Tools for Thought

Faculty of Arts

Department of Philosophy

Tools for Thought

Free Energy Instrumentalism in an Enactive Framework

Thesis for the degree of Doctor in Philosophy at the University of Antwerp by

Thomas van Es

Supervisor:

Prof. Dr. Erik Myin

Antwerp, 2021

Faculteit Letteren en Wijsbegeerte

Departement Wijsbegeerte

Denkgereedschap

Vrije energie instrumentalisme in een enactivistisch kader

Proefschrift voorgelegd voor het behalen van de graad van Doctor in de wijsbegeerte aan de Universiteit Antwerpen te verdedigen door

Thomas van Es

Promotor:

Prof. Dr. Erik Myin

Antwerpen, 2021

Voor mijn moeder.

Get real, they tell me. If only they knew,
how real this life really gets.
— Sean ‘Slug’ Daley

Table of Contents

Acknowledgments

This thesis has been nearly four years in the making, and I couldn't have done it without the help of all the amazing people around me. First and foremost I want to express my gratitude to my wife Alessandra, whose support and love has pulled me through many tough times, and has for multiple times been a sounding board for many of my ideas; thank you so much. Many thanks to my supervisor Erik Myin who has been an excellent guide through this first stage in academia (hopefully of more to come), and whose strict but fair criticism continues to push me forward. I also want to thank the many colleagues in Antwerp for reading and commenting on my work, as well as the helpful discussions during lunchtime when that could still happen, such as Karim Zahidi, Zuzanna Rucińska, Ludger van Dijk, and especially my pre-COVID office mate Luca Roccioletti for the many tangentially related discussions that took up far more time than they should've —it's just difficult to let it be when you disagree — I suppose that's a blessing and a curse for a philosopher. Thanks to Jo Bervoets for working through some thoughts on autism with me in our co-written paper, it's been equal parts theoretical and personal exploration, and I hope we can continue working together.

I also want to thank the team in Wollongong for having been such excellent hosts in what was just a terrible time to have some good ol' chats on philosophy: after the raging fires, amidst a pandemic. Many thanks to Michael Kirchhoff who magically found the time to supervise me in all the chaos. Thanks to Inês Hipólito who helped my wife and I find our way and became a collaborator and good friend. I'm pretty proud of our work together, and I have the feeling there's plenty left for us to say. Thanks also to Mads Dengsø, I hadn't thought to find a fellow appreciator of the fine delicacy of salmiak and salty licorice in Australia. Also many thanks to the readers of earlier drafts of the work presented here that haven't been mentioned yet such as Mel Andrews, Manuel Baltieri, Kristien Hens and Victor Loughlin, and Sander Van de Cruys.

Finally, I want to thank the mother who raised me to be the person I am today, who I know was always proud of everything I did, but was not able to see me finishing my doctorate. Bedankt mam, voor alles. Woorden schieten tekort om je liefde en zorg te beschrijven, en eigenlijk is een doctoraatsthesis daar ook niet de juiste plek voor. Ik mis je en denk elke dag aan je. Ik draag dit werk op aan jou, zoals eigenlijk elke prestatie van mij impliciet weer een ode is aan je opvoeding, aan je visie en aan je veerkracht. Rust zacht, mam.

Much love to everyone else involved that I have forgotten to mention,

Thomas

Abstract

This thesis explores the philosophy of cognitive science. It argues for an *instrumentalist* approach to the *free energy principle* (FEP) within an *enactivist* framework. Enactivism is an approach to cognitive science that stresses the importance of an organism's embodiment, their environmental embedding, and their interactional history. In enactivism, one explains organismic activity not with a study of neural activity, but instead with a study of the organism's biological features, behavior patterns and the relevant historical trajectory. The FEP provides a formalism with which to model organismic activity in an environment, that unearths statistical relations between an organism's and its environment's dynamics. The formalism's interpretation remains disputed.

Part I discusses the representationalist and realist interpretation of the FEP where the formalism is implemented in the brain to make predictions about the environment called predictive processing. Chapter 2, co-written with Erik Myin, argues that the representational status is unwarranted by providing a counterexample to the definition of representation used in predictive processing literature. Moreover, Chapter 3 argues that, even granting the representational status, predictive processing is internally inconsistent, inviting a contradiction where systems will be required both to represent *and not* represent aspects of their environment. The predictive processing project thus seems conceptually confused. Chapter 4 studies the embedded view of vision and argues that it provides a powerful non-representational explanation of the phenomena predictive processing targets.

