exclude that having causally relevant semantic properties is *necessary* in order for a posit to qualify as a representation.

properties *just are* properties of the vehicle; namely the features that make the vehicle structurally similar to a relevant target (see O'Brien 2015a; Williams 20017; Williams and Colling 2017; Lee 2018). And, if the relevant structural similarity is exploited, these properties are *guaranteed* to be the properties that are causally relevant to the system's behavior (Gładziejewski and Miłkowski 2017). So, structural representations seem to be exactly the kind of posits that *should* meet the Job Description Challenge. If, as I've argued, they do not meet it, then there is probably something wrong with the Job Description Challenge itself. Maybe it is too demanding. Maybe it still hangs to a non-naturalistic conception of intentionality and content. At any rate, if *no* candidate representational posit is able to meet the Job Description Challenge, then the problem is likely to be the Job Description Challenge itself, rather than any candidate representational posit in question. Compare: if *all* the students *always* fail their tests, then one would be inclined to think that the problem is the tests, rather than the students.

However, I do not think that my argument entails such a *reductio* of the Job Description Challenge. To start, my argument, if correct, only shows that structural representations do not meet the Job Description Challenge. It is silent on whether other types of representations meet it. Maybe they do or maybe they don't, but adjudicating this issue lies significantly beyond the scope of this chapter.

Moreover, alongside structural representations, there is another kind of representation that is widely supposed to meet the Job Description Challenge, namely *input-output representations* (see Ramsey 2007 pp. 68-77). These are representations of the values and arguments a computational system is supposed to compute upon. For instance, if *really* feedforward artificial networks acting as recognition models compute the probability of a label given (i.e. conditioned over) an input vector, they will need to manipulate vectors (arrays of variables or values) and probabilities (a value ranging from 0 to 1), which are mathematical objects. Since physical systems cannot manipulate (at least *prima facie*) mathematical objects, they must

manipulate something that stands-in for them, and that represents, in an appropriate way, the relevant mathematical objects. These are input-output representations (see Ch. 2: §3.4).

As far as I can see, my argument does not change this state of affairs: if really input-output representations meet the Job Description Challenge[163], they meet it whether my argument is correct or not. And, if input-output representations meet the Job Description Challenge, the Job Description Challenge *can* be met. It would thus be false that *all* students *always* fail the test.

But what if it turns out that input-output representations fail the Job Description Challenge too? Wouldn't that show that there is something wrong with the Job Description Challenge? Maybe yes. Yet notice: I'm not claiming that input-output representations fail the Job Description Challenge. The claim that input-output representations fail the Job Description Challenge might be a *reductio* of the challenge, but that claim is not defended here, and so the argument offered in *this* chapter is, as far as I can see, no *reductio* of the challenge.

Moreover, even if it turns out that *no* candidate class of representational posits meets the challenge, the charge of *reductio* strikes me as excessive. Discovering that no representational posit meets the Job Description Challenge would be a *reductio* of the challenge only given a strong prior representationalist assumption. But one could also have some prior inclination towards anti-representationalism, and conclude that the Job Description Challenge yielded a correct result in each case. Now, I do not wish to adjudicate here whether one should be inclined more towards representationalism or anti-representationalism. I will only notice that, insofar representationalism and anti-representationalism are not taken to be *a priori* truths, but rather empirical research programs (or at least the conceptual bedrocks of empirical research

[163] Importantly, this at least partially depends on the theory of computational implementation one endorses. Here, I will stay neutral on the issue. Notice, however, that many (I suspect the majority of) theories of computational implementation try to avoid *pancomputationalism*; namely, the view that any complex physical system implements a number of (or perhaps all) computations (see Piccinini and Maley 2021 for a review). The important point to notice, for present purposes, is this: that many accounts of computational implementation *would not* deem sufficient, for a physical system to compute a function, that the causal goings-on internal to the system systematically "mirror" the transition between computational states. Thus, if the idea common to these accounts is correct, input-output representations need to be *more* than causal mediators allowing a system to "march in step" with some relevant computable function.

programs), we should be open to revise our representationalist or anti-representationalist inclinations.[164] Thus, even if it were true that no candidate representational posit meets the Job Description Challenge (a strong claim that this chapter does not support), that fact alone would not *necessarily* lead to a *reductio* of the challenge. It might also lead to a revision of one's representationalist commitments.

4 - Conclusion

In this chapter, I have argued that structural representations, as Gładziejewski defines them, do not meet the Job Description Challenge. In other terms, physical structures satisfying (a) to (d) function *merely* as causal mediators within the systems deploying them, and do not behave in a recognizably representational way. Hence, even if an argument to vindicate the claim that generative models meet condition (a) were provided, there would still be a reason *not* to consider generative models as representations (structural or otherwise).

Suppose my claims, thus far, have been on the right track, and generative models really are not (structural) representations. It is now natural to wonder what generative models *are*. In the next section, I will answer that question, claiming that generative models are *non-representational* structures instantiating an agent's sensorimotor mastery.

[164] This claim is typically made by philosophers leaning towards anti-representationalism (e.g. Chemero 2009; Ramsey 2017). But the rationale behind it works both ways: if anti-representationalism is *not* an *a priori* truth, one ought to revise one's own anti-representationalist commitment in the light of the relevant empirical evidence.

Chapter Six - Generative models as nonrepresentational structures instantiating sensorimotor mastery[165]

1 - What could generative models be, if not structural representations?

In the last two chapters, I have argued that, at present, we have no compelling reason to think that generative models qualify as structural representations (Ch 4) and that, even if we *had* such a reason, their representational status would at least be dubious, as they would not meet the Job Description Challenge (Ch 5). But what could generative models be, if not structural representations?

Here, I try to answer, claiming *generative models are non-representational structures instantiating an agent's sensorimotor mastery*. I will argue generative models are *implemented* in physical structures that are not representational vehicles; and that these physical structures *instantiate* the agent's tacit knowledge (or "practical know-how") of the regular ways in which bodily movements change the incoming flux of sensory stimulation. To substantiate this claim, I examine the simplest PP system capable of active inference I know of, in the form of a simple robotic "brain". I show that nothing in that "brain" qualifies as a representational vehicle, and that such a conclusion *likely generalizes* to all PP systems. This will naturally make the physical structures implementing generative models appear as *non-representational* structures instantiating an agent's sensorimotor mastery.

The next section concisely exposes three necessary features the obtaining of which identifies representational vehicles. Section three introduces the robotic "brain", clarifying its functioning. Section four argues that such "brain" hosts no structure qualifying as a representational vehicle, whereas section five argues that the same verdict is *likely to generalize*

[165] This chapter is based upon, and expands on, Facchin, M. (2021a). Predictive processing and anti-representationalism, *Synthese*, https://doi.org/10.1007/s11229-021-03304-3

to PP systems more generally. In this way, generative models, as PP conceives of them, will naturally appear to be non-representational structures instantiating an agent's sensorimotor mastery. Lastly, section six considers and allays some worries my claim might raise.

2 - Representational vehicles: three necessary features

Here, I quickly rehearse some commitments of representationalism (already detailed in Ch. 2 and 3), showing how each commitment spells out a condition that identifies representational vehicles; that is, a condition any item *must* satisfy to qualify as a representational vehicle. Hence, jointly, these commitments yield a set of (at least) *necessary* conditions the satisfaction of which (at least partially) identifies representational vehicles.

2.1 - Distality and determinacy

Representations are type-identified by their contents, which are both distal and determinate. Representations "are about" *well-specified worldly targets*, rather than the proximal conditions by means of which these targets are causally encountered (see Chapter 2, §2.2). Hence, representational vehicles can always be assigned a determinate and distal content, given a theory of content.

As seen in (Ch.2: §2.2), the relevant senses of "distality" and "determinacy" are the ones at play in the horizontal disjunction/stopping problem (e.g. Neander 2017; Artiga and Sebastian 2018; Rosche and Sober 2019). 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 an appropriately "out there" target. 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.

Notice that distality and determinacy are *necessary* features of representational contents. This is because if content is not distal and determinate, misrepresentation becomes problematic, if not impossible (e.g. Godfrey-Smith 1989). But representations are partially defined by their ability to misrepresent (e.g. Dretske 1986). So, *necessarily*, they can misrepresent (see again the discussion of Fodor's crude causal theory provided in Ch.2: §2.2)

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 is not really a 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:

> **Distality and determinacy**: if a candidate vehicle V really is a representational vehicle, then there is a correct theory of content C such that, according to C, V represents a *well determinate and distal target* T.

It seems obvious that to assess whether candidate vehicles really are vehicles using distality and determinacy one needs a correct theory of content C. This, *prima facie*, poses a problem: namely that of determining which is the correct theory of content.

2.2 - Exploitable structural similarity

In this context, determining the "right" theory of content is surprisingly easy. As things stand, there are only *two* promising theories of content: teleo-informational semantics and structural similarity based accounts (cfr. Neander 2017; Shea 2018).[166] But, if the arguments provided in (Ch. 5: §§ 2.1.1 - 2.1.2) are on the right track, the former *reduce* to the latter: teleo-semantics is a *specie* of structural-similarity based semantics, and indication is a special case of structural similarity.[167] So, we are left with just *one* account of content, imposing a crisply defined condition any candidate vehicle must satisfy in order to be identified as a vehicle; namely exploitable structural similarity. Thus, by requiring that V represents T *if and only if* V bears an exploitable structural similarity with T, one captures the candidate vehicles that *actually* qualify as vehicles according to the theory. Hence, the second necessary condition:

> **Exploitable structural similarity**: if a candidate vehicle V really is a representational vehicle, then it satisfies distality and determinacy in virtue of an exploitable structural similarity it bears to a relevant target T[168]

As before, I understand structural similarities as second-order structural resemblances.

---

[166] I do not think that this claim is contentious: to be sure, there are other theories of content (such as purely informational or purely causal accounts, as well as accounts based on functional/computational role or interpretational semantics). But all these accounts face terribile and well known challenges (see Cummins 1996: Ch. 3 and 4; Artiga and Sebastian 2018), and, at least as far as I can see, no compelling answer to these challenges has been provided. This seems also a fairly widespread belief, given that, in current philosophy of cognitive science, these theories are hardly endorsed.

[167] Notice that indication is a special case of structural similarity because, whereas indication entails second order structural resemblance, the inverse is not true: second order structural resemblance does not entail indication (cf Shea 2018: 137-140). Now, perhaps one could leverage this point to argue that there's a real sense in which structural similarity and indication are distinct content-grounding relations (if they were *identical*, we would expect *every* case of structural similarity to involve, or be, a case of indication), and thus that teleo-informational account do not *really* reduce to structural similarity based accounts. I'm fairly neutral about this move, as it does no damage to the argument I'm building here. Even if teleo-informational accounts and structural similarity based accounts were distinct, it would still be true that indication entails a form of structural similarity, and that is the only thing that matters for my argument.

[168] Notice that the teleosemanticsist insisting that teleo-informational accounts do not reduce to explitable structural similarity based accounts *must* find this necessary condition fairly *liberal*: after all, since they leverage the fact that there are exploitable structural similarities that do not involve indication, they ought to insist that not all exploitable structural similarities ground contents. Hence they ought to concede that there are instances of contentless exploitable structural similarity; which entails that the necessary condition I'm developing here generates false positives (hence the liberality).

Notice that, in this context, understanding structural similarities in terms of second-order structural resemblances makes the condition *easier* to obtain. This is because second-order structural resemblances are *easier* to obtain if compared with other popular "unpackings" of structural similarity, such as isomorphism or homomorphism (see O'Brien and Opie 2004). Hence, by unpacking structural similarity in terms of second-order structural resemblance I'm placing a (relatively) low bar on what must be the case in order for a candidate vehicle to really qualify as a vehicle.

Recall now the canonical definition of *exploitability* (Shea 2018: 120) discussed in (Ch. 4: §1.2):

> **Exploitability**: V bears an *exploitable* structural similarity to T if, and only if:
>
> (a) The relevant (i.e. structural similarity-constituting) relations holding among the constituents of V are such that the system's processing is systematically sensitive to them; &
>
> (b) The relevant target constituents and their relations are of significance to the system

Condition (a) imposes that the functioning of the system in which the candidate vehicle is tokened must be systematically sensitive to the relevant relations holding among the vehicle constituents. This means that the obtaining, or not obtaining, of a similarity-constituting relation among two or more vehicle constituents must affect the outputs produced by the device. The idea is that the similarity *itself* (and the degree to which it obtains) must govern the success of the system relying on the candidate structural representation to organize its behavior (Gładziejewski and Miłkowski 2017). The more the candidate vehicle is structurally similar to its target, the more the system is likely to (non-accidentally) succeed. And the less the candidate vehicle is structurally similar with its target, the more the system is likely to (non-accidentally) fail.

Condition (b) mentions the fact that the relevant target must be "of significance" to the system in which the candidate representation is tokened. Here, significance should be unpacked

in terms of the *task functions* of the system; roughly, the outputs the system is *disposed* to produce robustly and that it is *supposed* to produce in virtue of some stabilization mechanism having operated over it. The concept of a task function can be unpacked further (see Shea 2018: Ch. 3). Task functions have two main ingredients: robustness and stabilization.

Robustness indicates a property of the *outputs* produced by a system. Here, "outputs" should be understood broadly, as encompassing movements, actions and their consequences (Shea 2018: 55). An output is said to be robust if, and only if, it is produced in response to a range of different *kinds* of input and in different external conditions. What counts as different kinds of input and different external conditions clearly depends on the system in question and its activities. But, in general, for two inputs to be of different kinds their difference must be detectable by the system (e.g. two colored patches that differ only for how they reflect *invisible* light are different inputs for the mantis shrimp but not for me) and they must not be "groupable" together under the head of stimulus generalization (e.g. Pavlov's dog salivating in response to the chime of bell A *or* bell B is not responding to different kinds of inputs). Equally broadly, external conditions count as different only when the variation affects the system's ability to achieve an outcome (e.g. if the outcome is walking in a straight line, "being on Earth" and "being on Jupiter" count as different conditions, as the difference in the gravitational force makes me more or less likely to produce the outcome. A difference in temperature of 0.3° wouldn't, and so does not count as a different condition). (Shea 2018: 55-56).

Stabilization also indicates a property of the outputs, namely that of being produced because they lead to good consequences (Shea 2018: 56). Very broadly, an output is stabilized if the consequences of having produced an output in the past, provide an account for the existence of systems producing that output in the present. According to Shea, we should understand stabilization *disjunctively*: there are many distinct ways in which an output can be stabilized. Natural selection is one such way. With an overused example: hearts are now present and

disposed to pump blood because the blood-pumping of ancestral hearts was evolutionary advantageous for certain organisms, which thus survived and *reproduced* - that is, produced (among other things) other blood-pumping hearts. Learning with feedback is another way in which outputs can be stabilized (Shea 2018: 59-62). An individual organism's history of reinforcements clearly accounts as for why the organism is disposed to behave in the ways in which it is disposed to behave.

Importantly, explicit design can be an alternative way to stabilization (Shea 2018: 64-65). A system might be disposed to produce an output leading to good consequences not because of its past history (or the evolutionary history of its lineage), but because it has been engineered to be so disposed. This is probably the most obvious explanation as for why something is disposed to produce desirable (in some sense) outputs, and perhaps the *only* explanation available before the discovery of natural selection processes (Dennett 1996).

Summarizing: a structural similarity is *exploitable* by a system when (and only when): (a) the degree of structural similarity between vehicle and target influences the chances of the system's success, and (b) the target *matters* to the system, given its functions: the outputs that it produces robustly either in virtue of some stabilization process or explicit human design.

2.3 - Mathematical contents constrain representational contents

The last necessary feature individuating *genuine* representational vehicles is that their representational content (as determined by the relevant exploitable structural similarity) must *at least be coherent* with their mathematical content, as determined by the computations implemented by the system in which they are tokened. Thus, mathematical contents *constrain* representational contents.

> **Mathematical contents constrain representational contents**: if a candidate vehicle V really is a representational vehicle, then its representational content must *at least be coherent* with its mathematical content (i.e. computational role).

This might sound odd: surely, the structural-representationalist interpretation of PP buys into this constraint (Ch. 3: §4.4), but why should anyone else? Yet, the constraint is pretty innocent: recall that mathematical contents are determined by the relevant computations implemented in a system (Ch. 2: §3.4): each vehicle is thus assigned its mathematical content *in virtue of* its computational role. And surely every representationalist endorsing computationalism *must* accept that the content of representational vehicles must *at least be coherent* with its computational role.[169] Otherwise, representational and computational explanations come apart, forcing us to choose between the two.

To see why, consider the following: suppose our best theory of content ascribes to a signal carrying only two bits of information the representational content conveyed by all the volumes of the *Encyclopedia Britannica*. Since all that representational content cannot possibly "squeezed" in just *two* yes/no questions, and the physical shape of the vehicle must allow it to carry the representational content it carries (Dretske 1981: 41; Cao 2012), either the theory of content used *or* the informational description of the system is wrong. I take this to be an entirely unproblematic claim: it seems obvious that the informational capacity of a signal places an upper bound on what the signal can represent.

Now, as signalled in (Ch. 2: §1), I'm here understanding computation as *generic computation*, and transmitting some bits of information from a source to a receiver qualifies as generic computation (Piccinini and Scarantino 2011). Once generic computation is in place, the conclusion that the relevant theory of content must license ascriptions of content that are *at least compatible* with the computational capacities of the system under scrutiny (and the computational capacities of the vehicles tokened within that system) is easily reached. And, again, this conclusion strikes me as entirely unproblematic. If our best theory of content were

[169] One might now wonder what justifies the commitment to mathematical content: isn't computational role enough? I personally think it is (also because I'm now skeptical about mathematical contents, see Facchin *submitted*), and the argument I'm going to present can be entirely rewritten in terms of computational roles. But the structural-representationalist view of PP is committed to mathematical contents, and so I will use it here.

to force us to say that a two-layer perceptron is representing (either *x* or *y*), then something in the theory of content has gone awry: two layer perceptrons simply cannot compute the exclusive disjunction (e.g. Kruse *et al*. 2016: 19-20)

Notice also that a *far stronger* version of this constraint was unproblematically adopted in classical cognitive science. Indeed, it was embedded in the conception of the mind as a syntactic engine *emulating* a semantic engine. Bluntly, the idea was that albeit computational systems cannot be sensitive to semantic properties, they can be sensitive to syntactic ones; namely, the physical features of vehicles upon which computational processes are defined (Fodor 1980; Dennett 1987). Thus, by arranging syntactic (computational) properties and semantic ones so that they "march in step", computational systems can *behave as if* they were sensitive directly to meaning. Hence, Hageland's (1989: 106) *dictum*: "If you take care of the syntax, the semantics takes care of itself". Mathematical contents and representational ones had to "fit" each other to a non-trivial extent. Thus surely mathematical contents constrained representational ones.