Part II explores alternative interpretations to the FEP. Chapter 5 suggests that the usage of the term 'model' in the literature supports only the weaker instrumentalist interpretation, whilst the realist alternative demands additional argumentation. Chapter 6, co-written with Inês Hipólito, supplements this by providing a systematic overview of possible interpretations, and argues that only instrumentalism is tenable. An interesting corollary is that the question of representationalism for models in the FEP moves to the background.

In Part III instrumentalism is put to work. First, Chapter 7, co-written with Jo Bervoets, argues that an enactivist conceptualization of autism as primarily a sensorimotor atypicality aids in formulating an ethical appeal for inclusion. Chapter 8, co-written with Michael Kirchhoff, elaborates on the explanatory power of enactivism and attempts to incorporate its core concepts into the FEP's formalism. The results are mixed: certain features fit seamlessly in the formalism, whereas others do not. The conclusion reflects the general conclusion of the

thesis: while the FEP provides statistical detail, it is explanatorily stronger when embedded within an enactivist framework.

Samenvatting

In deze thesis verken ik de filosofie van de cognitiewetenschappen. Ik verdedig een *instrumentalistische* interpretatie van het *vrije energie principe*, binnen het theoretisch kader van het *enactivisme*. Het vrije energie principe fungeert hier als aanvullend analytisch gereedschap binnen een enactieve conceptualisering van cognitie. Het enactivisme is een stroming in (de filosofie van) de cognitiewetenschappen die zich sterk afzet tegen het doorgaanse neurocentrisme en de alomtegenwoordige computermetafoor van het brein en cognitie. In plaats van het brein, stelt het enactivisme dat het organisme-in-haar-omgeving het correcte doel van de studie van organismische activiteit is. Indien we willen verklaren waarom een bepaald organisme bepaalde zaken op een bepaalde manier doet, heeft het, volgens de enactivist, weinig zin het organisme in een breinscanner te leggen. We kunnen beter het organisme in de omgeving en de interactionele geschiedenis bestuderen om te achterhalen hoe bepaalde patronen ontstaan. In het enactivisme zijn wij *niet* ons brein, maar is het brein slechts een orgaan in het lichaam dat handelt in haar omgeving. Deze verandering van het perspectief heeft verregaande implicaties voor hoe we cognitief wetenschappelijk onderzoek doen.

Het vrije energie principe is een wiskundige beschrijving van bepaalde kenmerken eigen aan *zelforganiserende* systemen zoals organismen: systemen die in onevenwicht stationaire toestand (Engels: *non-equilibrium steady-state*) zijn met hun omgeving. Dit slaat grofweg op systemen die over een bepaalde duur zich in dezelfde staat bevinden door middel van voortdurende interactie met de omgeving. Bij benadering vallen organismen hieronder, bijvoorbeeld. Dit formalisme kan op welk systeem dan ook toegepast worden om de specifieke relaties tussen de verschillende aspecten van het bestudeerde systeem uit te lichten. In het formaliseringsproces wordt er een model opgesteld van het systeem in de vorm van een opsomming van variabelen en hun statistische relaties tot elkaar. Bij zelforganiserende systemen zal je telkens zien dat de variabelen over tijd binnen een bepaalde band blijven, wat wiskundig equivalent is aan het laag blijven van de waarde van de vrije energie. Dit betekent dat, zolang een systeem in stand blijft, de waarde van de vrije energie laag blijft.

In één bepaalde interpretatie van het vrije energie principe genaamd *predictive processing* (letterlijk: predictief verwerken) betekent dit dat het systeem zelf het model gebruikt om voorspellingen te doen om de vrije energie laag te houden en dus in stand te

blijven. Specifiek wordt er gedacht dat het brein dit model implementeert, en dat het model een beschrijving bevat van de causaal-probabilistische structuur van de wereld en de rest van het lichaam. Met dit model kan het brein de juiste voorspellingen doen om zichzelf in leven te houden. Zo wordt dus gedacht dat het brein de enige relevante actor is in een geheel organisme, dat zelf probeert uit te vinden hoe het lichaam het best bestuurd kan worden om in leven te blijven. Hierin staat het dus haaks op het enactivisme.