Whilst I don't endorse that view, it is surely worth mentioning it here to highlight *how very innocent* the third criterion is, and how deeply it is woven in the very fabric of computational and representational explanations.[170] Indeed, in order for *computational and representational* explanations to make sense, it seems necessary that computation and representation must "march in step", hence the former must place *at least some* constain on the latter.[171]

Tacking stocks: if a candidate vehicle V really is a vehicle, then it has a determinate and distal content, which bears in virtue of the exploitable structural similarity that V bears with a distal and determinate target T, and such a representational content is at least coherent with its mathematical content (that is, its computational role).

---

[170] See also (Piantadosi 2021; Mollo 2021) for two recent defenses of the idea that representational content should at least be constrained by computational roles.

[171] Yet notice that the opposite need not be true: for example, it would not be true if computation does not require representation, as some argue (e.g. Piccinini 2006).

In the next two sections, I will introduce a simple generative model capable of active inference, and argue that *none* of its components satisfies that description. Generative models will thus appear as *non-representational* structures instantiating the system's sensorimotor mastery.

3 - A simple robotic "brain" capable of active inference

Here, I introduce a simple robotic "brain" capable of active inference. First, I introduce the architecture and its functioning principle. Then, I consider its operations "*in vivo*" by means of an example. An important note concerning the experimental methodology closes the section.

3.1 - The architecture and its functioning principle

According to PP, generative models are physically instantiated by patterns of neural activation and axonal connections (Friston 2005: 819-820; Buckley *et al.* 2017: 57); these are the relevant candidate vehicles. Connectionist systems are thus ideally suited to examine the representational commitments of PP (Dołega 2017; Kiefer and Hohwy 2018; 2019).

Consider the network Bovet (2007) engineered as a control system for robotic agents, enabling them to execute 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 and Bovet 2009) or successfully navigating simple T-mazes (Bovet and Pfeiffer 2005a; 2005b).

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.

The overall structure of the network is displayed in **figure 5**

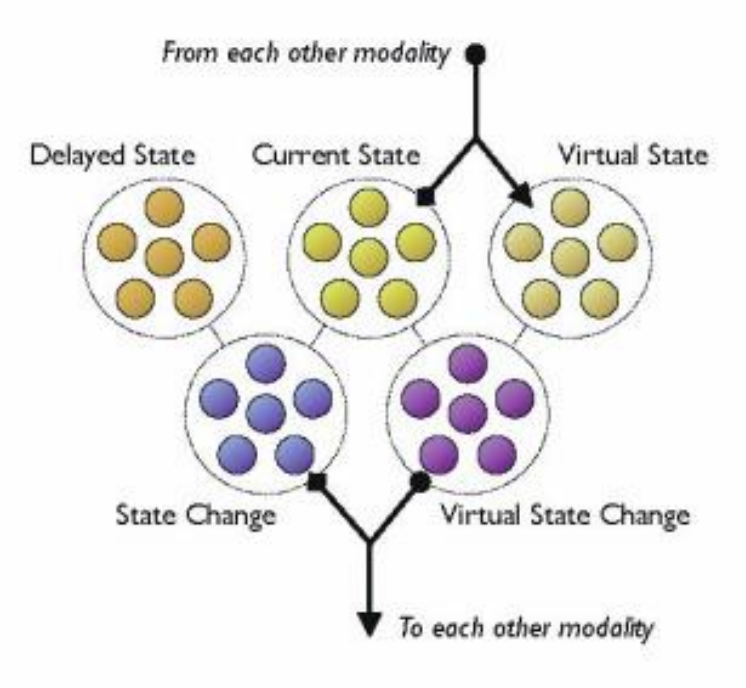

Figure 5. Implementation of the model (one modality). © IEEE. Reprinted with permission from (Bovet and Pfeiffer 2005b).

The number of neurons in 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.[172] 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 (i.e. (VC)s and (VS)s are fully connected).

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.[173] 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 modality *a* and *b*: there will be patterns of co-activation 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

[172] Hence, the basic function computed within each modality is vector subtraction.
[173] 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.

expected, given the activity in all (SC)s.[174] 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 (on generative models and sensorimotor contingencies see Ch. 1; §4).

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.[175]

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.[176] In this way, Bovet's networks qualify as minimal PP systems, able to "actively infer" the sensory states expected in all modalities.

---

[174] 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.

[175] 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 2008:16). Intriguingly, such an horizontal message-passing is rarely implemented in robotic models inspired by PP, see (Ciria *et al*. 2021).

[176] 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.

3.2 - The functioning *in vivo*: an illustrative example

Now, to see this simple generative model in action, consider the following experiment in which the network enabled a form of "phonotaxis"[177] comparable with that of female crickets (Bovet 2007: 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[178]; 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, Hedwig and Webb 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 visceroception. 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

[177] 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.

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

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 competition between "phonotactic" and optomotor behaviors can be seen in (Bovet 2007: 90, figures 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".[179] Hence, the synaptic coupling between visual and motor modality constituted a simple forward model[180]; and, more generally, the

[179] 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).

[180] 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.

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).[181]

3.3 - A note on synthetic methodology

Before I move forward, I need to place a *caveat*. 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, given how Bovet describes his experiments and his overall commitment to a *synthetic methodology* of research. I briefly elucidate both points in order.

First, Bovet describes his systems in two different ways. *Networks* are described proximally, as receiving *signals from sensors and motors* and correlating those signals. For instance, when introducing the general architecture of the network, Bovet writes:

> "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*. 2) All populations of artificial neurons are homogeneously coupled to each other through artificial synapses, whose plasticity follows a simple rule well-known to biologists: 'neurons that fire together wire together'" (Bovet 2007:12, emphasis added)

Robots, however, are described distally, in terms of environment-involving behaviors. The

[181] On experience-dependent adaptability, see in particular (Bovet and Pfeiffer 2005a; 2005b).

"phonotactic" examples provided above clarifies the point: the robot is described (for instance) as *approaching the "sound source"*, or as *manifesting the optomotor reflex*. So, there are two levels of descriptions at play here. One, pertaining to the networks, is proximal. The other, pertaining to the robot, is distal. The *proximally describable* goings-on in the network give rise to the distally describable behavior of the robot. For instance (again, using the "phonotaxis" experiment described above), the mismatch between the patterns of activation of viscerоceptive (CS) and (VS) generates prediction error, which modifies the states of auditory and motor (VS), thereby *giving raise to the robot's* "phonotactic" behavior.[182]

Now, on Bovet's synthetic methodology of research. In extremely succinct terms, the aim of the synthetic methodology is that of "understanding by building" (see Pfeiffer and Bongard 2007, Ch. 1 and 3, Tani 2016, Ch. 5, Hoffman and Pfeifer 2018). The basic idea animating it is that our *best* way to understand the mechanical underpinning of some (cognitive) behavior of interest is that of creating a *real world* artifact able to exhibit that behavior. Proponents of the synthetic methodology put forth two reasons as to why building *real* artifacts to understand cognitive behaviors should be our way of proceeding.

One is that it avoids excessive abstraction and idealization. Surely, robotic and artificial neural networks models *are* abstract and idealized (e.g. they typically have less degrees of freedom than their biological counterparts). Yet, since the products of synthetic methodology are *real world artifacts*, these artifacts must comply with physical laws that are usually ignored in cognitive models, in a way that might sometimes strain the outputs of research (see Pfeiffer and Bongard 2007: 68-70).

The other is the "law of uphill analysis and downhill invention":

> "It is pleasurable and easy to create little machines that do certain things. It is also quite easy to observe the full repertoire of behaviors of these machines - even if it goes beyond what we originally planned, as it often

[182] Notice that this is a simple explanation of the robot's distally characterized behavior that does not involve any usage of the term "representation". So, I'm delivering what I've promised in (Ch. 5: §3.4.2).

> does. But it is much more difficult to start from the outside and try to guess the internal structure just from the observation of behavior. [...] A psychological consequence of this is the following: when we analyze a mechanism, we tend to overestimate its complexity. In the uphill process of analysis, a given degree of complexity offers more resistance to the workings of our mind than it would if we encountered it downhill, in the process of invention. " (Braitenberg 1984: 20 - 21)

If this law is correct, then the best (i.e. the simplest and most productive) way we have to understand intelligent behavior and the mechanisms generating it is to *try to build* the simplest possible mechanisms generating the target behavior.

Bovet's work is an *extreme* example of synthetic methodology. For, typically, researchers adopting it have a target *biological* behavior that they wish to replicate in a robot (e.g. Webb 1994) or at least a target behavior that they wish the robot to exhibit. But Bovet's robotic models *have no such target*. Indeed, Bovet's models *have no purpose*:

> "Before the agent starts interacting with its environment, all synaptic weights of the network are initialized to zero. In other words, sensors and motors initially do not interact: *the agent has no built-in reflexes, nor any similar behavioral primitives provided by an external designer*. [...] *Similarly, the system has no goal*." (Bovet 2007: 30, emphasis added)

The same holds for their controllers:

> "The main characteristic of the neural architecture is the absence of explicit control or regulation mechanisms." (Bovet 2007: 29)

And in fact Bovet describes what its networks are designed to do in squarely proximal terms.

The importance of these methodological points will become apparent in the following.

4 - The "brain" hosts no representational vehicle.

Thus far, I have identified three necessary conditions a candidate vehicle must satisfy in order to be rightfully identified as a vehicle, and introduced a simple robotic "brain" capable of active inference. It is now time to check whether the structures in that "brain" qualify as vehicles according to the criteria given above.

4.1 - Activation patterns are not representational vehicles

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, Bengio and Courville 2016: Ch. 15). But the network has no hidden layers.[183] It is thus doubtful whether we should consider its activity patterns as candidate representational vehicles.

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 - so cheap they can be arbitrarily defined (Shea 2018: 112-113). Hence, it is extremely likely that the patterns of activation of the 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 (from § 2.2 above) that *exploitability* is canonically defined as the conjunction of two requirements. A vehicle V bears an exploitable structural similarity with T just in case (a) the structural similarity-constituting relation or relations holding among vehicle constituents are such that the system is systematically sensitive to them, and (b) the target T is "of significance" to the system; that is, T *matters* (in the broadest possible sense of the term) for the system's task-functions. Recall further that task-functions are outputs that a system produces robustly, either because of some stabilization process (natural selection, learning) or because of *explicit design*. Since Bovet's networks (i.e. the systems in which the patterns of activation are tokened) are designed, their relevant task functions are determined by their designer.

---

[183] A reviewer of the journal *Synthese* objected that (VS)s should be counted as hidden layers, because they do not receive inputs from sensors or motors (and so they are not input layers) nor they convey outputs to other networks or effectors (and so 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.

However, as seen above (§ 3.3) Bovet describes his networks in squarely *proximal* terms. He says, for instance, that (CS)s only produce patterns of activation that capture the state of a 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: 528, italics added). Similarly, he states (SC)s have been designed so as to reflect how the *sensory inputs* have changed in a timestep. Equally proximal descriptions are in fact given for each neural population. Notice that, when he so claims, Bovet is telling how the network has been *designed* to operate. That is, he is expressing the task functions of the network and its components.

But then, *by design*, the network's task functions target only proximal states, and therefore only proximal states will be of significance to it. And 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*. Hence, distality and determinacy fail to obtain. As a result, it should be concluded that the candidate vehicle is not *really* a vehicle. Conversely, if we assign candidate vehicles distal targets, then exploitable structural similarity fails to obtain, for the candidate vehicle is *not* assigned a distal (and determinate) content in virtue of an *exploitable* structural similarity.

In sum, the relevant candidate vehicles (patterns of activation in Bovet's network) appear to be unable to satisfy distality and exploitable structural similarity *in conjunction*. As a result, these candidate vehicles are not really representational vehicles.

The same holds if instead of single patterns of activations we focus on the entire activation space (e.g. Churchland 1989; 2012): 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.

Notice also that this verdict does not change if instead of focusing on the entire activation space or single activation pattern one focuses on the activity of *single units*. It is surely possible to construe single units as receptors, and thus as elements that must bear an exploitable structural similarity to something. The problem is that, as Bovet's citation provided above testifies, single units are receptors *only* of proximal states, such as the values of single camera pixels or the activity level of some motor. Hence, just entire activation spaces and single patterns of activity, they fail to bear an exploitable structural similarity to a *distal* target. Hence they fail distality, hence they are not representational vehicles.

What if one focuses on the task-function of *the robots*, rather than the task functions of *the network*?[184] 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.

Yet, as highlighted above, Bovet very clearly states that, by design, his robots *have no function*. There is nothing they are *supposed to do*: they are not designed to produce any target distally characterized behavior, and so there is no output (in the relevant sense) that they are designed to produce. Indeed, Bovet describes his robots as: "artificial systems endowed with a self-developing dynamics, yet *without any particular task or motivation*" (Bovet 2007: 8, emphasis added). Given that robots are artificial systems, and that the task functions of artificial systems are determined by their designer, it seems correct to conclude that Bovet's robots have just no task function, long-armed or otherwise. Hence considering *the robots'* (non-existent) task functions will not solve the problem with distality I just raised.

Couldn't perhaps the patterns of activation have acquired some distally characterized function through the network's individual learning history? A negative answer seems

[184] I owe this objection to an anonymous reviewer of the journal *Synthese*.

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 1988; Shea 2018: 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; 2017).[185] *Mutatis mutandis*, the same reasoning seems to apply to entire robotic agents. Notice that this is actually entirely compatible with the way in which Bovet characterizes the "developmental trajectory" of its robots:

> "The Hebbian learning rule, which modifies the synaptic weights of the network, *is not modulated by any value system that would define a particular goal*. It is worth noting at this point that the term learning can be slightly misleading: the Hebbian learning rule is *not a learning strategy allowing the agent to achieve a given task or optimizing a given fitness function;* rather, it is an arbitrary rule that defines the synaptic plasticity of the network [...]" (Bovet 2007: 30 emphasis added)

Maybe we should assign content to single activation patterns (or, single units activation) in a different way. Wiese (2018: 219-223) has in fact recently suggested a different procedure to do so. He suggests that albeit the (generative) model as a whole represents the causal structure of the world in virtue of the exploitable structural similarity holding between the two, 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: 34). That is, a neuronal response represents the events that, on

[185] 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.

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: 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: 96; Shea 2018: 125; Kiefer and Hohwy 2018: 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 ($v_a...v_n$) map one-to-one onto T's constituents ($t_a...t_n$) in a way such that the same *pattern* of spatial relations holds among both ($v_a...v_n$) and ($t_a...t_n$). 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 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).[186]

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 by *modus tollens* 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.

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.

4.2 - Connections are not representational vehicles

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

[186] I will expand upon this point down below, in § 6.3

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: 26-29). The mathematical content of these connections (i.e. their weight value) is thus exclusively determined by factors lying *inside* the system and carrying only proximal content (if any). 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.

The argument above is not conclusive, so I'm forced to concede that it *might* be possible to assign determinate and distal contents to weighted 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 standard notion of superpositionality (see Van Gelder 1991; 1992; Clark 1993: 17-19), superposed representations cannot be structurally similar to their targets (this was anticipated in Ch. 4: § 4.2).

Recall the relevant notion of superpositionality at play (Ch. 2: § 3.3). The notion is defined in terms of a vehicle being conservative over a target (Van Gelder 1991: 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", and "John" has no "representational space" left to represent something other than John. Conversely, "John plays" is not conservative over John: to represent "John" I need not minimally token "John plays", which as a matter of fact *has* the representational space to represent something more 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 $T_a...T_n$ just in case V is conservative over each member of $T_a...T_n$. Notice the

plural: superposed representations are always, by definition, conservative over more than one target (Van Gelder 1992; Clark 1993: 17-19).

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 $v_a...v_n$ 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 and Opie 2004: 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.[187]

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.

---

[187] To my dismay, I discovered that this claim is not original: "In general, schemes of representation define a space of allowable representations and set up a correspondence with the space of items or contents to be represented. We are accustomed to thinking of such schemes *as setting up a roughly isomorphic correspondence* [...]. The notion of *superposed representation overthrows this whole familiar picture*, for superposition aims precisely at finding *one point in the space of representation that can serve as the representation of multiple contents*". (Van Gelder 1991: 45-46, emphasis added).

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). This argument has already been presented at length in (Ch. 4: § 4.5) and will not be discussed further here.

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.

4.3 - The network as a whole is not a representational vehicle

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 and Hohwy 2018: 2394-2395; Wiese 2018: 219). Albeit I think this is a fair point, I fail to see how it might challenge my conclusion.

For one thing, as already noted in (Ch. 4: §§4.6-4.8), we should not confuse the vehicles tokened in representational systems with the systems in which vehicles are tokened. Although it is certainly true that there is a sense of the word "representation" according to which it makes sense to say that entire systems represent, this sense of "representation" is distinct from the one at play when we refer to things that *are* representations (that is, the material objects that function as representational vehicles). Hence, although it is entirely correct to speak of networks (and other computational systems) representing some target, this sense of the world "representation" is distinct from the one at play when we say that a pattern of activity of a network (or some other computationally relevant state of a computational system) represents a

target. In this second sense, the term "representation" denotes a representational vehicle: when one says that a pattern of activity represents, one is (typically) saying that the pattern of activity is a representational vehicle tokened within a system and bearing a certain content; and, in fact, rather than saying that the pattern of activity represent a target, one can easily say that the pattern of activity *is a representation of* that target. Conversely, when one claims that a network represents a target, one is typically *not* claiming that the network is tokened in some larger system in which it functions as a representational vehicle; rather, one is claiming that some specific vehicle is currently being tokened within the network; and, in fact, one could hardly *sensically* say that a network is a representation of a target. Entire networks are computational systems within which representational vehicles are tokened. It follows that they are distinct from the representational vehicles tokened within them.

Notice, however, that even if the distinction above were spurious, there would still be reasons as to why the entire network would fail to qualify as a representational vehicle. After all, it would still be 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, a candidate vehicle exploitably structurally similar to its target, 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, whilst the network instantiates a simple generative model "knowing" the robot's sensorimotor contingencies, the structures implementing that model do not qualify as representational vehicles. Hence, they are *non-representational structures*. Nevertheless, as the example discussed above shows, these non-representational structures manifestly *instantiate* the

system's knowledge of its own sensorimotor contingencies. The network here examined is thus a *non-representational* structure that instantiates the relevant system's knowledge of its own sensorimotor contingencies.

But what about *all other* PP architectures instantiating generative models?

5 - Will it generalize?

I think the most obvious objection to the verdict I have provided above is that it will not generalize. Bovet's architecture is *really* a simple architecture, and PP systems are typically far more complex than that. Moreover, it is just *one case*, so it is definitely not a strong inductive basis. Indeed, as previously noticed (Ch. 4: §4.5), there are many different network architectures, some of which far removed by the simple network considered above: why should one believe that no genuine vehicle could be found in those?

These are simple, *but powerful*, reasons to believe that my verdict will not generalize. And I do think they are solid reasons. But I also think that my verdict *is likely to generalize*, and in this section I am going to provide a cluster of arguments to this effect. To be clear, none of my arguments will be conclusive. But this, I think, is fine: after all, I'm basically providing an inductive generalization, and inductive generalizations are defeasible by their very nature.

That being said, what reasons are there to believe that my verdict will not generalize?

5.1 - The model is not deviant

Surely one reason is that Bovet's network is *very unlike* other networks implementing PP systems. So, perhaps it is a *deviant* PP architecture full of idiosyncratic features, and the outcome of its analysis does not easily transfer to other PP architectures.

A problem with this objection is that it presupposes that there are non-deviant PP architectures. But *all* PP architectures are deviant to a degree, as there just is no standard

implementation of PP, especially when it comes to robotic models (Ciria *et al.* 2021). So, for instance, there are PP architectures that do not have distinct prediction and error units (O'Reilly, Wyatte and Rohrlich 2014), or that have no distinct set of ascending and descending connections (Matsumoto and Tani 2020), or that incorporate specific "parametric bias" units to govern their computational functioning (Tani 2014). There are also architectures that do not represent *neither predictions nor prediction errors* (Thornton 2017). All these architectures are significantly different from the one proposed by Rao and Ballard (1999).[188] But they are nevertheless PP architectures in the full sense of the term. Now, if this is the case, and the relevant insights that the analysis of these "deviant" networks offer generalize, why *shouldn't* the insights obtained analyzing Bovet's network generalize too?

5.2 - Missing ingredients do not block the generalization

Maybe, then, the problem is that Bovet's networks lack a key ingredient of PP - one which, when taken into account, would force the revision of the anti-representationalist verdict I have offered here. But what could that ingredient be?

Hierarchy is plausibly the most obvious candidate. Bovet's networks are non hierarchical, whereas the majority (but by no means all, see Tani 2014; Lanillos and Cheng 2018) of PP systems are. However, I simply fail to see how hierarchy would force me to revise my verdict.

To start, adding hierarchy means adding hidden layers and connections to (and from) these layers. But these connections would *most likely* be superposed representations just as the connections of Bovet's network. The "most likely" qualification is important: I cannot in principle exclude that in some particular connectionist implementation of PP connections will turn out to be structurally similar to the (distal and determinate) target domain the network has

[188] It is also worth noting that there is no *intrinsic* reason as to why Rao and Ballard's (1999) network should be taken to be the *canonical* connectionist implementation of PP. As far as I can see, Rao and Ballard's network is just a convenient (hence, often cited) example.

been trained to operate upon. But, at present, I see no *positive reason* to believe that such an exceptional connectionist system exists (or that it will be produced). Weight matrices of connectionist systems have traditionally been considered *superposed* representations, and the only argument I know of claiming that they are structural, rather than superposed, representations is rather weak (see Ch. 4: § 4.3). Hence, when it comes to considering the putative representational role of weighted connections in hierarchical PP systems, *the argumentative burden is not mine*; rather, it is carried by those who wish to claim that connections qualify as representational vehicles in the light of the (fairly minimal) necessary conditions stated above.

But what about the hierarchically higher layer of units? Won't they qualify as representational vehicles? There are, I believe, several reasons suggesting a negative answer. One is that it is *very* doubtful that their patterns of activity can be assigned appropriate distal contents (O'Regan and Degenaar 2014). From a strictly computational point of view, hierarchically higher layers are just models of the layers directly *below* them, as they must only learn to predict how the layer directly below them will behave.[189] To give a concrete (but not overly complex) example, in hierarchically stacked recurrent neural networks encoding a generative model, the hierarchically higher layer only estimates the rough behavior *of the recurrent network directly below it* in the next timestep (see Tani 2003).[190] Notice that the same *purely proximal* ascription of content carries over in the informal presentation of the computational activity of these layers. In fact, these layers are often informally characterized as producing "abstract statistical summaries of the *original visual input*" (Bulow *et al.* 2015: 5-6; emphasis added) or representing "increasingly sophisticated aspects of the *original input*"

[189] (see, for instance Hinton 2007b; Eliasmith 2013: 92; Simione and Nolfi 2015; Dołega 2017: 12-13; Spratling 2017: 97). See also (Ch. 1; §2.1).

[190] More precisely, the hierarchically higher layer estimates the *parametric bias* vector: an input vector of the lower layer that basically determines the sensory predictions of the lower layer (by determining its predictive activity).

(Foster 2019: 33). Notice that this seems a case in which mathematical contents (i.e.the computational activities of the higher level under scrutiny) *are constraining* the representational content ascribed to these layers. And they are doing so in favor of a *proximal* characterization. This poses a challenge to a representationalist reading of these layers, as it strongly suggests that their content would not be distal (see also Orlandi and Lee 2019).[191]

But what about the standard account of representations in hidden (i.e. hierarchically higher) layers of artificial neural networks? Often, careful mathematical analysis of the activity of hierarchically higher layers reveals that the activity of these layers exhibits a *structure preserving* mapping g holding between patterns of activation and features of the *distal* domain the network has been trained to operate upon (e.g. Elman 1991; Shagrir 2012; Churchland 2012). If the output of these analyses are correct, then *there is* a (likely exploitable) structural similarity holding between hierarchically higher layers and appropriately distal and determinate targets. This would be enough to block the generalization I'm trying to defend.

In reply, I wish to notice two things. One is that although it is surely *possible* that analytic techniques *will* reveal such structural similarities, it is also possible that they *will not* reveal them. So, when it comes to finding structural similarities in hierarchically higher layers, both the representationalist and the anti-representationalist are *making a bet*. And, surely, the representationalist can marshall some empirical evidence favoring its position. But the anti-representationalist can do it too. Consider, for instance, the mathematical analysis of the networks used for the neuro-robotic experiments reviewed in (Tani 2003; 2014; 2016: Ch. 6-9). The networks are simple recurrent neural networks with a parametric bias, busy predicting the sensory stream the robotic agent will experience at the next timestep. Importantly, these networks have three kind of input units: "normal" input units (reflecting the state of the relevant

[191] Notice that, at least sometimes, defenders of the structural representationalist reading of PP acknowledge this point. For instance: "One key task performed by the brain, according to these models, is that of guessing the *next states of its own neural economy*" (Clark 2013a: 183).

sensor or motor), "context" input units (which "recycle" part of the net's output as input at the next timestep) and *parametric bias* units, whose state can be either externally set or generated by the networks' dynamics depending on the task. The mathematical analysis of these networks often reveals the presence of a structure-preserving mapping in between the networks' activity and the robot's domain of operation. However, such a mapping is often *very* rough (e.g. in Tani 2003 and Tani 2016 Ch. 8 the mapping only consists in the bi-partition of a single plane), and holds between the level of activation of the parametric bias units and the robot's actions *expressed in terms of the robot's sensory input* (e.g. joint angles of the robot's actuators). Hence, the relevant structure-preserving mapping holds between a relatively small portion of the network's structure and a *proximal* target (roughly, the state of the robot's sensors and motors). Surely this kind of structure-preserving mapping does not support a structural-representationalist reading of the network - even if it does illuminate the network's functioning in an explanatorily useful way.

The second thing I wish to notice is that, even when an appropriately structure-preserving mapping is revealed by means of mathematical analysis, its mere existence might not be enough to substantiate the desired structural representationalist reading (see Ch. 4: §4.4). For, such mappings often hold not among *individual* patterns of activation and targets, but *the entire activation space* and the target domain. But the activation space *is not* a vehicle that can be tokened within a connectionist system: it is an abstract mathematical space that describes its behavior. Moreover, activation spaces-target domains structural similarities are often defined in a way that does not take into account weighted connections. However, if V is a structural representations of T, then changes to V that make it more accurate *just are* changes to V that make it more structurally similar to T. And, in the case of artificial neural networks, the change that makes them more accurate surely includes changes in their connections weights, hence it is correct to say that *if* artificial neural networks are structural representations, *then* weights are

elements in the relevant structural similarity. But, in the case at hand, we are denying weights any role in the relevant structural similarity. Hence accepting it should lead us to conclude (by *modus tollens*) that artificial neural networks are not structural representations.

I do not think that any of my two remarks provides *conclusive* evidence in favor of the claim that hierarchically higher layers do not qualify as representations. So my argument is again 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.

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 2012a) 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.

A further "missing ingredient" that might block the generalization depends on the fact that

Bovet's networks are relatively *ancient*, and so do not take advantage of many techniques that are used in more recent connectionist systems (e.g. max-pooling, long-short term memories). While this is certainly true, I do not think that the fact that Bovet's networks do not take advantage of such techniques blocks the generalization I'm proposing. The reason is that such techniques do not seem to be part of the computational/algorithmic apparatus of PP (cfr. Millidge, Seth and Buckley 2021). And though it is most likely possible to *augment* PP systems by means of such techniques (e.g. Ororbia and Kelly 2021), these techniques seem extraneous to the computational core of PP (at least as exposed in Ch.1).

5.3 - Considering non-aritficial PP systems does not solve distality

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 and 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).[192] In fact, the discussion about what PP systems are supposed to do is typically couched in *proximal* terms, such as avoiding sensory states with high surprisal[193] or encountering the sensory states predicted by the model (see Hohwy 2020b). 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[194] 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

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

[193] 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.

[194] This assumption is a corollary of the assumption that sensory states are under-informative in respect to their worldly causes (see Orlandi 2014; Anderson 2017 for discussion).

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, Clark and Thornton 2012).

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.[195]

5.4 - Representation hunger

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 able to confront more complex (less "low level" sensomotoric) task domains, my verdict would have been different, as confronting with such more complex task domains simply *requires* representations. The idea is thus that there are two different "tiers" of cognition: low-level sensorimotor coordination and everything else.

This challenge is typically made precise by invoking "representation hungry" cognitive

[195] In the discussion above, the relevant notion of function at play was that of task-functions, as defined by Shea (2018). But that clearly isn't the *only* notion of function on offer at the philosophical market (see Garson 2016 for an overview). So one might wonder what would happen if a *different* notion of function were adopted: would a different notion of function allow one to identify genuinely representational vehicles? Here, I clearly *cannot* consider each and every individual notion of function to check whether it would allow to vindicate representationalism about PP. But I will provide a quick remark to motivate a negative answer. Consider an alternative account of functions *F*. Is the account *normative* (i.e. able to tell apart function from malfunction) or does it identify functions with causal dispositions (e.g. Cummins 1975)? If the former, then the remarks about normativity made above seem to apply. If the latter then it seems that there is a problem when it comes to account for misrepresentation. One reason as for why accounts of representations based on functions are attractive is that the normativity of functions allows to understand misrepresentation in a straightforward way: if V is, for instance, supposed to indicate T, then a tokening of V can be easily said to misrepresent whenever its tokening does not indicate what it is supposed to indicate. Yet, if functions are not normative, this attractive move is precluded, and an alternative account of misrepresentation is needed. The non-normativist about function is thus faced with a challenge, which, as far as I know, has not been met (yet).

domains; that is, task domains in which *no environmental signal* can guide the agent performance (Clark and Toribio 1994). These task domains are defined either by the *physical absence* of the cognitive target (e.g. planning one's next summer vacation) or the sensitivity to *parameters whose physical manifestation is complex and unruly* (e.g. discriminating dogs and cats). Since, in these task domains, there seems to be no unambiguous environmental signal able to lead the agent's performance, then the agent's performance must be due to some knowledge-structure internal to the agent; that is, a representation.

Hence, the generalization I'm proposing is blocked because cognition is divided in two distinct "tiers"[196]: low-level sensorimotor coordinations and representation-hungry cognitive tasks. The model I have here considered is a model able to deal only with cognitive task domains belonging to the first "tier". As such, any putative anti-representationalist conclusion extracted from it does not generalize to systems able to master representation-hungry task domains. And, had I considered a model able to master such task domains, *I would not* have drawn an anti-representationalist conclusion (as mastering these task domains *requires* representations to be in place).

I see two reasons as to why the representation hungry challenge cannot block the generalization I'm proposing here. One is purely factual: as a matter of fact, Bovet's networks *can* enable a robot to achieve mastery in a representation-hungry cognitive task domain. The other is conceptual, and it has to do with the explanatory breadth PP is supposed to have. I unpack both points in turn.

---

[196] It is not entirely clear what these "tiers" are. A natural reading is that these "tiers" are two distinct *kinds* of cognitive processes: Indeed, interpreting the two "tiers" as kinds would noticeably bolster the representation-hungry argument. In fact, conclusions reached by means of induction surely cannot be transferred from one kind to another - if we inductively establish that all crows are black, we have not *thereby* established that all cows are black. However, neither Clark and Toribio (1994) nor their commentators (Degenaar and Myin 2014; Zahnoun 2019) characterize the two "tiers" as different kinds. Moreover, Clark has often (Clark 2013b: 153-154; 2015b, 2015c, Linson *et al*, 2018 Constant, Clark and Friston 2021) suggested that the two "tiers" of cognition are in an important sense *continuous*. Yet, the more their continuity is stressed, the less potent the pull of the "representation hungry" argument: if, at the end of the day, the two tiers are essentially the same thing, then it is not clear what what exactly is blocking the generalization from "low-level" sensorimotor coordinations to "representation hungry" cognition.

5.4.1 - Factual problems with "representation hunger"

Factually speaking, in numerous simulations and robotic experiments systems guided by Bovet's architecture *self-initiated* their behavioral routines. This is because the network controlling these systems expected a stream of input that the environment *didn't deliver*, thereby leading to the generation of prediction error and its minimization through active inference (Bovet and Pfeiffer 2005a; 2005b; Bovet 2006). Is the agent guided by an unambiguous environmental signal emanating from its target? If not, then it seems that the network has enabled the agent to achieve some mastery in a cognitive task domain in which the target is absent. If yes, then what is the relevant signal triggering the agents' actions? It seems to me that the only candidate is the flow of the external stimulation that, mismatching the networks' expectations, generates the prediction error that triggers the agents' actions (*via* active inference). But this surely is a parameter whose physical manifestation is unruly and open-ended, as it encompasses *all* the sensory states that are not the expected ones. Thus, in both cases, the networks enabled the robot to cope with a "representation hungry" task domain.

Moreover, the network architecture Bovet engineered is capable of *delayed reward learning* in the context of T-maze tasks (Bovet and Pfeiffer 2005a; 2005b). The experimental procedure is easily explained as follows: at the beginning of each trial, the agent is placed at one end of a "T" shaped maze, typically the one at the end of the long arm of the "T". The agent is then let free to roam the maze - ideally, it should head to one specific arm of the "T", in which a reward[197] is placed. In order to discriminate in which arm to turn, the agent can use a cue - in

[197] 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, Adams and Montague 2012). It is thus worth noting that Bovet himself acknowledges that, when it comes to his experiment, "reward" and "punishment" are *arbitrary* tags, which he uses to simplify the exposition of the experimental procedure. 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; 2005b). 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.

the case at hand, a laterally placed "bumper" stimulating the robot's tactile modality.

Now, albeit delayed reward learning in the context of T-mazes[198] is a simple (and highly standardized) task, it is worth noting that it is typically considered to be a *working memory* task, as, to correctly solve it, the agent has to associate cue and motor behavior, recall the association when the outcome (reward or punishment) is received, and forge a novel association, which must be remembered for the entire duration of the experiment. Thus, solving the relevant T-maze task *prima facie* qualifies as an instance of "representation hungry" cognition. And, in fact, the "memory space" needed to solve such a task has been precisely quantified (in bits, see Kim 2004), and it is often thought that such a task cannot be mastered by agents lacking some form of internal memory storage (Carvalho and Nolfi 2016).[199]

Nevertheless, Bovet's non-representational network managed to enable a robot to solve the task with a high degree of accuracy only by learning a set of relevant sensorimotor associations. More precisely (but see Bovet and Pfeiffer 2005a; 2005b and Bovet 2007: 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 expected visual flow. But by turning, the robot also entered in the correct arm of the T-maze, thus "stumbling upon" the reward.[200] It thus seems correct to conclude that Bovet's networks are non-representational systems able to successfully master at least some representation-hungry cognitive tasks. Thus, the anti-representationalist conclusion I am defending *can* generalize to representation-hungry

[198] These results have been replicated in other kinds of mazes as well, ranging from Y-mazes to more exotically shaped asymmetrical mazes in which each arm has several turns (see Bovet 2007, figures 7-13).

[199] However, it is worth noticing that the experimental data collected by Carvalho and Nolfi positively show that even very simple agents lacking any memory storage can solve both T-maze and *double* T-maze tasks with a decent degree of success.

[200] 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 (e.g. Bruineberg, Chemero and Rietveld 2019).

cognitive domains, and I have obtained such a conclusion by analyzing a system capable of representation-hungry cognition.

5.4.2 - Conceptual problems with "representation hunger"

As said above, I think that there is also a conceptual reason as to why the representation hungry challenge fails to block the generalization I'm proposing here. To put it bluntly, I think that the "representation hungry" line of defense is not open to the representationalist willing to endorse PP, *at least if PP has the explanatory breadth it is said to have*. Supporters of PP often claim that PP delivers us a "cognitive package deal in which perception, understanding, dreaming, memory and imagination may all emerge as variant expressions of the same underlying mechanistic ploy" (Clark 2016: 107). Such a view is widespread in the PP literature, (Seth 2015, Spratling 2016, Pezzulo 2017), and it often gives rise to the claim that PP *explains everything about the mind* (Friston 2009; 2010; Hohwy 2015). Indeed, PP (at least in its most popular, Frinston-inspired) formulation is a *grand unified theory* of brain functioning and cognition (cf Anderson and Chemero 2013). But if this is correct, and 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: there can be no difference regarding the representational status of the two "tiers" of cognition. If the posits of PP are representational (as many believe), then cognition, as PP describes it, is representational *in its entirety*; "low-level", environmentally-driven sensorimotor coordinations included.

And if, as I have argued here, these posits are *not* representational, then cognition, according to PP, will be non-representational in its entirety; "representation-hungry" cognition included.