In *Part I* (Deel I) van de thesis analyseer ik de *predictive processing* interpretatie en beargumenteer ik dat het perspectief niet gerechtvaardigd is in haar representationalistische beginselen en bovendien contradictoir is. Ik sluit het deel af met een bespreking van een gematigd representationalistisch voorstel, wat vooruitzicht biedt voor een geheel non-representationeel perspectief. Hoofdstuk 2, geschreven met Erik Myin, betoogt dat de definitie van *representatie* die gebruikt wordt in de *predictive processing* literatuur ontoereikend is. Kort gezegd bieden we een tegenvoorbeeld, dat aan alle vier genoemde voorwaarden voldoet zonder intuïtief als representatie te gelden. In Hoofdstuk 3 graaf ik dieper in *predictive processing* en beargumenteer ik dat, zelfs als we ervan uitgaan dat de theorie valide is, ze intern contradictoir is. Specifiek beargumenteer ik dat een onderdeel van een systeem medeonderdelen wel én niet zal moeten representeren, afhankelijk van of we het systeem als geheel beschouwen, of de onderdelen als systeem op zich beschouwen —twee perspectieven die binnen *predictive processing* beide belangrijk zijn.

In Hoofdstuk 4 bekijk ik de *embedded view* (letterlijk: ingebed zicht) theorie, waarin niet het brein op zich maar juist de inbedding in de omgeving centraal staat. De theorie biedt een non-representationele verklaring voor een groot deel van onze cognitie, maar behoudt de noodzaak voor representatie voor een selecte groep activiteiten die te maken hebben met zaken die 'afwezig' zijn. In dit hoofdstuk beargumenteer ik dat de nood voor representatie misplaatst is en dat ook de activiteiten die met zogenoemde 'afwezige' zaken goed te verklaren zijn met hetzelfde mechanisme waarmee we de andere activiteiten verklaarden. Dit betekent dat, sprekend met hoofdstukken 2 en 3, representaties in *predictive processing* ongerechtvaardigd zijn en problemen opleveren en, sprekend met Hoofdstuk 4, ze ook onnodig zijn.

Na de bespreking van de *predictive processing* interpretatie van het vrije energie principe in *Part I*, verbreed ik het blikveld en worden andere perspectieven op het vrije energie principe bekeken in *Part II* (Deel II). In Hoofdstuk 5 analyseer ik het gebruik van het woord 'model' in de vrije energie literatuur en ontwaar ik een spraakverwarring. Zo wordt het enerzijds gebruikt als een model zoals wij dat opmaken van systemen die we willen bestuderen, en anderzijds als een model dat wordt gebruikt door het systeem zelf, zoals in *predictive*

processing geopperd wordt. De verwarring doet zich specifiek voor wanneer bevindingen in het wetenschappelijk gevormde model direct toegepast worden op het systeem zelf zonder aanvullende argumentatie waarom we ervan uit zouden moeten gaan dat dit model daadwerkelijk door het organisme geëxploiteerd wordt —dit is een onderliggende aanname. Dit heeft tot gevolg dat er grootse doch ongerechtvaardigde conclusies worden getrokken. Hoofdstuk 6, geschreven met Inês Hipólito, borduurt hierop voort. We maken een systematisch overzicht van de verschillende mogelijke interpretaties van het vrije energie principe over de assen van representationalisme/non-representationalisme, waarin vastgesteld wordt of in eender welke interpretatie het model gezien wordt als *representatie* van een doelobject of niet, en realisme/instrumentalisme, waarin wordt vastgesteld of in eender welke interpretatie het model als reëel onderdeel van de te beschrijven wereld ziet (in *realisme*), of slechts als wetenschappelijk *werktuig* waarmee we de wereld kunnen vatten (in *instrumentalisme*). Uit onze analyse volgt dat slechts de instrumentalistische kijk het hoofd boven water houdt. Een interessante implicatie hiervan is dat de vraag over representationalisme van minder belang wordt. Als het model slechts een wetenschappelijk werktuig is, vervormt de representationalismekwestie tot een algemeen wetenschapsfilosofisch vraagstuk over ons gebruik van modellen en of die al dan niet representationeel van aard zijn. Echter, voor cognitief wetenschappelijk onderzoek is dit niet belangrijk. Wat uitmaakt is dat het model ons helpt het doelsysteem beter te begrijpen en dit staat los van de conceptuele analyse van die modellen.