5.5 - Two-tiering predictive processing

Perhaps, however, there is a way to defuse the argument I have just given and construe PP as a two-tier account of cognition. A forthcoming paper by Gładziejewski could be read as doing just that.

As I understand it, Gładziejewski (*forthcoming*) aims to "fuse" representationalist and anti-representationalist readings of PP in a single overarching conceptual framework. His proposal is that *some* instances of predictions, through non-representational in nature, could be rightfully treated as representational by invoking *implicit representations*, thereby subsuming them under an overarching representational framework. Other instances of predictions, however, require full-blown *explicit* representations, and are literally representational. Or so, at least, Gładziejewski (forthcoming) argues.

What matters here is that this proposal suggests that there are two rather different senses of predictions at play in predictive processing (see also Anderson and Chemero 2013). According to the first sense, predictions *are not* representations, although it is useful to treat them as if they were, by deploying the theoretical construct of an "implicit representation". According to the second sense, predictions *are* literal representations built using an internal generative model of the world. Provided with this distinction, one could divide predictive processing in two distinct "tiers". In the first tier, there are non-representational predictive processes, whereas in the second tier there are representational predictive processes. It is important to note that these two tiers are inhibited by different predictive mechanisms with different functional roles.

In the first (nonrepresentational) tier, predictions are realized by hardwired mechanisms whose main task is that of encoding signals efficiently. To do so, the physical shape of these mechanisms "embodies" some specific expectations. For example, retinal ganglion cells are wired so as to fire only when some discontinuity in the light intensity is detected, and in this sense, we can treat them as predicting uniform light intensity and as reporting only deviations

from the predicted values. Similarly, the fact that in the primary visual cortex neurons tuned to cardinal orientations outnumber neurons tuned to oblique orientations suggests that our visual cortex implicitly predicts natural orientation statistics (both examples come from Gładziejewski *forthcoming*: §3.1). What makes these cases instances of "first tier" prediction is the absence of any discretely identifiable representational vehicle (hence the need for *implicit* representations) and the fact that information flows through these mechanism in only *one* direction; that is, they do not exhibit the bidirectional flow of information due to the interplay of prediction and prediction errors (see Ch. 1: §2.3).

In the second (representational) sense, predictions are realized by mechanisms which can alter their state as new information comes from the *bidirectional* flow of predictions and prediction errors. Here, there is an easy to individuate vehicle of the prediction (i.e. the state of the mechanism) and predictions are not used only to encode signals, but also to update one's estimates concerning the sources of the sensory signals (Gładziejewski *forthcoming*: §3.2).

Now, given the difference of mechanisms involved, it is reasonable to suppose that conclusions established in regard to tier-one prediction will not transmit to tier-two prediction (and *vice versa*). Hence, if the networks here considered were a case of tier-one prediction, Gładziejewski's argument would block the generalization I'm here advocating for. So, the relevant question is: are Bovet's networks capable of tier-two prediction, or are they only capable of tier-one prediction?

I'm not sure. On the one hand, Bovet's networks are non hierarchical, so there is no *bidirectional* flow of information.[201] This would put them in tier-one. However, Bovet's networks are not hardwired, they do *not* (only) encode signals efficiently, they update their estimates based on the errors they receive, and there are clearly identifiable states (such as

[201] This is not entirely true, as there still is a bidirectional *horizontal* flow of information (modality a to modality b *and vice versa*). It's not clear whether Gładziejewski will consider it bidirectional in the relevant sense. I assume he won't, so as to make his case as strongly as possible.

activation vectors) that carry predictions and prediction errors. This seems to put them into tier-two. So, unless hierarchy were *necessary for* tier-two prediction, Bovet's networks seem to fall more in the tier-two camp than in the tier-one camp. Hence, even adopting Gładziejewski's distinction, the generalization I'm proposing does not seem to be blocked. It *would* be blocked if adding hierarchy were to substantially modify the representationalist credentials of PP systems, and I've already argued (§5.2) above that this is *most likely* not the case.

Now, I'm willing to concede that the judgment on whether Bovet's networks realize tier-one or tier-two predictions might still be uncertain. But I'm equally willing to offer some reasons *not* to adopt Gładziejewski's proposed partition between the "two tiers".

5.5.1 - The distinction between the two tiers does not generalize well

One reason not to adopt it is that it does not provide clear cut distinction in many cases. Indeed, it seems to apply *only* to Gładziejewski's chosen example! As seen above, the partition does not apply particularly well to Bovet's network - but that is far from the only case in which the partition does not apply neatly (if at all applicable)

Consider, for instance, the networks proposed by Spratling (2016) to show that PP is able to simulate empirical data collected from humans engaged in sophisticated cognitive tasks (including categorization, context dependent task switching and naive-physics reasoning), thereby suggesting that PP can accounts for these sophisticated cognitive domains. The network Spratling proposes are hierarchical and token easily identifiable prediction states, *but are also hardwired*. Would this make their predictions tier-one or tier-two predictions? As far as I can see, Gładziejewski's framework offers no crisp answer to this question. Consider further models of PP relying on continuous-time recurrent neural networks (see Tani 2016: Ch. 9-10). The connections in these networks are hardwired; and to change them modellers typically rely on genetic algorithms, and so their change occurs only through simulated

generations of networks (e.g. Harvey *et al.* 1997). Does this make these models hardwired? Presumably yes, as even the number of cells in the visual cortex sensitive to oblique orientations or the lateral connections of retinal ganglion cells can change through generations, and these are the examples Gładziejewski offers of hardwired mechanisms. But then, does this mean that all PP models using continuous time recurrent networks are capable only of tier-one prediction? These networks are very powerful, and capable of mimicking complex cognitive capacities - capacities a representationalist would typically like to indicate as genuinely representational. Again, it seems that Gładziejewski's framework offers no crisp answer in this case.

5.5.2 - Explicitating the troubles with implicit representations

A second reason not to adopt Gładziejewski's framework is that it presupposes a workable distinction between implicit and explicit representations. Yet, no such distinction is offered. When it comes to explicit representations, Gładziejewski characterizes them referring back to his own work on generative models as structural representations (Gładziejewski 2016), and I have already argued at length that the argument offered there is flawed and cannot be easily ameliorated (Ch. 4). As for implicit representations, Gładziejewski's characterization is far too liberal. In fact, he seems to adopt a conception of implicit representations based on Dennett's (1978: Ch. 6) example of the chess-playing program (see Gładziejewski *forthcoming*: 10-11). Without entering into the details of the example, Dennett's point is that *sometimes* it is legitimate to ascribe belief and/or desires to systems which have no corresponding internal states, if this allows us to interact better with said systems. Gładziejewski elaborates the idea as follows:

> "Although there is no localized, separable, causally active internal vehicle that bears this content, the overall dispositional pattern of the program's behavior embodies the desire implicitly. [...] on this notion, (implicit-)representational ascriptions are grounded in the fact that a given

> information-processing system is wired in a way that allows it to embody dispositions to behave or respond 'as if' it (explicitly) represented the world t be certain way" (Gładziejewski *forthcoming*: 11)

There are several points worth unpacking here. First, the passage oscillates between implicit representations as dispositional patterns within systems and representational *ascriptions*. But these are not the same things: representational ascriptions typically ascribe *explicit* representations[202], and dispositions within a system (assuming that they are the vehicles of implicit representations) can be there even in absence of any ascription.

Notice also that to equate implicit representations to ascriptions makes their content non-natural: ascriptions require ascriptors, which are presumably intentional systems having contents in their own rights. Since I doubt that Gładziejewski wants to allow non-natural representations in his framework, I will only focus on the dispositional patterns, equating them with implicit representations.

Now, it is safe to say that implicit representations cannot be simply identified with dispositions (Ramsey 2007). Salt is soluble in water, but its microphysical structure does not represent "solubility". Wood and alcohol are inflammable, but it would be peculiar to say they represent flames.

The same holds if we restrict the scope of the claim to disposition *of information processing* systems: every information processing system has a myriad of dispositions that cannot possibly be (or ground) any implicit representations. My brain has the disposition to fry, if subjected to appropriately high temperatures. But surely it does not represent *fryability*. It also has the disposition to splat, if hit with sufficiently high forces, but it is very hard to claim my brain implicitly represents "splattability".

The same still holds if we restrict the scope of the claim to disposition of information processing systems *that causally contribute to the intelligent behavior of such systems*, for this

[202] Indeed, in Dennett's original example, what gets ascribed are full blown personal level propositional attitudes.

set still includes an inordinate amount of dispositions, and many of them are presumably non representational at all. For example, the slippery aluminium "paws" of the robot dog *Puppy* causally contributes to *Puppy*'s stabilization and gait change (Pfeiffer and Bongard 2007: 98-99), but it would be far-fetched to say that *Puppy*'s paws being made of aluminium represents anything.

Perhaps, then, the claim should be understood as the claim that dispositions of the *processors* of information in information processing systems are implicit representations. But, again, the claim overgeneralizes: processors have all sorts of dispositions (e.g. breakability), and even the ones that *matter* for their information processing capacities (e.g. channel capacity, degrees of freedom of their input gates, memory capacity, *etc*.) are not all obviously representational.

So, it seems that in order to make the claim that implicit representations are dispositions tenable we need some further constraint. As far as I can see, Gładziejewski provides none.

To conclude: Gładziejewski's proposed distinction between tier-one and tier-two predictions is, *prima facie*, unable to block the generalization I'm proposing. Moreover, at least as Gładziejewski articulates it, the distinction between tier-one and tier-two prediction is a bit nebulous, and does not give clear cut results in many cases. Lastly, the conceptual bedrock Gładziejewski relies on to articulate the distinction is not solid, as it provides no tenable conception of an implicit representation. These, I think, are compelling reasons not to accept the proposed distinction between tier-one and tier-two prediction in the first place.

***

There can be other arguments by means of which one might try to resist the sweeping generalization I'm proposing here. But, at present, I can think of no such argument.[203] Hence, as things stand, I think it is safe to conclude that the generalization I'm proposing holds. In the

[203] Nor could the two anonymous reviewers of the journal *Synthese*.

following, I will thus consider and allay some worries my claim can raise.

6 - Allaying some worries

Thus far, I've argued that the anti-representationalist verdict I provided examining Bovet's networks is likely to generalize to other PP systems. But my anti-representationalist verdict might raise other worries. For example, isn't the argument I just provided extremely dependent upon a sectarian and idiosyncratic understanding of what cognitive representations are? And, even supposing that my argument is on the right track, does it imply that PP is a radically revisionist theory of cognition? Am I suggesting that global anti-representationalism is correct? In this section, I will briefly consider a number of similar worries, and try to allay them all.

6.1 - Is this a "Hegelian argument"?

Chemero (2009, Ch. 1) notices a recurrent argumentative pattern in cognitive science. First, a research program or modeling tool is proposed, and starts gaining attention. Then, it is thoroughly analyzed (typically by researchers hostile to it), and it is concluded that it *cannot possibly* provide a satisfactory account of its explanatory target, and that it is therefore best abandoned. Strikingly, Chemero notices, such analyses are made on entirely *a priori* grounds, and are supported by little to no empirical evidence. Moreover, they typically rely on *contentious assumptions*, which are often not endorsed by the proponents of the research program/modeling tool under attack. Hence, they systematically fail to persuade their target audience and/or advance the discussion. Chemero dubs these arguments "Hegelian arguments", and urges us *not* to make them. Haven't I just made one? Here's two reasons to answer negatively.

First, the argument I offered does not target a research program or modeling tool. It considers the *philosophical interpretation* of PP, suggesting that such a *philosophical*

interpretation is best abandoned. But the philosophical interpretation of a theory should not be confused with the theory. Compare: one could attack a specific *philosophical interpretation* of certain chemical processes, suggesting that they do not support strong emergentist claims but only weak emergentist ones. Such an argument surely won't force us to revise our practice of chemistry: whether salt strongly or weakly emerges from sodium and chloride, salt would remain NaCl, it would retain all of its familiar properties and it would still respond in the exact same way to experimental manipulations.

Secondly, and, I believe, most importantly, the premises of my argument are not contentious. In general, their acceptance is widespread, and the structural-representationalist account of PP endorsed them all. Perhaps one could argue that the third necessary condition I imposed is contentious, since not everyone endorses computationalism. However, PP is a neurocomputational theory, and examining it at least *prima facie* presupposes a commitment to computationalism. Weren't computationalism true (in some form or another), we would have little reason to care about PP.

One might further argue that endorsing computationalism does commit one to the existence of mathematical contents. Given that the third condition I imposed is spelled out in terms of mathematical content, then it is contentious even if computationalism is assumed. Yet, although it is correct to say that endorsing computationalism does not in and by itself commit one to the existence of a special kind of content (mathematical content), the third condition can be rewritten dropping mathematical content in favor of *computational roles*. And surely every computationalist must admit that, if a candidate vehicle bears representational content, then that content must be *at least coherent* with the computational role of the vehicle. The "contentiousness" associated with mathematical contents is thus both very minimal and easily eliminated.

I thus conclude that the argument I provided is not a "Hegelian argument".

6.2 - "two-level attribution" *versus* non-representationalism

My arguments against Wiese's (2018) appeal to Eliasmith's theory of content *presuppose* that vehicles can be assigned contents in only one way (§ 4.2). But what if, as a reviewer of the journal *Synthese* asked, 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 and respond to different explanatory aims, they need not be mutually exclusive. Such a "two-level attribution"[204] of content can thus allow us to follow Wiese's suggestion, without *thereby* inviting the problems I raised before. What could be said in response?

To start, I wish to point out an ambiguity. Talking of "attributing content" is ambiguous between two readings. On a first reading, content attributions are not *mere* ascriptions: the vehicle *really bears* multiple contents 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. On a second reading, content attributions 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: 109).

I wish to consider both readings, and to argue that each reading yields an outcome

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

unfavorable to the "two-level attribution" view.

Consider, first, the realist reading. There is a sense in which the realist reading is deeply compatible with the structural-representationalist reading of PP. In fact, the structural-representationalist reading of PP *already* ascribes multiple contents to vehicles, as it claims that vehicles have both *mathematical* and *representational* contents. When one is interested only in the computational goings-on inside a PP system, one will be concerned just with the mathematical contents associated with its computational structure. This happens, for instance, when researchers strive to determine whether prediction error signals are computed by division or by subtraction (see Spratling 2017), or when they wonder whether the probabilities represented in PP systems (if any) should be interpreted as Bayesian priors or in simpler information-theoretic terms (e.g. Thornton 2017; 2020). Conversely, when one is interested in how PP systems interact with the environment, one needs to be concerned with the representational contents present in PP systems (if any). This looks like a "two-level" attribution of the kind suggested above.[205]

Now, if such a "two-level attribution" works, then why can't one hold that vehicles have multiple contents in virtue of the fact that they satisfy multiple content-grounding relations, and simply pick up the content which is most relevant given one's explanatory interests?

Because there is a problem with the "two-level attribution" thus interpreted. Suppose V jointly satisfies 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*. According to a realist "two-level attribution", V *really and objectively* represents T *as well as* T*. Thus V has *two* representational contents, and we are free to "pick one" based on our explanatory needs.

[205] And perhaps it is, but an important difference should nevertheless be noticed. Mathematical and representational contents are different kinds of content (Egan 2014: 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" here examined assigns different contents of the same kind (representational) to the same vehicle. And it does so by appealing to two (intuitively competing) content-grounding relations, rather than a content-grounding relation and the computational profile of the vehicle.

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:

> *Case 1*: A tokening of V is a misrepresentation when T, and only T, is not the case (*mutatis mutandis* for T*)
>
> *Case 2*:A tokening of V is a misrepresentation when at least one among T and T* is not the case
>
> *Case 3*: A tokening of V is a misrepresentation when both T and T* are not the case

Which is the correct one? For my purposes here the answer does not matter, because considering each case more closely makes evident that all three options end up assignin one, and *only* one, content to V.

If *case 1* is correct, 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 *case 2* is correct, 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 *case 3* is correct, 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[206], the "two-level attribution" view entails that V does not have *multiple* contents, but only a single (perhaps disjunctive or "conjunctive") content. Moreover, in two cases out of three, that content is *not* determined by *any* of the theories of content

[206] A reader might wonder why I have not considered option (b) when considering Wiese's proposal. The answer is that I did so merely for ease of exposition. 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*). And, as far as I can see, there is no principled reason to choose one option over the other.

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 now 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 one considers *case 2* or *case 3* as the correct case, 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 an objectivist and realist reading of the "two-level attribution" view is untenable. There might be a convincing reply to the argument I have just put forth. But, as far as I can see, such a reply has still to be provided. At present, then, the realist and objectivist reading of "two-level attribution" view does not really seem viable.

Does the content pragmatist reading of the "two-level attribution" fare any better? I doubt it, because content pragmatism strikes me as a form of anti-representationalism in disguise. Whilst it is surely true, as Mollo (2020: 108) rightfully notices, that content pragmatism differs from content eliminativism (i.e. anti-representationalism) because content pragmatism holds that content *cannot* be eliminated from cognitive-scientific explanations, content pragmatists do not take content ascriptions to be *part* of these explanations. Rather, content pragmatists interpret them as a strictly speaking unnecessary gloss over cognitive-scientific explanation proper (Egan 2014: 127-128; 2020: 33-34; Mollo 2020: 105). On their view, ascriptions of content only provide a "user-friendly", *but explanatory idle*, tool to highlight, in an intuitively perspicuous manner, how the internal goings-on of a computational system relate to the environment. Content pragmatists are thus *explanatory anti-representationalists*: they hold that, strictly speaking, cognitive-scientific explanations do not posit representational vehicles

bearing representational contents. Content pragmatists are also *metaphysical anti-representationalists*: they hold that there is no fact of the matter about what vehicles represent, because representational contents are not really "in the system", but are merely ascribed from an external observer.

In sum, content pragmatists hold that representations are not *really* posits of our cognitive-scientific theories, that representations are not *really* necessary to the explanation of our cognitive capacities, and that representational contents are not "*really* in there". How this position qualifies as a form of representationalism is, for me, a mystery.[207]

6.3 - Aren't generative models still representations in some sense?

One could object that my argument relies on a too narrow conception of representation. Perhaps a representationalist can accept some degree of content indeterminacy (e.g. Ramsey 2020: 57-58). Perhaps purely proximal contents are fine. Perhaps not all vehicles need to bear an exploitable structural similarity with their targets.[208] At any rate, my too narrow conception of representation pushed me to impose necessary conditions that are *just too strict*. So, I need to relax them. And once those are relaxed, my anti-representationalist verdict might no longer hold.

So, should I relax these three conditions? I think the answer is negative.