Part III (Deel III) bespreekt de compatibiliteit van het instrumentalistische perspectief op het vrije energie principe met het enactivisme. Eerst, in Hoofdstuk 7, laten co-auteur Jo Bervoets en ik de verklaringskracht van het enactivisme zien. We construeren een conceptualisering van autisme als een atypisch patroon van sensorimotorische interactie. Atypische sensorimotorische interacties leiden vanzelf in interactie met een sociale omgeving ook tot atypische sociale interactiepatronen. Omdat die patronen vormen in interactie met de omgeving, is ook de omgeving constitutief voor de gevormde interactiepatronen. Zo kunnen we zowel de homogeniteit van veel basale interactiepatronen als de heterogeniteit van sociale interactiepatronen in autisten ondervangen. Bovendien is onder deze conceptualisering autisme een gelijkaardige atypischheid als bijvoorbeeld atypisch lang zijn, of blind zijn. Vanuit deze gelijkenis formuleren we ook een ethisch appel voor verdere inclusie van autisten in onze samenleving, analoog aan hoe we ook onze samenleving inclusiever pogen te maken voor bijvoorbeeld blinden.

Na deze enactivistische uiteenzetting is het interessant om te zien tot in hoeverre het enactivistisch perspectief ondervangen kan worden in het formalisme van het vrije energie principe. Dit pogen Michael Kirchhoff en ik te doen in Hoofdstuk 8. We analyseren specifiek de enactivistische notie van autonomie als precaire operationele sluiting (Engels: *precarious operational closure*) en bestuderen of het vrije energie principe deze noties kan emuleren. Het resultaat is gemengd. Er zijn bepaalde kenmerken van operationele sluiting die zich goed laten vangen in het vrije energie formalisme zoals zelf-individuering (de capaciteit van operationeel gesloten systemen om zichzelf van hun omgeving te onderscheiden), maar andere kenmerken zoals de precariteit vallen buiten het net van het vrije energie formalisme. Dit geeft een indicatie van de limieten van wat we met het vrije energie principe kunnen beschrijven. Dit heeft tot gevolg dat het belangrijk is om buiten de conceptuele grenzen van het formalisme te treden als we de activiteit van organismen willen begrijpen. De bevindingen in deze thesis ondersteunen dus een instrumentalistische benadering van het vrije energie principe met een enactivistische insteek.

1 Introduction

In the morning, I wake up and get ready for work. When I turn on the light in the living room, I startle our pet rabbit in case my stumbling hasn't woken him up yet. I brush my teeth, do some stretches, pack my bag, feed the rabbit, get properly dressed before I proceed to get my bicycle for the half hour ride it takes me to appear at university. Most of these activities I do with relative ease as I have grown quite accustomed to the usual succession of my morning routine here in Antwerp. I'm accustomed to brushing my teeth before just about anything else in the morning, and I'm also accustomed to getting dressed before heading out to work, a habit that has never failed to be incredibly rewarding. Each of these activities in themselves is also stuffed full with aspects that I am attuned to, like the walking distance from our bed room to the bathroom, the lightswitch inconveniently placed *outside* of that bathroom; in fact, I'm even attuned to the usual distance I travel, the usual amount I close in on my target, with each and every step. On the flipside, my pet rabbit has grown accustomed to its own habits, and they do not always line up with mine, which sometimes leaves him startled, other times he's sitting ready at the place in the kitchen I usually go to when getting his food.

Of course, the specifics of my morning routine are not important for my thesis — although it's fun to know it involves a pet rabbit, no? What *is* important is the ubiquity of our attunement to interactional regularities displayed on multiple levels in even the 'simplest' of our daily activities. In the cognitive sciences, part of what we hope to explain and understand is how humans, but also rabbits and even single-celled organisms can display this well-attuned activity in their navigation of their respective environments. The *prima facie* obvious way to tackle this, is to do empirical studies, and investigate the particular activities of the particular organisms you're interested in. It is not self-evident how philosophy could contribute. Yet in 'doing the science', we (implicitly) draw on a theoretical framework, and this is crucial to what our science can tell us about our object of study. Depending on the framework we base our work in, we may not only draw wildly different conclusions from the same data (which is significant enough for the value of philosophical investigations), but also gather different data from the same experiment, or even construct an entirely different experiment to begin with. In fact, even what we consider the object of study depends on the theoretical framework: if I want to understand my pet rabbit's activities, do I need to observe him as he explores a new area, or do I only need to observe the spiking patterns in a neuroimaging study? The framework we adhere to will determine which aspects of the object of study are of relevance.