To start, these three conditions are typically accepted, as I have repeatedly argued. The claim that they need revision needs to be argued for. But, as far as I can see, there are few, if any,

[207] A reader might object that, at least as articulated in Egan (2014; 2020) content pragmatism is committed to a form of explanatory and metaphysical representationalism in regards to representations of mathematical contents. As a matter of historical reconstruction, the observation is correct. Yet, as both Ramsey (2020) and I (Facchin *submitted*) have argued elsewhere, a combination of realism for mathematical contents and antirealism for representational contents is not a stable position: it "slides into" full-blown content realism or content irrealism.

[208] Notice that albeit I never found the previous two claims articulated in the literature, representationalists *do often argue* that anti-representationalists operate with a too strict conception of representation; see (Clark and Toribio 1994; Clark 2015a, 2015b; Williams 2017).

arguments to that effect.[209] Indeed, even Ramsey (2020: 57-58) does not provide any *argument* to think that some degree of content indeterminacy is acceptable. He simply states that it is. And that is fine, given that his objective is that of arguing that, *even if present*, content indeterminacy *wouldn't* be a problem for representational realism.

Yet, perhaps, reasons to reject, or weaken, these three conditions *could* be provided; and the fact these conditions are widely accepted does not guarantee their truth. However, I believe that it is *highly unlikely* that these reasons *could* be provided, even in principle. In fact, it seems to me that weakening even one of my three conditions has very nasty consequences, which are at least *very hard* to accept.

Consider, first the distality and determinacy condition. Could it be outright rejected? I doubt it. To outright reject it is to reject that content must be determinate. But content determination is constitutively connected with the possibility of misrepresentation (Ch. 2: §2.2; §6.2 above). If content is radically indeterminate, then misrepresentation becomes problematic or impossible. But this is a problem for the representationalist, given that the obtaining (or non obtaining) of a representation's conditions of satisfaction is what should explain the non-accidental success (or failure) of a system (e.g. Godfrey-Smith 2006; Shea 2018; Gładziejewski and Miłkowski 2017). Accepting that content can be indeterminate *cripples* the explanatory power of representations.

What about distality, then? Could we accept that representational content is purely proximal? The answer seems negative for similar reasons. Content is supposed to meaningfully connect the internal goings on of a system with the environmental contingencies relevant to that system (e.g. Egan 2014; 2018; Shea 2018: 31-36). This is why content is explanatory powerful. But proximal contents *do not* connect a system with its environment: they only

[209] This is no longer true: Bergman (2021) *has* offered an explicit argument to the effect that content indeterminacy is not a problem for sub-personal cognitive representations (as opposing to mental ones). Owing to space limitations, I will not discuss it here. Yet, in my view, the considerations I offer in the rest of the paragraph are sufficient to resist Bergman's argument.

connect a system with its internal and peripheral states.

What if the determinacy (or distality) requirement were just *weakened*, rather than rejected? Perhaps content need not be *entirely* determinate or distal for representational explanations to work. The problem I see with this suggestion is that distality and determinacy are not *graded* properties: they are all-or-nothing properties. If we accept that there is no way to determine whether, say, the fuel gauge of a car indicates the amount of gasoline in a tank, or the amount of liquid in the tank (or the actual height of float in the tank), how can we put an end to the disjunction? How can we exclude that the gauge indicates (the amount of gasoline or the amount of liquid or the height of the floater or the amount of air in the tank or the state of the connected potentiometer or the state of some other intermediate component in the system)? As far as I can see, there is no way to exclude this long and cumbersome disjunction. And once such a disjunction is in place, the system's ability to misrepresent seems compromised.

Moreover,it should be noted that allowing contents to be indeterminate or proximal does not provide a good fit with the empirical practice of cognitive science. When cognitive scientists mention contents, these contents are typically determinate and distal. So, if the point of the philosophical debate on representations is to analyze and account for the kind of representations invoked in the explanatory practice of cognitive science, allowing contents to be indeterminate and proximal *amounts to* changing the end goal of the whole endeavor.

Couldn't, then, the second requirement be rejected or weakened? Perhaps exploitable structural similarity is not the relevant relation that will yield us determinate and distal contents. It is hard to see how it could be weakened. I've already deployed the weakest notion of structural similarity on offer. Moreover, the less demanding the notion of structural similarity, the more content determinacy is problematic. And the same problem makes weakening exploitability even less attractive: as things stand, *exploitability* is what makes the content of structural similarity-based theories determinate (Shea 2018: 119-126). Thus, weakening the

exploitability requirement invites indeterminacy.

What about rejecting the second requirement, then? I think this might be a promising move, provided that one has a strong theory of content at one's disposal. Yet, as I have been at pains to argue in the previous chapter, our strongest alternatives to structural similarity-based theories of content (teleo-informational theories) require a vehicle-target exploitable structural similarity too. So, even if one wants to avoid reducing teleo-informational theories of content to structural similarity based theories of content, the second best option suffers from all the same problems of the first best option. And, as far as I can see, there is no further alternative: a robustly realistic interpretational semantics either makes content proliferate uncontrollably (Cummins 1991) or collapses into structural similarity (as shown in Shagrir 2012). Consumer-based teleosemantics is turning out to be just another variation on structural similarity (Millikan 2020).[210] Functional role semantics seems unable to account for misrepresentation (see Cummins 1996: 29-53). Purely informational accounts of content are plagued by the challenges of distality and determinacy (Artiga and Sebastian 2018; Rosche and Sober 2019). As far as I know, there is no other naturalistic theory of content on the market.

Lastly, one could weaken or abandon the third requirement, namely that mathematical contents (or computational role) must place some constraint on representational contents. But it is hard to see how it can be weakened, given that all that constraint requires is that the two must be *at least coherent*. And, as argued above (§ 2.3) if computational role and representational content part ways, we are forced to choose between the two. If computational role and representational content do not cohere, we are forced to *either* trust the computational description of the system *or* our theory of content of choice.

[210] It might be worth noting, as an historical aside, that Millikan (1984) already partially defined content in terms of vehicle-referent *mappings*, and that her explicit aim was to defend a *picture theory of language* (and signs more generally). Indeed, she defended a *wittgenstenian* picture theory of language, according to which: "The fact that the elements of a picture are related one another in a determinate way represents that things are relates one another in the same way" (Wittgenstein 1921/2013 §2.15). This is precisely how structural representations represent. On Millikan's picture theory and its connections with PP, see (Sachs 2018).

For all these reasons, I conclude that my anti-representationalist verdict does not hinge on a too narrow conception of representations, and that generative models are not "in some sense" still representations.

6.4 - Does my anti-representationalist verdict entail a radical revision of cognitive science?

My anti-representationalist verdict might be taken to entail a strong form of revisionism in regard to the explanatory practices and/or lexicon of cognitive science.[211] So, is my verdict revisionist? And, if yes, to what extent? The answer to the first question is positive: anti-representationalism *is* a revisionist position. But, it is not *radically* or *dangerously* revisionistic. Or so, at least, I want to argue.

The explanatory practices and the explanatory lexicon of cognitive science surely *seem* strongly committed to representationalism. Anti-representationalism is thus a revisionary position. But the revision might not be as deep as it might *prima facie* appear.

To start, albeit in cognitive-scientific explanations it is commonplace to use the word "representation", it is not always clear whether that word *designates* representations. Ramsey (2007), for instance, rightfully notices that many cognitive scientists routinely use the word "representation" to refer either to simple causal mediators or dispositions of cognitive architectures. Jacobson (2003; 2013) notices that cognitive neuroscientists often refer to *instantiations* as representations. She points out, for instance, that a cognitive neuroscientists might refer to a pattern of activation of the amygdala as *representing* a subject's fearful state, while plausibly intending to claim that activations of the amygdala *instantiate* the subject's fearful state. Cognitive scientists are starting to notice this issue, as well as the theoretical problems connected to it, and are striving to incorporate a philosophically robust notion of

[211] Thanks to an anonymous reviewer of *Synthese* for having raised these questions.

representation in their explanatory practices (e.g. Brette 2019; Poldrack 2020; Backer, Lansdell and Kording 2021). Although I sincerely applaud this effort, I cannot help but notice that it *at least implicitly admits* that many non-representational structures have in fact been wrongly labeled as representational in the past years. Hence, cognitive-scientific explanations might not be *as* reliant on representations *as* their expression in public language suggest.[212]

Notice that such an over-usage of representational terminology affects PP too. In the PP literature, the term "model" has been used to refer to: (a) the whole brain (see Ch. 4: §§ 4.7, 4.8), (b) axonal connections (Ch.4: §4.4), (c) functionally specializes neural circuits (e.g. the mirror neuron system as body-model, se Kilner, Friston and Frith 2007), (d) neuronal responses and connections (Buckely *et al*. 2017), single hierarchical layers anatomically individuated (e.g. Kiefer and Hohwy 2019: 387), (e) the alpha motor neurons of the spinal cord that directly innervate muscles (Friston 2011: 491), and (f) single neurons (Palacios *et al*. 2019). It is *at least plausible* that not all elements in (a) to (f) are considered representations by PP theorists, even if they normally use the term "model" when referring to them.[213]

At this point, then, it is important to notice that the claim that models (both generative and inverse) are *essentially controllers* mediating agent-environment interactions is commonplace both in the literature on the free-energy principle[214] and PP (e.g. Seth 2015; Baltieri and Buckley 2019; Kirchhoff and Kiverstein 2019; Corcoran, Pezzulo and Hohwy 2020). In the theoretical vocabulary PP deploys, "model" primarily denotes control structures. Given the "good regulator theorem" (Conant and Ashby 1970), such control structures must be homomorphic to what they control; namely the generative process.[215] This much is accepted

[212] It is also worth noting that some cognitive scientists are explicitly anti-representationalists, see, for instance, (Chomsky 1995).
[213] Indeed, one of PPs architects, namely Karl Friston, seems to endorse some form of anti-representationalism. See (Ramstead, Kirchhoff and Friston 2019).
[214] See, for example (Bruineberg and Rietveld 2014; Bruineberg, Kiverstein and Rietveld 2018; Baltieri, Buckley and Bruineberg 2019; Tshantz Seth and Buckley 2020). Notice also that, due to space limitations, I will not discuss the complex relationship between the free-energy principle and PP.
[215] Does this entail that there is a mistake in Ch. 4, and that generative models *really* are structurally similar to

both in my account and the structural-representationalists' ones. My account part ways from the structural-representationalist ones when they claim that, models-as-controllers are also models-as-structural-representations; that is, when the structural-representationalist claims that *on top of* the homomorphism entailed by the law of requisite variety there is *also* a content-constituting exploitable structural similarity holding between the controller and some well determined and distal worldly target, such that the model-as-controller is *also* a structural representation of that target (e.g. Pezzulo 2017; Williams 2018a: 117-124).[216] Thus, my claim that generative models are non-representational structures instantiating an agent's sensorimotor mastery is entirely compatible with the theoretical commitments of PP. Thus, no revision of the theory is forced here.

Noticing that my anti-representationalist verdict is also entirely compatible with PP's commitment to models-as-controllers also has the addit benefit of enabling us to *continue using* a model-based vocabulary when discussing matters related to PP, without having to adopt a fictionalist (or otherwise irrealist) stance towards models. This, I think, is a very positive result. We avoid a painful revision of our scientific *vocabulary*: we are not forced to substitute every occurrence of "model" with some cumbersome, *ad hoc* expression. We also avoid an equally painful, and conceptually suspect (see Sprevak 2013; Ramsey 2020) revision of the *usage* of said vocabulary: we can still *use* the word "model" to refer to objectively present and robustly real models-as-controllers (rather than mysterious posit of some fictionalist/irrealistic framework).[217]

---

the environment, and thus that a structural-representationalist reading of them is warranted? No, it does not. For one thing the generative process is *not* the environment, unless "the environment" designates everything but the brain. The generative process also includes an agent's active body, as well as physiological bodily states. And I have *never* denied that the generative model is structurally similar to the generative process (see Ch. 4: §4.7). I've only argued that such a structural similarity is insufficient to substantiate a structural-representationalist reading of generative models.

[216] Notice that this does not completely exhaust the additional commitments of the structural-representationalist reading of generative models. As seen in Ch. 4, this reading is also committed to the model being *decouplable form* and *allowing for representational error detection in regard to* its target.

[217] What, then, about *inference*? PP is, at some level of description, committed to a view of cognitive systems approximating Bayesian inferences. But inferences require representations; perhaps even language-formatted

Notice further that the anti-representationalist verdict I am proposing is compatible with (and, indeed, motivated by) the current explanatory practices of cognitive science. As discussed above (§ 6.1) my anti-representationalist verdict does not hinge on an idiosyncratic understanding or representations. Indeed, the three necessary features I have listed are fully compatible with the philosophically rich notion of representations that is being deployed *by cognitive scientists* (as opposed to *philosophers of cognitive science*, see e.g. Brette 2019; Poldrack 2020; Backer, Lansdell and Kording 2021). Moreover, I reached my verdict by examining an artificial neural network, and artificial neural networks are surely central in the current empirical practice of cognitive science. So, my verdict is not fueled by, nor promotes, alternative research programs in cognitive science, such as ecological psychology (e.g. Kelso 1995; Chemero 2009), or some variant of enactivism (Hutto and Myin 2013; 2017).

These, it seems to me, are solid reasons to believe that my anti-representationalist verdict does not in any way force us to strongly revise the lexicon and/or the explanatory practices of cognitive science as it is currently practiced.

6.5 - Does my anti-representationalist verdict entail a radical revision of our self conception?

The last worry I wish to consider (and allay) is that my non-representationalist verdict might be taken to entail some radical form of revisionism regarding our own self-conception as rational agents. Nicholas Shea articulates this worry beautifully when he writes:

> "Some want to eliminate the notion of representational content from our

ones. So, it seems I'm forced to be a revisionist about the *inferential* lexicon of PP. Yes, but revisionism about inferences is not nearly as painful as revisionism models. For one thing, connectionists already revised the concept of inference equating it to the concept of "network reaching a stable state" (see McClelland, Rumelhart and the PDP research group 1986). That is a concept of inference I'm simply free to adopt. Moreover, the structural-representationalist reading of PP *is forced to revise the commitment to inference too*. Here's three examples. Clark adopts a notion of inference-as-action-selection (see Clark 2016: 15; 176-194). Hohwy (2019: 200) candidly admits that, to be literally inferential, PP must utilize a revised notion of inference with a far broader extension. Kiefer, claims that even simple systems such as bacteria are literal inference engines (Kiefer 2020: 7). It seems clear that in all these instances "inference" does not denote a truth-preserving transition between lingua-formatted states expressing propositions.

> theorizing entirely, perhaps replacing it with a purely neural account of behavioural mechanisms. If that were right, it would radically revise our conception of ourselves as reason-guided agents since reasons are mental contents. That conception runs deep in the humanities and social sciences, not to mention ordinary life." (Shea 2018:6)

Does my anti-representationalist verdict force us to revise our conception of ourselves as rational agents guided by reasons? I think the correct answer to this question is: no more than the structural-representationalist reading of PP.[218]

Indeed, as far as I can see, nowhere my argument entails the form of content nihilism that would force us to abandon our conception of ourselves as agents guided by reasons. In fact, the claim that generative models are non-representational structures does not, *by itself*, entail a form of global anti-representationalism about the *mind*. It does not even entail, by itself, global anti-representationalism about *cognition*. There are two reasons as to why this is the case.

One is that the claim has no *direct* entailment when it comes to personal-level mental content (e.g. what I recall when I recall the melody of my favorite song; what my beliefs and desires are about). In general, we should be wary of projecting the sub-personal level onto the personal level and *vice versa* (Dennett 1991; Hurley 1998). The content of personal-level mental states might, *but need not*, have a sub-personal level counterpart. Hence, the fact that generative models are non-representational structures (and are thus devoid of content) need not imply that the personal level is devoid of content too. And our self-image as rational agents sensitive to reasons is an image describing us at the *personal* level. Hence, it is not threatened by the claim that generative models are non-representational structures.

The other is that the claim that generative models are non-representational structures does not *by itself* entail that the sub-personal level is devoid of such structures. To be sure, proponents of PP often claim that PP is an account of cognition *in general*, and that it can explain all cognitive processes (e.g. Friston 2009, 2010; Hohwy 2015; Clark 2016; Spratling

[218] On the relationship between PP and folk-psychology, see (Dewhurst 2017).

2016; Pezzulo 2017). But the evidence for this claim is far from conclusive. Surely, the simulations detailed in (Spratling 2016) show that *handcrafted* PP models can master a variety of cognitive domains *in isolation*. But these data do not entail that a *single, generic* generative model can master a variety of cognitive domains *at once*. Moreover, the only PP architecture for general cognition I know of (Ororbia and Kelly 2021)[219] introduces a declarative memory module and a working memory module that *are not* generative models. Those might be representational structures - nothing in the arguments here offered imply that they are not.

Moreover, PP suffers from an at least *prima facie* problem when it comes to account for cognitive processing targeting cognitive domain with no clear sensory manifestation (Williams 2020). It is hard to see how a generative model of sensory observations could be used to cognize about justice, or the number 27, or the law of excluded middle. Williams considers two possible ways in which PP could be expanded so as to account for these cognitive processes. One way is that of modifying the cognitive architecture I described in Ch.1, perhaps adding some specialized components, like a module for "abstract thought". The other is to claim that "abstract thought" depends on our mastery of cultural practices, and in particular linguistic ones (Clark 2016: Ch. 9; Fabry 2015; 2018). It has long been speculated that the mastery of natural languages can modify one's cognitive architecture, installing a "virtual machine" capable of entirely novel cognitive processes (Dennett 1991; Clark 1993, Ch. 8).

Now, to recommend which strategy to follow lies far beyond the scope of the present treatment. The only thing I wish to highlight is this: that in the theoretical space lying between outright modifying the standard PP architecture and specifying how to install a virtual machine on it, it is *at least possible* that genuine cognitive representations will be encountered.

Thus, *in and by itself*, my anti-representationalist verdict does not entail, nor invite, content

[219] Importantly, as things stand, such an architecture still needs empirical validation. At present, the architecture is just a blueprint.

nihilism.

Conclusion

In this dissertation, I have examined the metaphysical status of generative models in the PP framework. I have argued that they are not structural representations. If I'm right, they are best understood as *non-representational control structures*, whose main task is that of coordinating agent-environment interactions.

My claim has been defended as follows. First (Ch.4), I've argued we presently lack any compelling reason to think generative models qualify as structural representations. For, structural representations are partially identified by the similarity tying them to their target, and there's no reason to believe generative models bear such a similarity.