In this thesis, I shall primarily be concerned with two distinct but interestingly complementary cognitive science research programmes: radical embodied cognitive science (or more specifically enactivism), and Bayesian cognitive science (or more specifically the free energy principle). My aim in this thesis is to explore their mutual compatibility by looking at different theoretical proposals from the literature, analyzing their respective advantages and disadvantages, separating the key insights from the unwanted unclarities, incoherences or stark contradictions. Towards the end of this thesis, I will take the residue from each chapter, and construe the outlines of a novel approach to cognitive science that is deeply embedded in both enactivism and the free energy principle, yet unlike previous attempts to wed the two frameworks that exist in the literature. The very basic idea is that enactivism supplies the epistemology as well as the conceptual framework through which the study of organismic activity should occur, whereas the free energy principle provides the statistical toolkit that allows for formalization of the process under study, which may unearth statistical relations between the organism and its environment that are crucial for our understanding. Each chapter comes fully equipped with its own introduction into the exact theoretical details relevant for the respective topic of discussion. Yet before jumping right into the deep end, I want to briefly introduce the two frameworks so as to give a general idea of the direction the remainder of the thesis is headed in. I shall start with a broad outline of what I call the 'classical' approach. This will provide a reference point for both enactivism and the free energy principle so that we can see not only how they differ from one another, but also how both frameworks deviate from what is typically seen as the initial vantage point.[1]

According to what I call the classical approach, cognition is what happens between sensory stimulation, and acting based upon that stimulation. Following Hurley (2001), this is called the *sandwich model*. First, the agent in question gets an *input* from the environment via their senses, then, the agent processes this input allowing for the *planning* of an *action*, in which the agent changes the environment around them: the *sense-plan-act* sandwich, in which only the *plan* aspect is dubbed cognition (see for example Marr 1980 and Fodor 1975). This processing of the input entails forming an internal reconstruction of the input, attempting to represent the external environment as well as possible, which can be manipulated to plan a possible mode of action. Think of the pet rabbit that I startle when turning on the light. The

[1] None of these descriptions should be taken as exhaustive, nor should they be taken as representative of every position that could broadly be classified as an approach of that type. Many proponents of classical, enactive and free energy approaches disagree amongst one another concerning key issues, and there are many approaches that could be seen as a middle passage between different directions. However, they do give a general idea.

light radiating off the lamp, bouncing off of the objects populating the living room enters through his eyes. His brain receives the signal through the nervous system and forms an internal representation of the suddenly lit room. This causes an abrupt, and extreme change in the internal representation of the environment by way of the starkly increasing luminosity. Extreme and abrupt changes are, one could imagine, perhaps categorized as potentials for danger, so the current representation can be labeled as *dangerous*. Once the labeling is done, the brain can compute an *output*. Due to the *danger* label, my pet rabbit's brain will put the previously installed alertness mechanisms to work so as to plan a course of action. The course of action will involve quickly sitting upright and stretching out the body so as to be receptive to a larger area of causes of inputs, so that the potential threat can be properly recognized, and further actions may be planned accordingly. And thus, the sense-plan-act sandwich is complete.

This is a bit of a toy example, but it shows clearly some of the central theoretical commitments of the classical approach. The classical approach is *neurocentric*: cognition happens in the brain. The only sense in which the rabbit's environment matters is by way of the inputs it delivers, and the only sense in which the rabbit's rest of the body matters is by way of the types of inputs it is sensitive to, and the types of outputs it affords. As the brain is thought to act as an agent independent of the body, the classical approach is *seclusionist*: the brain is cut off from direct contact with the environment, as it only has access to the inputs delivered to it by the senses. As the agent is thought to be secluded from its environment with no direct access, the classical approach is *representationalist*: only by incorporating the inputs into an internal reproduction or representation of the external world can the brain cognize or plan the appropriate mode of action onto the world. The classical approach is *computationalist*: the manipulation and exploitation of the representations stored in the brain are achieved by way of a computational process. This can be e.g. by computing the distance of particular objects relative to the agent, computing the trajectories of moving objects or computing the muscle movements necessary to produce any specific output. As a computation is a move made within a formal system, the classical approach is *inferentialist*: cognition is a form of inference executed by the brain.

Even among proponents of the classical approach, nearly each of these terms is contested. What exactly does the term mean and how would the feature it describes work in a living agent? The questions concerning the what and how of representation, computation and inference all lie at the heart of the philosophy of cognitive science. Yet for our current purposes, it suffices that this brief overview provides a baseline to contrast the two relevant approaches with. I shall first discuss the enactive approach, which displays a strong contrast with the

classical approach, and finish with the free energy principle, which is a bit of an odd duck as its proponents' commitments vary wildly across the board.

Under the enactive approach, cognition is a concept that describes activity that displays adaptivity to the environment the activity is performed in, and cannot be understood without understanding the environment, the organism, and their respective interactional history (Di Paolo 2005; Di Paolo et al., 2017; see also Varela et al. 1991). To see how enactivism differs from the classical approach, let us return again to the startled pet rabbit. The pet rabbit is, before I switch on the light, very strongly attuned to the specific regularities pertaining to our living room at night time. This involves, for example, a rather dark and mostly still environment with the soft mechanical background buzz of machinery native to city life. His particular position is one that places him at about the center of the room, elevated about a meter above the floor level on a piece of furniture. As I switch on the light, the tight attunement to the environmental dynamics breaks off abruptly. A sudden break requires quick re-calibration. After all, in the rabbit's evolutionary and developmental interactional history, abrupt and severe breaking points in the environmental attunement are regularly followed by dangerous situations, and so, re-calibrating and checking out what's up may save the rabbit his life. As such, he sits up straight to survey the surroundings so as to reestablish grip on its environment, whilst displaying readiness to flee *via* either side of the piece of furniture that he is on.

The theoretical commitments that we can glean from this reframing of the toy example are as follows. The enactive approach conceives of cognition as *embodied* and *embedded* as opposed to classical neurocentrism. This means that the brain plays a role in cognition as part of the agent's body, which always figures in a specific environment. Pulling apart the body and its environment in analysis will not help us from understanding cognition any better. The agent is in *direct* contact with its environment, and is thus not classically secluded. As the brain is not an agent in itself, but the organism in its environment is the object of study, the question of 'access' disappears, as the organism is thought to be directly coupled to the environmental dynamics. By extension, there is no need to represent the environment internally: why would one use one's limited resources making an internal reproduction of the environment when the environment itself is *right there*? The entire process is also non-computational. Computation is a specific skill in our sociocultural heritage that did not exist before our ancestors started engaging in it. This means that agents do not compute unless they have been taught to do so by others in their community, broadly construed. Though my pet rabbit is clever, I have not spent my time trying to teach him computation. The same goes for *inference*, which is a broader term, capturing more distinct activities than computation does. Empirical research could show that

my pet rabbit is capable of inference —again, he is rather clever— but the currently described situation does not call for this. Crucial here is that inference is not thought to be an essential part of cognitive activity. Further, the aforementioned sense-plan-act sandwich has disappeared, or rather, it is eaten in full bites instead of its components being analyzed in separation from one another (Hurley 2001; O'Regan and Noë 2001). This is to say that the agent's sensing and acting appear as sensorimotor engagements, inseparable from one another.

The free energy principle is the odd one out among the three theoretical frameworks discussed here, as its proponents have wildly varying ideas as to what the principle entails with regards to the epistemological status of the agent, where to draw the agent's boundaries, or whether the principle even says anything about these matters at all. At heart, it is a formal definition of the dynamics that are essential to self-organizational systems, those that retain their structural integrity over time (Friston 2013). It starts with constructing a model of the object of study by organizing the variables that we associate with the target. The formalism allows us to pick out particular relations that appear over time between the variables associated with the organism itself and those associated with the environment. Yet it remains a matter of dispute what any of this means for the actual object of study. Nonetheless, there are specific proposals associated with the free energy principle that *are* classifiable. I shall briefly lay out a blunt version of the *predictive processing* description of the startled pet rabbit, and use that to broadly delineate some of the different positions held in the FEP literature.

Predictive processing takes the FEP's formalism to be literally at work in an agent's brain (Hohwy 2013; Clark 2016). The brain thus constructs, exploits and updates a representational model of itself and the environment. On the basis of this model, it produces probabilistically weighted (*Bayesian*) predictions of the input that it will encounter *via* the senses. The brain's primary (or only) business is to minimize the error of these predictions. As such, only prediction errors produce salience, and require interaction, when everything is predicted correctly, no action is needed. When the pet rabbit is asleep on its piece of furniture, its brain is continuously producing predictions with little error as it correctly predicts the relative darkness, the soft, mechanical background hum, as well as the homeostatic state of the rabbit's well-fed, well-exercised body. Yet when I switch on the lights, this creates a huge prediction error for the brain to process. This means that the brain needs to update its model of the current environmental dynamics. As its focus is on minimizing prediction error, it aims to quickly gain a grasp of the situation with the information that is available. This means that the brain selects actions for the body to maximize its grasp on the situation. The best way to be able to correctly predict what is happening is for the brain to get sensory stimulation from as

wide of an area as possible, so the brain decides for the rabbit's body to sit up straight and analyze the situation until it sorted out what is going on, and its prediction errors recede again.