And, even if such a reason were provided, we *still* would lack any compelling reason to think generative models are structural representations. For, as currently characterized, their functional profile is not a *representational* functional profile (Ch. 5).

In fact, scrutinizing a simple connectionsit implementation of a generative model (Ch.6) readily shows generative models being nonrepresentational structures instantiating an agent's sensorimotor skills. And this verdict, I argued, can be easily generalized well beyond the case here directly observed.

Suppose I'm correct: *where from here*? What's the natural next step of this line of inquiry? Interest is rising about PP as a theory of *phenomenal consciousness* (e.g. Seth 2021). And the arguments provided here can be readily put in contact with phenomenal consciousness. After all, the line of thought sketched here is broadly consonant with sensorimotor enactivism, which *just is* a theory of phenomenal consciousness.

Yet, being a closet-illusionist, *I don't care* about phenomenal consciousness. I care, however, about (cognitive) representations. And in Ch.6: §6.5 I've been a bit undecided on them. On the one hand, I've argued that my non-representationalism *can* face the challenge posed by " "representation hungry" cognition (Clark and Toribio 1994), which supposedly

keeps *global anti-representationalism* at bay. On the other hand, I've argued there are also reasons to resist the conclusion to *global* anti-representationalism. Maybe PP does not explain *everything* about the mind (e.g. Williams 2020), and to account for what PP leaves unexplained we will need to posit (or discover) representational structures.

Are these reasons to resist the conclusion to global anti-representationalism solid? I'm not entirely sure. Could generative models account for intelligence *writ large*? The question is largely empirical - but, *pace* Williams (2020), I'm not sure the answer is negative. On the one hand, his clever argument *cannot* attack non-representational versions of PP, as it is entirely predicated on generative models having an iconic *representational format*. On the other hand, some, admittedly very limited, empirical results paint what, to me, looks an unexpectedly rosy picture (Hua and Kunda 2020, Kunda 2021).

So, how far does the anti-representationalism defended here goes? Can we *really* cognize without representing *at all*? And if not, how do the representational and non-representational pieces of the thinking machinery interface with each other? These are questions I'd love to answer in the near future. Wish me luck.

References

Adams, R. A., Shipp, S., & Friston, K. (2013). Predictions, not commands: active inference in the motor system. *Brain Structure and Function*, 218(3), 611-643.

Adams, R. A., Stephan, K. E., Brown, H. E., Frith, C. D., & Friston, K. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, 4: 47.

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

Aikins, K. (1996). Of sensory systems and the aboutness of mental states. *Journal of Philosophy*, 93(7), 337-372.

Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & de Lange, F. P. (2013). Shared representation for working memory and mental imagery in early visual cortex. *Current Biology* 23(15), 1427-1431.

Anderson M. (2017). Of Bayes and bullets. An embodied, situated, targeting-based account of predictive processing. In T. Metzinger, W. Wiese (Eds). *Philosophy and Predictive Processing*: 4. Frankfurt am Main, The MIND Group. https://doi.org/10.15502/9783958573055.

Anderson, M., & Chemero, T. (2013). The problem with brain GUTs: conflation of different senses of "prediction" threatens metaphysical disaster. *Behavioral and Brain Sciences*, 36(3), 204.

Anderson, M., & Chemero, T. (2019). The world well gained: on the epistemic implication of ecological information. In M. Colombo, M. Stapleton, L. Irvine (Eds.), *Andy Clark and His Critics*.(pp. 161-173). New York: Oxford University Press.

Artiga, M. (2021). Strong liberal representationalism. *Phenomenology and the Cognitive Sciences*, https://doi.org/10.1007/s11097-020-09720-z.

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.

Backer, B., Lansdell, B., Kording, K. (2021). A philosophical understanding of representations for neuroscience. *ArXiv:* 2102.06592.

Baltieri, M. (2019). *Active inference: building a new bridge between control theory and embodied cognitive science*. Ph.D. Dissertation, University of Sussex. Accessed at http://sro.sussex.ac.uk/id/eprint/84970/ Last accessed 20/10/2020.

Baltieri, M., & Buckley, C. (2019). Generative models as parsimonious descriptions of sensorimotor loops. *Behavioral and Brain Sciences*. https://doi.org/10.1017/s0140525x19001353,e218.

Baltieri, M., Buckley, C., & Bruineberg, J. (2020). Predictions in the eye of the beholder: an active inference account of watt governors. *ALIFE 2020*, 121-129, https://doi.org/10.1162/isal_a_00288

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, 11(7), 280-289.

Bar, M. (2009). The proactive brain: memory for prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521), 1235-1243.

Bar, M., Kassam K. S., Ghuman, A. S:, Boshyan, J., Schmid, A. M., Dale, A. M., Hämäläinen M. S., Marnkovic, K., Schacter, D. L., Rosen, B. R., & Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, 103(2), 449-454.

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, 16(7), 419-429.

Bartles, A. (2006). Defending the structural concept of representation. *THEORIA – Revista de Theoria, Historia, Y Fundamentos de la Ciencia*. 21(1), 7-19.

Bastos A. M., Ursey, W. M., Adams, R. A., Mangun G. R., Fries, P., & Friston, K. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76(4), 695-711.

Bechtel, W. (2008). *Mental Mechanisms. Philosophical Perspectives on Cognitive Neuroscience*. New York: Routledge.

Bergman, K. (2021). Should the teleosemanticist be afraid of content indeterminacy? *Mind and Language*. https://doi.org/10.1111/mila.12395.

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

Bickhard, M. H. (2009). The interactivist model. *Synthese*, *166*(3), 547-591.

Blackmore, S. J., Wolpert, D. M., & Frith C. D. (1998). Central cancellation of self-produced tickle sensation. *Nature Neuroscience*, 1(7), 635-640.

Blackmore, S. J., Wolpert, D. M., & Frith C. D. (2000). Why can't you tickle yourself?. *Neuroreport*, 11(11), R11-16.

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

Bolstad, W. M., & Curran J. M. (2017). *Introduction to Bayesian Statistics* (3rd edition). Hoboken, N. J.: Wiley and Sons.

Bongard, J., Zykov, V., & Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science*, 314(5802), 1118-1121.

Boone, W, & Piccinini, G. (2016). the cognitive neuroscience revolution. *Synthese*, *193*(5), 1509-1524.

Botvinick, M., & Touissant, M. (2012). Planning as inference. *Trends in Cognitive Sciences*, 16(10), 485-488.

Bovet, S. (2006). Emergence of insect navigation strategies from homogeneous sensorimotor coupling. In *Proceedings of the 9thInternational Conference on Intelligent Autonomous Systems (IAS 9)*, pp. 525-533. Tokyo.

Bovet, S. (2007). *Robots with Self-Developing Brains*, Dissertation, University of Zurich https://www.zora.uzh.ch/id/eprint/163709/1/20080298_001884101.pdf. Accessed 25 February 2020.

Bovet, S., & Pfeifer, R. (2005a). Emergence of delayed reward learning from sensorimotor coordination, *Proc. IEEE/RSJ Int. Conf. On Intelligent Robots and Systems*, https://doi.org/10.1109/IROS.2005.1545085

Bovet, S., & Pfeifer, R. (2005b). Emergence of coherent behaviors from homogeneous sensorimotor coupling, *ICAR '05 Proceedings 12th International Conference on Advanced Robotics,* https://doi.org/10.1109/ICAR.2005.1507431

Braitenberg, V. (1984). *Vehicles. Experiments in Synthetic Psychology*. Cambridge, MA.: The MIT Press.

Brette, R. (2016). Subjective physics. In El Hady A. (Ed.), *Closed Loop Neuroscience* (pp. 145-170). London: Elsevier.

Brette, R. (2019). Is coding a relevant metaphor for the brain?. *Behavioral and Brain Sciences*, 42, e:215, 1-58. https://doi.org/10.1017/S0140525X19000049.

Brown, H., Adams, R. A., Parees, I., Edwards, M., & Friston, K. (2013). Active inference, sensory attenuation and illusion. *Cognitive Processing*, 14(4), 411-427.

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

Brown, H., Friston, K., & Bestmann, S. (2011). Active inference, attention, and motor preparation. *Frontiers in Psychology*, 2:218.

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 an optimal grip on a field of affordances. *Frontiers in Human Neuroscience*. 8: 559.

Bruineberg, J. Kiverstein, J., & Rietveld, E. (2018). The anticipating brain is not a scientist. The free energy principle from an ecological enactive perspective. *Synthese*, 195(6), 2417-2444.

Bubic, A., Von Cramon, Y., Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, 4: 25.

Buckley, C. L., Kim, C. S, McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: a mathematical review. *Journal of Mathematical Psychology*, 81, 55-79.

Bulow, P., Solodkin, E., Thagard, P., Eliasmith, C. (2015). Concepts as semantic pointers: a framework and a computational model. *Cognitive Science*, *40* (5), 1128-1162.

Cao, R. (2012). A teleosemantic approach to information in the brain. *Biology and Philosophy*, *27*(1), 49-71.

Cao, R. (2020). New labels for old ideas: predictive processing and the interpretation of neural signals. *Review of Philosophy and Psychology*, *11*(3), 517-546.

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*. Cambridge, MA.: The MIT Press.

Cheney, D. L., & Seyfarth, R. M. (1985). Vervet monkey alarm calls: manipulation through shared information?. *Behavior*, *94*(1-2), 150-166.

Chomsky, N. (1983). Mental Representations. *Syracuse Scholar*, 4(2), 2.

Chomsky, N. (1995). Language and nature. *Mind, 104*(413), 1-61.

Churchland, P. M. (1989). *A Neurocomputational Perspective*. Cambridge, MA.: The MIT Press.

Churchland, P. M. (2012). *Plato's Camera. How the Physical Brain Captures a Landscape of Abstract Universals*. Cambridge, MA.: The MIT Press.

Churchland, P. S. (1987). Epistemology in the age of neuroscience. *The Journal of Philosophy*, *84*(10), 544-553

Churchland, P. S., & Senjowski, T. J. (1992). Neural representation and neural computation. *Philosophical Perspectives*, 4, 343-382.

Ciria, A., Schillaci, G., Pezzulo, G., Hafner, V. V., Lara, B. (2021). Predictive processing in cognitive robotics: a review. *Neural Computation*, *33*(6), 1402-1432.

Clark, A. (1989). *Microcognition*. Cambridge, MA.: The MIT Press.

Clark, A. (1993). *Associative Engines. Connectionism, Concepts, and Representational Change*. Cambridge, MA.: The MIT Press.

Clark A. (1997). The dynamical challenge. *Cognitive Science*, 21(4), 461-481.

Clark, A. (2010). Memento's revenge: the extended mind, extended. In R. Menary (Ed.), *The Extended Mind*, (pp. 43-66). Cambridge, MA.: The MIT Press.

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

Clark, A. (2013b). *Mindware. An introduction to the Philosophy of Cognitive Science* (2nd edition). New York: Oxford University Press.

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

Clark, A. (2015b). Radical predictive processing. *The Southern Journal of Philosophy*, 53, 3-27.

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

Clark, A. (2016). *Surfing Uncertainty*. New York: Oxford University Press.

Clark, A. (2018). Strange inversions. In Huebner, B. (Ed.), *The Philosophy of Daniel Dennett*. New York: Oxford University Press.

Clark, A. (2020). Beyond desire? Agency, choice, and the predictive mind. *Australasian Journal of Philosophy*, 98(1), 1-15.

Clark, A., & Grush, R: (1999). Towards a cognitive robotics. *Adaptive Behavior*, 7(1), 5-16.

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

Conant, R. C., & Ashby W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systemic Science*, 1(2), 89-97.

Constant, A., Clark, A., & Friston, K. (2021). Representation wars: enacting an armistice through active inference. *Frontiers in Psychology*, *11*: 598733.

Corcoran, A., Pezzulo, G., & Hohwy, J. (2020). From allostatic agents to counterfactual cognizers: active inference, biological regulation, and the origins of cognition. *Biology and Philosophy*, *35*(3), 1-45.

Craik, K. (1943). *The Nature of Explanation*, Cambridge: Cambridge University Press.

Cummins, R. (1975). Functional analysis. *The Journal of Philosophy*, 72(20), 741-765.

Cummins, R. (1991). *Meaning and Mental Representation.* Cambridge, MA.: The MIT Press.

Cummins, R. (1996). *Representations, Targets and Attitudes*. Cambridge, MA.: The MIT Press.

Cummins, R. (2010). *The World in the Head*. New York: Oxford University Press.

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

Dayan, P. (2003). Helmholtz machines and sleep-wake learning. In M. A. Arbib (ed.), *The Handbook of Brain Theory and Neural Networks*, (pp. 522-525). Cambridge, MA.: The MIT Press.

Dayan, P., & Abbott, L. F. (2001). *Theoretical Neuroscience. Computational and Mathematical Modeling of Neural Systems.* Cambridge, MA.: The MIT Press.

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

Dayan, P., & Hinton, G. E. (1996). Varieties of Helmholtz machines. *Neural Networks*, 9.8, 1385 – 1403.

Degenaar, J., & Myin, E. (2014). Representation-hunger reconsidered. *Synthese*, *191*(15), 3639-3648.

Dennett, D. C. (1978). *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA.: The MIT Press.

Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, MA.: The MIT Press.

Dennett, D. C. (1991). *Consciousness Explained*. New York: Little Brown.

Dennett, D. C. (1996). *Darwin's Dangerous Idea*. New York: Simon Schuster.

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

Dewhurts, J. (2017). Folk Psychology and the Bayesian brain. In T. Metzinger and W. Wiese (Eds.), *Philosophy and Predictive Processing: 9*. Frankfurt am Main, The MIND Group. https://doi.org/10.15502/9783958573109.

Dewhurst, J. (2018). Individuation without representation. *The British Journal of the Philosophy of Science*, *69*(1), 103-116.

Donnarumma, F.,Costantini, M., Ambrosini, E., Friston, K., Pezzulo, G. (2017). Action perception as hypothesis testing. *Cortex*, *89*, 45-60.

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. (2017). Radical sensorimotor enactivism & predictive processing. In T. Metzinger, W. Wiese (Eds.). *Philosophy and Predictive Processing*: 10, Frankfurt am Main: The MIND Group, https://doi.org/10.15502/9783958573123.

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

Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge, MA.: The MIT Press.

Dretske, F. (1986). Misrepresentation. In Bodgan R. (Ed.). *Belief: Form, Content and Function*. (pp. 17-36). New York: Oxford University Press.

Dretske, F. (1988). *Explaining Behavior. Reasons in a World of Causes*. Cambridge, MA.: The MIT Press.

Dretske, F. (1994). The explanatory role of information. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*. 349(1689), 59-70.

Egan, F. (2010). Computational models: a modest role for content. *Studies in History and Philosophy of Science Part A*, *41*(3), 253-259.

Egan, F. (2012). Representationalism. In E. Margolis, S. Samuels, P. Stich (Eds.), *The Oxford Handbook of Philosophy of Cognitive Science*, (pp. 250-272). New York: Oxford University Press.

Egan, F. (2014). How to think about mental content. *Philosophical Studies*, 170(1), 115-135.

Egan, F. (2019). The nature and function of content in computational models. In M. Sprevak, M. Colombo (Eds.), *The Routledge Handbook of the Computational Mind*. (pp. 247-259). New York: Routledge.

Egan, F. (2020). A deflationary account of mental representation. In J. Smortchkova; K. Dołega, T. Schlicht (Eds.). *What are Mental Representations?* (pp. 26-53), New York: Oxford University Press.

Eliasmith, C. (2000). *How Neurons Mean: a Neurocomputational Theory of Representational Content*. Ph.D. Dissertation, Washington University in St. Louis, MO.

Eliasmith, C. (2005). A new perspective on representational processes. *Journal of Cognitive Science*, 6, 97-123.

Eliasmith, C. (2013). *How to Build a Brain*. New York: Oxford University Press.

Elman, J. (1991). Distributed representations, simple recurrent networks and grammatical structure. *Machine Learning*, 7(2-3), 195-225.

Fabry, R. (2015). Enriching the notion of enculturation: cognitive integration, predictive processing, and the case of reading acquisition - a commentary on Richard Menary. In T. Metzinger, J. Windt (Eds.), *Open MIND*: 25(c). Frankfurt am Main: The MIND Group. https://doi.org/10.15502/9783958571143.

Fabry, R. (2018). Betwixt and Between: the enculturated predictive processing approach to cognition. *Synthese*, *195*(6), 2483-2518.

Facchin, M. (2021a). Predictive processing and anti-representationalism, *Synthese*, *199*(3-4), 11609-11604.

Facchin, M. (2021b). Are generative models structural representations?. *Minds and Machines*, *31*, 277-303.

Facchin, M. (2021c). Structural representations do not meet the job description challenge, *Synthese*, *199*(3-4), 5489-5508

Facchin, M. (*submitted*). Troubles with mathematical contents. Unpublished manuscript submitted to *Erkenntnis*. Preprint at http://philsci-archive.pitt.edu/19808/

Falandays, J. B., Nguyen, B., & Spivey, M. J. (2021). Is prediction nothing more than multi scale pattern completion of the future? *Brain Research*, 147578.

Feldman, G. A. (2009). New insights in action-perception coupling. *Experimental Brain Research*, *194*(1), 39-58.

Feldman, J. (2016). What are the "true" statistics of the environment?. *Cognitive Science*, 41(7), 1871-1903.

Feldman, H. & Frsiton, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, 4: 215.

FitzGerald, T. H. B., Dolan, R. J., & Friston, K. (2014). Model averaging, optimal inference and habit formation. *Frontiers in Human Neuroscience*, 8: 457.

Flanagan, J. R:, & Wing, A. M. (1997). The role of internal model in motion planning and control: evidence from grip force adjustments during movements of hand-hold objects. *Journal of Neuroscience*, 17(4), 1519-1528.

Fodor, J. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences*, *3*(1), 63-73.

Fodor, J. (1987). *Psychosemantics*. Cambridge, MA.: The MIT Press.

Fodor, J. (1990). *A Theory of Content and Other Essays*. Cambridge, MA.: The MIT Press.

Foster, D. (2019). *Generative Deep Learning*. Sebastopol, CA.: O'Reilly.

Franklin, D. W., & Wolpert, D. M. (2011). computational mechanism for sensorimotor control. *Neuron*, 72, 425-442.

Freeman, W., & Skarda, C. (1990). Representations: who needs them?. In J. McGaugh (Ed.), *Brain Organization and Memory: Cells, Systems and Circuits*, (pp. 275-380). New York: Oxford University Press.

Fresco, N., Copeland, J., & Wolf, M. (2021). The indeterminacy of computation. *Synthese*, https://doi.org/10.1007/s11229-021-03352-9.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, 16, 1325-1352.