In many ways this is very much like the classical approach: it is neurocentric and seclusionist, and involves computational inferences in the form of predictions over a representational model of the external world and the body. Yet it does not usually subscribe to the sandwich model of cognition, and describes the perception and action as components of the same prediction error minimization process. Also crucial is that it places *prediction* front and center as opposed to merely being reactive: an attempt to stay ahead of the curve.

Contrary to the predictive processing approach, one could consider the formalism of the FEP to merely be a useful descriptive tool of the statistical properties of organism-environment dynamics as we described them in our model as scientists. This is called the *instrumentalist* approach to the free energy principle as defended in chapters 5 and 6 (van Es 2020; van Es and Hipólito 2021, but see also Baltieri et al. 2020; Bruineberg et al. 2020). A middle passage of sorts is called active inference. In active inference, the fact that the statistical dynamics that show up in our models of organisms correspond to particular statistical computational processes, is taken to imply that organisms must engage in these statistical computations themselves in some way or another.. In a different sense, they could be said to argue that the free energy principle's formalism reveals the true statistical properties of nature. Of course, among active inference theorists there is plenty of variety in how the model or its relation to nature is exactly to be cashed out, and to what extent the model describes or corresponds to reality.

Now that we have surveyed the general theoretical background, I shall outline the structure of this thesis. The remainder of this thesis will be separated into 3 different parts. Part I discusses the representationalist approach associated with the free energy principle called predictive processing and a non-free energy inspired, but conceptually related approach called the embedded view, with a specific focus on the issues that come with their respective representational and/or inferential commitments. Part II concerns the viability of a more enactivism-inspired take on the free energy principle, and argues that we need to take the formalism to be nothing more than a useful tool for scientists, not something that is used or exploited in some sense by the organisms we study in their navigation of their environment. Part III shows what the free energy principle can do for us, and how its formalism can help us gain a clearer understanding of the dynamical relations between an organism and its environments as they interact.

More specifically, Chapter 2, co-written with Erik Myin, contains an analysis of the representational commitments of predictive processing, and argues them to be untenable. Chapter 3 zooms in on the proposed hierarchical nature of predictive processing, and argues it to be incompatible with the seclusionist view put forward in the literature. Chapter 4 discusses the embedded view, which is Orlandi's (2014, 2012, 2013) attempt to explain how we can explain an organism's sensitivity to environmental regularities without invoking representations or inference, whilst retaining the need for representation in situations that involve relating to things that are absent. In the chapter, I argue that is unnecessary to retain representationalist commitments as the embedded view provides us with all the tools we need to account for our sensitivity to the regularities Orlandi considers to be absent. Chapter 5 kicks off the Part II with a general analysis of the usage of the world 'model' in the free energy literature, and suggests that interpretations beyond instrumentalism overstep the bounds of what is afforded by the formalism itself. Chapter 6, co-written with Inês Hipólito, provides a systematic overview of the different takes one could have with regards to the models in the free energy principle formalism, and argues in favour of instrumentalism. An interesting implication of this analysis is that the question of whether the models imply representationalism ceases to matter under an instrumentalist approach of the free energy principle. In Part III explores the compatibility of the free energy principle and enactivism. Chapter 7, co-written with Jo Bervoets, argues that an enactive conceptualization of autism as primarily a sensorimotor atypicality provides extra weight to an ethical appeal to further inclusion. This display of the explanatory strengths of enactivism sets the stage for Chapter 8, which explores the extent to which the free energy principle's formalism could accommodate the enactive conceptualization of autonomy. I finish with some concluding remarks.

References

Baltieri, M., Buckley, C. L., & Bruineberg, J. (2020). Predictions in the eye of the beholder: an active inference account of Watt governors. In *Artificial Life Conference Proceedings* (pp. 121-129). One Rogers Street, Cambridge, MA 02142-1209 USA journals-info@ mit. edu: MIT Press.

Bruineberg, J., Dolega, K., Dewhurst, J., & Baltieri, M. (2020). The Emperor's New Markov Blankets. *PhilSci Archive*

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Di Paolo, E., Buhrmann, T., & Barandiaran, X. (2017). *Sensorimotor life: An enactive proposal*. Oxford University Press.

Di Paolo, E. A. (2005). *Autopoiesis, adaptivity, teleology, agency. Phenomenology and the cognitive sciences, 4*(4), 429-452.

Fodor, J. A. (1975). *The language of thoug*ht. Harvard university press.

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Hurley, S. (2001). Perception and action: Alternative views. *Synthese, 129*(1), 3-40.