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 Comput Biol*, 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 motor control?. *Neuron*, 72(3), 488-498.

Friston, K. (2012a). Predictive coding, precision and synchrony. *Cognitive Neuroscience*, 3(3-4), 238-239.

Friston, K. (2012b). Prediction, perception and agency. *Journal of Psychophysiology*, 83(2), 248-252.

Friston, K (2013). Active inference and free-energy. *Behavioral and Brain Sciences*, 36(3), 132-133.

Friston, K., Adams, R. A., & Montague, R. (2012). What is value? Accumulated reward or evidence?. *Frontiers in Neurorobotics*, 4:6.

Friston, K., Adams, R. A., Perrinet, L., & Breakspear, M. (2012b). Perception as hypotheses: saccades as experiments. *Frontiers in Psychology*, 3: 151.

Friston, K., Clark, A., & Thornton, C. (2012). Free energy minimization and the dark room problem. *Frontiers in Psychology*, 3:130.

Friston, K., Danizeu, J., Kilner, J., & Kiebel, S. J. (2010). Action and behavior: a free energy formulation. *Biological Cybernetics*, 102(3), 227-260.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P. & Pezzulo, G (2017b). Active Inference: a process theory. *Neural Computation*, 29(1), 1-49.

Friston, K., & Frith, C. (2015a). A duet for one. *Consciousness and Cognition*, 36, 390-405.

Friston, K., & Frith, C. (2015b). Active inference, communication and hermeneutics. *Cortex*, 68, 129-163.

Friston, K, & Kiebel, S. (2009a). Predictive coding under the free energy principle, *Philosophical Transactions of The Royal Society B: Biological Sciences*, 364(1521), 1211-1221.

Friston, K., Levin, M., Sengupta, B., & Pezzulo, G. (2015). Knowing one's place: a free-energy approach to pattern regulation. *Journal Of the Royal Society Interface*, 12:20142383.

Friston, K., Lin, M. Frith, C. D., Pezzulo G. Hobson, J. A., Ondobaka S. (2017c). Active inference, curiosity and insight. *Neural Computation*, *29*,(10), 2633-2683.

Friston, K., Mattout, J., & Kilner J. (2011). Action understanding and active inference. *Biological Cybernetics*, 104(1-2), 137-160.

Friston, K., Parr, T. (2019). Passive motion and active inference. Commentary on "*Muscleless motor synergies and actions without movements: from motor neuroscience to cognitive robotics*" by Vishwanthan Mohan, Ajaz Bhat and Pietro Morasso. *Physics of Life Review*, *30*, 112-155.

Friston, K., Parr T., & de Vries B. (2017). The graphical brain: belief propagation and active inference. *Network Neuroscience*, 1(4), 381-414.

Friston, K. Schiner, T., FitzGerald, T., Galea, G. M., Adams, R., Brown, H., Dolan R. J., Moran, R., Stephan, E. K., Bestmann, S. (2012a). Dopamine, affordance and active inference. *PLoS Comput Biol*, *8*(1): e1002327.

Gallistell, C. R. (1980). *The Organization of Action: a New Synthesis.* Hillsdale, N. J.: Lawrence Erbaum Associates.

Garson, J. (2012). Function, selection and construction in the brain. *Synthese*, *189*(3), 451-481.

Garson, J. (2106). *A Critical Overview of Biological Functions*. Cam: Springer International Publishing.

Garson, J. (2017). A generalized select effect theory of functions. *Philosophy of Science*, *84*(3), 523-543.

Garzón, F., & Rodriguez, A. (2009). Where is cognitive science heading? Minds and Machines, 19(3), 301–318.

Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science*, *71*(5), 742-752.

Gładziejewski, P. (2015a). Action guidance is not enough, representations need correspondence too: a plea for a two-factor theory of representation, *New Ideas in Psychology* 40, 13-25.

Gładziejewski, P. (2015b). Explaining cognitive phenomena with internal representations: a mechanistic perspective, *Studies in Logic, Grammar and Rhetoric*, 40(1), 63-90.

Gładziejewski, P. (2016). Predictive coding and representationalism, *Synthese*, 193(2), 559-582.

Gładziejewski, P. (2017). Just how conservative is conservative predictive processing?. *Internetowy, Magazyn Filozoficzny Hybris*, 38, 98-122.

Gładziejewski, P. (2021). Un-debunking ordinary objects with the help of predictive processing. *The British Journal of Philosophy of Science*, https://doi.org/10.1086/715105.

Gładziejewski, P. (*Forthcoming*). Predictive processing, implicit and explicit. Retrieved from https://repozytorium.umk.pl/handle/item/6569, last accessed 28/08/2021

Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: causally relevant and distinct from detectors. *Biology and Philosophy*, 32(3), 337-355.

Goodfellow, I., Bengio, J., & Courville A. (2016). *Deep Learning*, Vol. I&II. Cambridge, MA.: The MIT Press.

Godfrey-Smith, P. (1989). Misinformation. *Canadian Journal of Philosophy*, 19(4), 533-550

Godfrey-Smith, P. (2009). Mental representations, naturalism and teleosemantics. In Macdonald, G., Papineau, D. (Eds.). *Teleosemantics*, (pp. 42-68), New York, Oxford University Press.

Goodman N. (1969). *The Languages of Art*, London: Oxford University Press.

Goschke, T., & Koppelberg, D. (1991). The concept of representation and the representation of concepts in connectionist systems. In W. Ramsey, S. Stich, D. Rumelhart (Eds.), *Philosophy and Connectionist Theory*, (pp. 129-162). New York: Routledge.

Grush R. (1997). The architecture of representation, Philosophical Psychology, 10(1), 5-23.

Grush, R. (2003). In defense of some Cartesian assumptions concerning the brain and its operations. *Biology and Philosophy*, *18*(1), 53-93.

Grush, R. (2004). The emulation theory of representation: motor control, imagery and perception. *Behavioral and Brain Sciences*, 27(3), 377-396.

Grush, R. (2008). Representation reconsidered by William M. Ramsey. Notre Dame Philosophical Reviews. http://ndpr.nd.edu/news/23327-representation-reconsidered/

Grush, R., & Mandik, P. (2002). Representational Parts. *Phenomenology and the Cognitive Sciences*, 1(3), 389-394.

Ha, D., Schmidhuber, J. (2018a). Recurrent world models facilitate policy evolution. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31* (pp. 2451-2463), Curran Associates.

Ha, D., Schmidhuber, J. (2018b). World models. *Preprint*. ArXiv:18.0310122.

Haruno, M., Wolpert, D. M., Kawato, M. (2001). MOSAIC model for sensorimotor learning and control. *Neural Computation*, 13(10), 2201-2220.

Haruno, M., Wolpert, D. M., Kawato, M. (2003). Hierarchical MOSAIC for motor generation. In T. Ono, G. Matsumoto, R. R. Llinas, A. Bethoz, R. Norgren, H. Nishijo, R. Tamura (Eds.), *Excepta Medica International Congress System* (Vol. 1250), (pp. 575-590). Amsterdam: Elsevier.

Harvey, I, Husbands, P., & Cliff, D. (1994). Seeing the light: artificial evolution, real vision, in D. Cliff, P. Husbands, J. A. Meyer & S. W. Winson (eds.), *From Animals to Animats 3* (pp. 392-401). Cambridge, MA.: The MIT press.

Harvey, I., Husbands, P., Cliff, D., Thompson, A, & Jakobi, N. (1997). Evolutionary robotics: the Sussex approach, *Robotics and Autonomous Systems*, *20*(2-4) 205-224.

Haugeland, J. (1989). *Artificial Intelligence*, Cambridge, MA.: The MIT Press.

Haugeland, J. (1991). Representational genera. In W. Ramsey, S. P. Stich, D. E. Rumelhart (Eds.). *Philosophy and Connectionist Theory*, New York: Routledge.

Haybron, D. M. (2000). The causal and explanatory role of information stored in connectionist networks. *Minds and Machines*. 10(3), 361-380.

Haykin, S. (2009). *Neural Networks and Learning Machines* (3rd edition). Upper Saddle River, N.J.: Pearson Education.

Hemion, N. J. (2016). Discovering latent states for model learning: applying sensorimotor contingencies theory and predictive processing to model context. *arXiv*: 1608.00359v1.

Hinton, G. E. (2005). What kind of graphical model is the brain?. *International Joint Conference on Artificial Intelligence.* Vol 5 (pp. 1765-1775), retrived at:

http://www.cs.toronto.edu/~hinton/absps/ijcai05.pdf. Last accessed 5/07/2020.

Hinton, G. E. (2007a). To recognize shapes, first learn to generate images. *Progress in Brain Research*, 165, 535-547.

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

Hinton, G. E. (2014). Where do features come from?. *Cognitive Science,* 38(6), 1078-1101.

Hinton, G. E., Dayan, P., Frey, B. J., & Neal R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science* 268(5214), 1158-1161.

Hinton, G. E., & Sejnowski, T. E. (1983). Optimal perceptual inference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 448.

Hobson, J. A., & Friston, K. (2012). Waking and dreaming consciousness: neurobiological and functional considerations. *Process in Neurobiology*, 98, 82-98.

Hoffman, M., & Pfeiffer, R. (2018). Robots as powerful allies for the study of embodied cognition from the bottom up. In A. Newen, L. De Bruin, S. Ghallagher (Eds.), *The Oxford Handbook of 4E Cognition* (pp. 841-862). New York: Oxford University Press.

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

Hohwy, J. (2015). The neural organ explains the mind. In T. Metzinger, J. M. Windt (Eds.), *Open MIND*: 19. 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. (2019). Quick'n'Lean or slow and rich? Andy Clark on predictive processing and embodied cognition. In M. Colombo, E. Irvine, M. Stapleton (Eds.). *Andy Clark and His Critics*. (pp. 191-205). New York: Oxford University Press.

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

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

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: an epistemological review. *Cognition*, 108(3), 687-701.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251-257.

Hua, T. Kunda, M. (2020). Modeling Gestalt visual reasoning on Raven's progressive matrices using generative image inpainting techniques. *Annual Conference on Advances in Cognitive Systems* (Palo Alto Research Centre, 2020).

Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction, and the functional architecture of the cat's visual cortex. *The Journal of Physiology*, 160(1), 106-154.

Hubel, D., & Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1), 215-243.

Hurley, S. (1998). *Consciousness in Action*. Cambridge, MA.: Harvard University Press.

Husbands, P., Harvey, I., Cliff, D. (1995). Circle in the round: state space attractors for evolved sighted robots, *Journal of Robotics and Autonomous Systems*, *15*, 83-106.

Hutto, D. & Myin, E. (2013). *Radicalizing Enactivism*. Cambridge, MA: The MIT Press.

Hutto, D:, & Myin, E. (2017). *Evolving Enactivism*. Cambridge, MA.: The MIT Press.

Iida, F. & Bovet, S. (2009). Learning legged locomotion. In A. Adamatzky, M. Komosinski (Eds.), *Artificial Life Models in Hardware*, (pp. 21-33).Verlag-London: Springer.

Jacobson, A. J. (2003). Mental representations: what philosophy leaves out and neuroscience puts in. *Philosophical Psychology*, *16*(2), 189-203.

Jacobson, A. J. (2013). *Keeping the World in Mind. Mental representations and the sciences of the mind*. Basingstoke: McMillan.

Kandel, E. R., Schwartz, J. H., Jessel, T. M., Siegelbaum, S. A., and Hudspeth A. J. (Eds.) (2012). *Principles of Neural Science* (5th edition). London: The MacGraw-Hill Companies.

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: a canonical neural computation. *Neuron*, 100(2), 424-435.

Kelso, J. A. S (1995). *Dynamic Patterns: The Self-organization of Brain and Behavior. Cambridge*, MA.: The MIT Press.

Kersten, D., Mamassian, P., Yuille A. (2004). Object perception as Bayesian inference. *Annu Rev. Psychol.*, 55, 271-304.

Kiebel, S., Daunizeau, J, & Friston, K. (2008). A hierarchy of time-scales in the brain. *PloS Comput Biol*, 4(11), e100209.

Kiebel, S., von Kriegstein, K., Daunizeau, J. & Friston, K. (2009). Recognizing sequences of sequences. *PLoS Comput Biol,* 5(8): e1000464.

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. (2020). Psychophysical identity and free energy. *Journal of the Royal Society Interface*. https://doi.org10.1098/rsif.2020.0370

Kiefer, A., & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese*, 195(6), 2387-2415.

Kiefer, A., & Hohwy, J. (2019). Representation in the prediction error minimization framework. In S. Robins, J. Symons, P. Calvo (Eds.), *The Routledge Companion to Philosophy of Psychology* (2nd Ed.) (pp. 384-410). New York: Routledge.

Kilner, J., Friston, K., & Frith, C. (2007). Predictive coding: an account of the mirror neuron system. *Cognitive Processing*, *8*(3), 159-166.

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

Kirchhoff, M. D., & Kiverstein, J. (2019). *Extended Consciousness and Predictive Processing: a third wave view*. New York: Routledge.

Kelin, C. (2018). What do predictive coders want?. *Synthese*, *195*, 2541-2557.

Knill, D., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Cognitive Science*, 27(12), 712-719.

Koski, T, & Noble, J. (2009). *Bayesian Networks: An Introduction*. Chichester: Wiley & Sons.

Kruse, R, Borgelt, C., Braune, C., Mostaghim, S., & Steinbrecher, M. (2016). *Computational Intelligence. A Methodological Introduction* (2nd edition). Verlag and London: Springer.

Kunda, M. (2021). AI, visual imagery, and a case study on the challenges posed by human intelligence tests. *Proceedings of the National Academy of Sciences*, *117*(47), 29390-29397.

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. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* https://doi.org/10.1109/IROS.2018.8593684

Lashley, K. S. (1929). *Brain Mechanisms and Intelligence*. Chicago, Ill.: Chicago University Press.

Lee, J. (2018). Structural representation and the two problems of content. *Mind & Language*, 34(5), 606-626.

Lee , T. S., & Mumford D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America, A*, 20(7): 1434-1448.

Le Hir, N., Sugaud, O., & Laflaquière A. (2018). Identification of invariant sensorimotor structures as a prerequisite for the discovery of objects. *Frontiers in Robotics and AI*, *5*:70.

Leinbewer, M., Ward, D. R., Sobczak, J. M., Attinger, A., & Keller, G. B. (2017). A sensorimotor circuit in the mouse cortex for visual flow predictions. *Neuron*, 95(6), 1420-1432.

Leitgeb, H. (2020). On non-eliminative structuralism: unlabeled graphs as a case study, part A. *Philosophia Mathematica*. https://doi.org/10.1093/philmat/nkaa001.

Levittin, J. Y., Maturana, H. R., McCulloch, W. S., & Pitts, W. H. (1959). What the frog's eye tells the frog's brain. *Proceedings of the IRE*, *47*(11), 1940-1951.

Lin, H. W., Tegmark, M., & Rolnick D. (2017). Why does deep and cheap learning work so well?. *Journal of Statistical Physics*, 168(6), 1223-1247.

Linson, A., Clark, A., Ramamoorthy, S., & Friston, K. (2018). The active inference approach to ecological perception: general information dynamics for natural and artificial embodied cognition. *Frontiers in AI and Robotics*, https://doi.org/10.3389/frobt.2018.00021

Lyre, H. (2016). Active content externalism. *Review of Philosophy and Psychology*, *7*(1), 17-33.

Maris, M., & Schaad, R. (1995). The didactic robots, *Techreport No. IIF-AI-95.09, AI Lab, Department of Computer Science, University of Zurich.*

Maris, M., & te Boekhorst, R. (1996). Exploiting physical constraints: heap formation through behavioral error in a group of robots, in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 1655-1660) Piscataway, NJ: IEEE Press.

Marr, D. (1982). *Vision. A Computational Approach*. San Francisco, CA: Freeman & Co.

Mataric, M. (1991). Navigating with a rat's brain: a neurobiologically inspired model for robot spatial representation. In J. A. Meyer and S. Wilson (Eds.), *From Animals to Animats 1* (pp. 169-75). Cambridge, MA.: The MIT Press.

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.

McClelland, J. L. (1998). Connectionist models and Bayesian inference. In M. Oaksford, N. Charter (Eds.), *Rational Models of Cognition*, (pp. 21-53). New York: Oxford University Press.

McClelland, J. L. (2013). Integrating probabilistic models of perception and interactive neural networks: a historical and tutorial overview. *Frontiers in Psychology*, 4:503.

McClelland, J. L., Rumelhart, D. E., and the PDP research group (Eds.). (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Voll. I & II. Cambridge, MA.: The MIT Press.

McCloskey, M. (1991). Networks and theories: the place of connectionism in cognitive science. *Psychological*

Science, *2*(6), 387-395.

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

McNamee, D., & Wolpert, D. M. (2019). Internal models in biological control. *Annual Review of Control, Robotics, and Autonomous Systems* 2, 339-364.

Menary, R. (2007). *Cognitive Integration: Mind and Cognition Unbound*. Basingstroke, Palgrave Macmillan.

Merckelbach, H., & van der Ven, V. (2001). Another white christmas: fantasy proneness and reports of hallucinatory experiences in undergraduate students. *Journal of Behavioral Therapy and Experimental Psychiatry*, 32, 137-144.

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

Miłkowski, M. (2017). Szaleństwo, a nie metoda. Uwagi o książce Pawła Gładziejewskiego "Wyjaśnianie za pomocą reprezentacji mentalnych". *Filozofia Nauki*, *25*(3(99)), 57-67.

Miller, K. J., Schalk, G., Fetz, E. E., den Nijs, M., Ojerman, J. G., & Rao, R. (2010). Cortical activity during motor execution, motor imagery, and imagery-based online feedback. *Proceedings of the National Academy of Sciences*, 107(9), 4430-4435.

Millidge, B., Seth, A. K., & Buckley C. (2021). Predictive coding: a theoretical and experimental review. *arXiv preprint arXiv:2107.12979*

Millidge, B., Tschantz, A., Seth, A. K., Buckley, C. L. (2020). Relaxing the constraints on predictive coding models. *arXiv preprint arXiv:2010.01047*

Millikan, R. G. (1984). *Language, Thought, and other Biological Categories.* Cambridge, MA.: The MIT Press.

Millikan, R. G. (2017). *Beyond Concepts*. New York: Oxford University Press.

Millikan, R. G. (2020). Neuroscience and teleosemantics. *Synthese*, https://doi.org/10.1007/s11229-020-02893-9.

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

Mollo, D. C. (2021). Why go for a computation-based approach to cognitive representations. *Synthese*, https://doi.org/10.1007/s11229-021-03097-5.

Mollo, D. C. (*forthcoming*). Deflationary realism: representation and idealization in cognitive science. *Mind and Language*, preprint (accepted version) at: http://philsci-archive.pitt.edu/17591/

Morgan, A. (2014). Representations gone mental. *Synthese*, 191(2), 213-244.

Morgan, A. (2019). Against neuroclassicism: on the perils of armchair neuroscience. *Mind and Language*, https://doi.org/10.1111/mila.12304.

Moser, E. I., Kropff, E., & Moser, M. B. (2008). Place cells, grid cells, and the brain spatiotemporal representations system. *Annu. Rev. Neuroscience*, 31, 69-89.

Murphy, K. P. (2012). *Machine Learning: a Probabilistic Approach*. Cambridge, MA.: The MIT Press.

Nair, V., & Hinton G. E. (2006). Inferring motor programs from images of handwritten digits. *Advances in Neural Information Processing Systems*, (pp. 515-522).

Namikawa, J., Ryunosuke, N., & Tani, J. (2011). A neurodynamical account of spontaneous behavior. *PLoS Comput Biol*, 7(10), e1002221.

Neander, K. (2017). *A Mark of The Mental*. Cambridge, MA.: The MIT Press.

Nirshberg, G., & Shapiro, L. (2020). Structural and Indicator representations: a difference in degree, not in kind. *Synthese*, https://doi.org/10.1007/s11229-020-02537-y.

Nolfi, S. (2002). Power and limits of reactive agents. *Neurocomputing*, *42*(1-4), 119-145.

Nolfi, S., & Parisi, D. (1993). Auto teaching: networks that develop their own teaching input. *Technical report PCIA91-03, CNR Rome*.

O'Brien, G. (2015). How does the mind matter? Solving the content causation problem. In T. Metzinger, J. M. Windt (Eds.). *Open MIND*: 28(T). Frankfurt am Main, The MIND Group. https://doi.org/10.15502/9783958570146.

O'Brien, G., & Opie, J. (2001). Connectionsit vehicles, structural resemblance, and the phenomenal mind. *Communication and Cognition*, 34(1/2), 13-38.

O'Brein, G., & Opie, J. (2004). Notes towards a structuralist theory of mental representations, in H. Clapin; P. Staines & P. Slezak (eds.), *Representation in Mind: New Approaches to Mental Representaion* (pp. 1-20). Oxford: Elsevier.

O'Brien, G., & Opie, J. (2006). How do connectionist networks compute?, *Cognitive Processing*, 7(1), 30-41.

O'Brien, G., & Opie, J. (2010). Representation in analog computation. In A. Newen, A. Bartles, E. de Jung (Eds.), *Knowledge and Representation*. Stanford, CA.: CSLI

O'Callaghan, C., Kveraga, K., Shine, J. M., Abrams, R. B., & Bar, M. (2017). Predictions penetrate perception: converging insights from brain, behavior and disorder. *Consciousness and Cognition*, 47, 63-74.

O'Reilly, R. C., Wyatte, D., & Rohrlich, J. (2014). Learning through time in the talamocortical loops. Retrieved from https://arxiv.org/pdf/1407.3432.pdf. Last accessed 18/07/2020.

O'Regan, J. K. (2011). *Why Red doesn't Sound Like a Bell: Understanding the Feeling of Consciousness*. New York: 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.

Orlandi, N. (2014). *The Innocent Eye: Why Vision is not a Cognitive Process*. New York: Oxford University Press.

Orlandi, N. (2016). Bayesian perception is ecological perception. *Philosophical Topics*, 44(2), 327-352.

Orlandi, N. (2020). Representing as coordinating with absence. In J. Smortchkova, K. Dołega, T. Schlicht (Eds.), *What Are Mental Representations?*, (pp. 101-134). New York: Oxford University Press.

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). New York: Oxford University Press.

Ororbia, A., Kelly, M. A. (2021). Towards a predictive processing implementation of a common model of cognition. *ArXiv Preprint*: 2105.07308. Last accessed 30/05/2020.

Palacios, E. R., Isomura, T., Parr, T., & Friston, K. (2019). The emergence of synchrony in networks of mutually inferring neurons. *Scientific Reports*, 9(1), 1-14.

Palmer, S. (1978). Fundamental aspects of cognitive representation. In E. Rosch, E. Lloyd (Eds.), *Cognition and Categorization*, (pp. 259-303), Hillsdale, N.J.: Erbaum.

Parr, T., & Friston, K. (2017). Working memory, attention and salience in active inference. *Scientific Reports*, *7*(1), 1-21.

Parr, T., & Friston, K. (2019). Attention or salience?, *Current Opinion in Psychology*, 29, 1-5.

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*, (167-177). Berlin and Heidelberg: Springer.

Peirce, C. S. S. (1938-51). *The Collected Papers of Charles Sanders Peirce* (vol. 1-8). A Burks, C. Hartstone, P. Weiss (Eds.). Cambridge, MA.: Cambridge University Press.

Penny, W. (2012a). Bayesian models of brain and behaviour. ISRN Biomathematics, 2012, 1–19. https://doi.org/10.5402/2012/785791.

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

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/10.15502/9783958573215.

Pezzulo, G., Donnarumma, F., Iodice, P., Maisto, D., & Stoianov I. (2017). Model-based approaches to active perception and control. *Entropy*, 19(6): 266.

Pezzulo, G., & Sims, M. (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

Pfeiffer, R., & Bongard, J. (2007). *How the Body Shapes the Way We Think. A New View of Intelligence*. Cambridge, MA.: The MIT Press.

Piantadosi, S. T. (2021). The computational origin of representation. *Minds and Machines*, *31*(1), 1-58.

Piccinini, G. (2006). Computation without representation. *Philosophical Studies*, 137(2), 205-241.

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

Piccinini, G. (2020). Nonnatural mental representations. In J. Smortchkova, K. Dołega and T. Schlicht (Eds.). *What Are Mental Representations?* (pp.254-286). New York: Oxford University Press.

Piccinini, G., & Maley, C. (2021). Computation in Physical Systems. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition). https://plato.stanford.edu/archives/sum2021/entries/computation-physicalsystems/ Last accessed 1/09/2021

Piccinini, G., & Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of Biological Physics*, 37(1), 1-38.

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., Nizard, A., Friston, K., Pezzulo, G. (2016). Active inference and robotic control: a case study. *Journal of the Royal Society Interface*, 13(122), 20160616.

Poldrack, R. (2020). The physics of representation. *Synthese*, https://doi.org/10.1007/s11229-020-02793-y

Raja, V., Valluri, D., Baggs, E., Chemero, A., & Anderson, M. L. (Forthcoming). The Markov Blanket trick: on the scope of the Free-energy principle and active inference. *Preprint*. Retrived at: http://philsci-archive.pitt.edu/18843/, last accessed 08/04/2021

Ramsey, W. (1997). Do connectionist representations earn their explanatory keep?, *Mind & Language*, *12*(1), 34-66.

Ramsey, W. (2003). Are receptors representations?, *Journal of Experimental & Theoretical Artificial Intelligence*, 15(2).

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

Ramsey, W. (2016). Untangling two questions about mental representation. *New Ideas in Psychology*, 40, 3-12.

Ramsey, W. (2017). Must cognition be representational?. *Synthese*, *194*(11), 4197-4214.

Ramsey, W. (2020). Defending representational realism. In J. Smortchkova, K. Dołega, T. Schlich (Eds.), *What are Mental Representations?* (pp. 54-78). New York: Oxford University Press.

Ramsey, W., Stich, S. P., & Garon, J. (1991). Connectionism, eliminativism and the future of folk psychology. In W. Ramsey, S. P. Stich, D. E. Rumelhart (Eds.), *Philosophy and Connectionist Theory* (pp. 199 - 228). New York: Routledge.

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

Rao, R., & Ballard, D. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation*, 9(4), 721-763.

Rao, R, & Ballard D. (1999): Predictive coding in the visual cortex:a functional interpretation of some extra-classical receptive fields. *Nature Neuroscience*, 2(1), 79-87.

Rao, R., & Senjowski, T. J. (2002). Predictive coding, cortical feedback, and spike-timing dependent plasticity. In R. Rao, B. A. Olshausen, M. S. Lewicki (eds.), *Probabilistic Models of the Brain: Perception and Neural Function,* (pp. 297-315). Cambridge, MA.: The MIT Press.

Rietveld, E., Denys, D., & van Westen, M. (2018). Ecological-enactive cognition as engaging with a field of relevant affordances: the Skilled Intentionality Framework (SIF). In A. Newen, L. De Bruin, S. Gallagher (Eds.), *The Oxford Handbook of 4E Cognition* (pp. 41-70). New York: Oxford University Press.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge, U. K.: Cambridge University Press.

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: a Parallel Distributed Processing approach*. Cambridge, MA.: The MIT Press.

Rogers, T. T., & McClelland, J. L. (2014). Parallel Distributed Processing at 25: further explorations in the microstructure of cognition. *Cognitive Science*, 38(6), 1024-1077.

Rowlands, M. (2006). *Body Language*. Cambridge, MA.: The MIT Press.

Rupert, R. (2018). Representation and mental representations. *Philosophical Explorations*, 21(2), 204-225.

Ryder, D. (2009a). Problems of representation I: nature and role. In S. Robins, J. SImons, P. Calvo (Eds.), *The Routledge Companion to Philosophy and Psychology* (2nd ed.) (pp. 233-250). New York: Routledge.

Ryder, D. (2009b). Problems of representation II: naturalizing content. In S. Robins, J. SImons, P. Calvo (Eds.), *The Routledge Companion to Philosophy and Psychology* (2nd ed.) (pp. 251-280). New York: Routledge.

Sachs, C. B. (2018). In defense of picturing: Sellars's philosophy of mind and cognitive neuroscience. *Phenomenology and the Cognitive Sciences*, 18(4), 669-689.

Sandborn, A. N, & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, 20(12), 883-893.

Schlicht, T., & Starzak, T. (2021). Prospects to enactivists approaches to intentionality and cognition. *Synthese*, *198*(1), 89-113,

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, J. Windt (eds.). *Open MIND*, 35. Frankfurt am Main, The MIND Group. https://doi.org/10.15502/9783958570108.

Seth, A.K. (2021). *Being You*. London: Faber&Faber.

Seth, A. K., & Critchley H. D. (2013). Extending predictive processing to the body: emotion as interoceptive inference. *Behavioral and Brain Sciences*, 36(3), 227-228.

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.

Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication*, Urbana, IL.: University of Illinois Press.

Sharot, T. (2011). The optimism bias. *Current Biology*, *21*(23), R491-R945.

Shea, N. (2007). Content and its vehicles in connectionist systems, *Mind and Language*, *22*(3), 246-269.

Shea, N. (2014). IV – Exploitable isomorphism and structural representation. *Proceedings of the Aristotelian Society*. Vol. 114, 2,2, (pp. 123-144). New York: Oxford University Press.

Shea, N. (2018). *Representation in Cognitive Science*. New York: Oxford University Press.

Shi, Y. Y., & Sun H. (2008). *Image and Video Compression for Multimedia Engineering. Fundamentals, Algorithms and Standards* (2nd ed.). New York: CRC Press.

Shipp, S. (2016). Neural elements for predictive coding. *Frontiers in Psychology*, 7: 1792.

Shipp, S., Adams, R. A., & Friston, K. (2013). Reflections on agranular architecture: predictive coding in the motor cortex. *Trends in Neurosciences*, 36(12), 706-716.

Simione, L., & Nolfi, S. (2015). Selection-for-action emerges in neural networks trained to learn spatial associations between stimuli and actions. *Cognitive Processing*, *16*(1), 393-397.

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.

Skansi, S. (2018). *Introduction to Deep Learning: from Logical Calculus to Artificial Intelligence*. Springer.

Smith, R., Ramstead, M. D., & Kiefer, A. (2021). Active inferences models do not contradict folk psychology. *Psyaxiv Preprint*, https://psyarxiv.com/kr5xf. Last accessed 06/06/2021

Smolensky, P. (1990). Tensor product variable binding and the representations of symbolic structures in connectionsit systems. *Artificial Intelligence*, 46(1-2), 159-216.

Sperry, R. W. (1950). Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of Comparative and Physiological Psychology*, 43(6), 482-489.

Sporns, O. (2010). *Networks in the Brain*. Cambridge, MA.: The MIT Press.

Spratling, M. W. (2015). Predictive coding. In D. Jaeger, R. Jung (Eds.), *Encyclopedia of Computational Neuroscience*, (pp. 2491-2494), New York: Springer.

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.

Spratling, M. W. (2019). Fitting predictive coding to neurophysiological data. *Brain Research*, *1720*, 146313.

Sprevak, M. (2011). Review of Representation reconsidered. *The British Journal of Philosophy of Science*, *62*, 669-675.

Sprevak, M. (2013). Fictionalism about neural representations. *The Monist*, 96(4), 539-560.

Sullivan, J., A. (2010). A role for representations in cognitive neurobiology. *Philosophy of Science*, *77*(5)

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese*. 87(3), 449-508.

Tani, J. (2003). Learning to generate articulated behavior through the bottom up and the top down interaction processes. *Neural Networks*, *16*(1), 11-23.

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: Actions, Symbols, and Consciousness as Self-organizing Dynamical Phenomena*. New York: Oxford University Press.

Tani, J., & Nolfi, S. (1999). Learning to perceive the world as articulated: an approach for hierarchical learning in sensory-motor systems. *Neural Networks*, *12*(7-8), 1131-1141.

Taylor, S. (1989). *Positive Illusions. Creative Self-Deception and the Healthy Mind*. New York: Basic Books.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure and abstraction. *Science*, 331(6602), 1279-1285.

Thornton, C. (2017). Predictive processing simplified: the infotropic machine. *Brain and Cognition*, *112*, 13-24.

Thornton, C. (2020). Predictive processing: does it compute?. In D. Mendoça, M. Curado, S. Gouveia (Eds.). *The Philosophy and Science of Predictive Processing* (pp.141-156). London: Bloomsbury Academic.

Todorov, E. (2009a). Efficient computation of optimal actions. *Proc. Natl. Acad. Sci. USA*, 106, 11478-11483.

Todorov, E. (2009b). Parallels between sensory and motor information processing. In M. Gazzaniga (Ed.), *The Cognitive Neurosciences* (*4th edition*) (pp. 613-624). Cambridge, MA.: The MIT Press.

Todorov, E., & Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, 5(11), 1226-1235.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189-208.

Tschantz, A., Seth, A. K., & Buckley, C. L. (2020). Learning action-oriented models through active inference. *PLoS Computational Biology*, *16*(4): e1007805.

Usher, M. (2001). A statistical referential theory of content: using information theory to account for misrepresentation. *Mind and Language*, 16(3), 311-334.

Van de Cruys, S., Friston, K., & Clark, A. (2020). Controlled optimism: reply to Sun and Firestone on the Dark Room Problem. *Trends in Cognitive Sciences.* https://doi.org/10.1016/j.tics.2020.05.012

Van Gelder, T. (1991). What is the "D" in "PDP"? A survey of the concept of distribution. In W. Ramsey, S. P. Stich, D. E. Rumelhart (Eds.). *Philosophy and Connectionist Theory*, New York: Routledge.

Van Gelder, T. (1992). Defining distributed representations. *Connection Science*, 4(3-4), 175-191.

Vásquez, M. J. C. (2019). A match made in haven: predictive approaches to (an unorthodox) sensorimotor enactivism. *Phenomenology and the Cognitive Sciences*. https://doi.org/10.1007/s11097-019-09647-0.

Vecchi, T., Gatti, D. (2020). *Memory as Prediction*. Cambridge, MA.: The MIT Press.

Von Eckardt, B. (1996). *What is Cognitive Science?*. Cambridge, MA.: The MIT Press.

Vold, K.; & Schlimm, D. (2020). Extended mathematical cognition: external representations with non-derived content. *Synthese*, *197*, 3757-3777.

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences*, 1464(1), 242-268.

Webb, B. (1994). Robotic experiments in cricket phonotaxis. In D. Cliff, P. Husbands, J.A. Meyer, S. WIlson (Eds.),

From Animals to Animats 3, (pp. 45-54). Cambridge, MA.: The MIT Press.

Webb, B. (2004). Neural mechanisms for predictions: do insects have forward models?. *Trends in Neurosciences*, *27*(5), 278-282.

Webb, B. (2006). Transformation, encoding and representation. *Current Biology*, 16(6), R184-R185.

Webb, B. (2019). The minds of insects. In M. Colombo, E. Irvine, M. Stapleton (Eds.). *Andy Clark and His Critics*, (pp. 254-265.). New York: Oxford University Press.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, 5(6), 598-604.

Wiese, W. (2017). What are the contents of the representations in predictive processing?. *Phenomenology and the Cognitive Sciences*, 16(4), 715-736.

Wiese, W. (2018). *Experienced Wholeness*. Cambridge, MA.: The MIT Press.

Wiese, W., & Friston, K. (2021). Examining the continuity between life and mind: is there a continuity between autopoietic intentionality and representationality?. *Philosophies*, *6*(1): 18.

Williams, D. (2017). Predictive Processing and the Representation Wars. *Minds And Machines*, 28(1), 141-172.

Williams, D. (2018a). *The Mind as a Predictive Modeling Engine: Generative Models, Structural Similarity, and Mental Representation*. Ph.D. Disseration, University of Cambridge, UK. Accessed at https://www.repository.cam.ac.uk/bitstream/handle/1810/286067/Daniel%20Williams%20PhD%20Thesis.pdf?sequence=1. Last accessed 17/09/2020.

Williams, D. (2018b). Predictive minds and small scale models: Kennet Craick's contribution to cognitive science. *Philosophical Explorations*, 21(2), 245-263.

Williams, D. (2018c). Pragmatism and the predictive mind. *Phenomenology and the Cognitive Sciences*, *17*(5), 835-859.

Williams, D. (2020). Predictive coding and thought. *Synthese*, *197*(4), 1749-1775.

Williams D., & Colling L. (2017). From symbols to icons: the return of resemblance in the cognitive science revolution, *Synthese*, 195(5), 1941-1967.

Wittgenstein, L. (1921/2013). *Tractatus Logico-Philosophicus*. New York: Routledge.

Wolpert, D. M:, Doya, I., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 358(1431), 593-601.

Wolpert, D. M., & Flanagan J. R. (2001). Motor prediction, *Current Biology*, 11(18), R729-732.

Zahnoun, F. (2019). On representation hungry cognition (and why we should stop feeding it). *Synthese*, *198,* 267-284.