Orlandi, N. (2012) Embedded seeing-as: Multi-stable visual perception without interpretation. *Philosophical Psychology, 25*(4), 555-573

Orlandi, N. (2013) Embedded Seeing: Vision in the Natural World. *Noûs 47*(4) 727–747

Orlandi, N. (2014) *The Innocent Eye: Vision is not a cognitive process*. Oxford University Publishing

van Es, T., & Hipólito, I. (2020). Free-Energy Principle, Computationalism and Realism: a Tragedy. *PhilSci Archive*

Part I

2 Predictive Processing and Representation: How Less Can Be More[2]

Authors

Thomas van Es 1

Erik Myin 1

1 Centre for Philosophical Psychology, Department of Philosophy, Universiteit Antwerpen, Belgium

Abstract

The ambitious, mathematically elegant unificatory proposal of Predictive Processing (PP) to account for perception and action seems to have taken the world by storm. Though many different varieties of PEM may be distinguished, most of them adhere to representationalism in one form or another. In this paper, we inquire into these representational foundations. We argue that PP is best understood in a non-representational way. We argue that the most popular way of construing representational content in PP, despite pretensions to the contrary, proliferates representations unacceptably. Next we show that PP's explanatory potential can be retained without positing representations. We thus show that PP can't have and doesn't need representations to do its explanatory work, and conclude that our efforts are better placed in furthering the programme of non-representational PP.

2.1. Introduction

Attempts to explain cognitive phenomena can be representational or non-representational. Representationalist approaches to all or some cognitive phenomena have been criticized for a few decades (see for example Gibson 1979; Varela et al. 1991; Hutto and Myin 2013 2017; Di Paolo et al. 2017). Nonetheless, wide-ranging representationalism continues to thrive as the mainstream position in cognitive science. In this paper, we investigate whether this is because the issues have been solved, with particular focus on predictive processing (PP). We argue that, despite significant effort, PP still can't have the representations it lays claim on. Further, we will defend that the relevant explanatory work can be done without representations. This shows that PP gains from a non-representational interpretation.

PP, in short, is a theoretical framework that places cognition, action and perception under a single banner of prediction error minimization (Hohwy 2013; Clark 2016). The standard interpretation is as follows. In order to maintain the system's homeostasis, the brain

[2] The text of this chapter has been published as van Es, T. and Myin, E. (2020) Predictive Processing and Representation: How Less Can Be More. In Mendonça, D., Curado, M. and Gouveia, S. S. (eds). *The Philosophy and Science of Predictive Processing*. Bloosmbury Academic

is thought to actively predict the barrage of stimuli entering the system. These predictions are tested against the actual stimuli, and the brain's primary (or only) function is to minimize the prediction errors that result from this testing. This prediction error minimization system is optimized using representational models that mirror the causal-probabilistic structure of the world. These models are continuously updated and fine-tuned as the brain detects prediction errors. Representations here are thought to perform their causal role because of their representational status (Gładziejewski 2016; Hohwy 2013; Clark 2016).

Whether it is legitimate to understand PP in a representational way is contested (Kirchhoff and Robertson 2018; Bruineberg et al. 2016; Hutto 2017). Central to the notion of representation is that it has *content*. On a standard understanding contentful representation is for a representation to have a target which it represents in a way as being such that it may not be so (Travis 2004). In having a target the contentful representation is intentional, in representing this target *in a way*, it is intensional. Supposing contents in a naturalistic framework requires being able to tell a naturalistically credible story of how both of the representational properties, intentionality and intensionality, can be natural properties. This requires being able to account for how these properties of intentionality and intensionality originate: how some naturally occurring properties genuinely *are* the representational properties. Moreover, it needs to show that the representations play a causal role *qua* representations—a causal role which derives from the fact that they are bona fide representations.

Many have argued that standard representational understandings of cognitive systems are inflated because they cannot provide an answer to these challenges about the natural origin and causal efficacy of contents (Hutto and Myin 2013; 2017; Rosenberg 2015). Nonetheless, some theorists have attempted to argue that these challenges can be met in the case of PP. According to a recent idea, representations in PP are analogous to cartographic maps, and attain their representational status by analogy to this prototype. This involves relating to the target by way of structural similarity, and further requires action guidance, decouplability and misrepresentation (Gładziejewski 2016; Gładziejewski and Miłkowski 2017). As it seems Gładziejewski's proposal has gained significant traction in the literature, we will take this particular proposal of representations in PP to be representative (Kiefer and Hohwy 2017; Lee 2018; Pezzulo et al. 2017; Wiese 2017; Wiese and Metzinger 2017; Williams 2018; Williams and Colling 2018, but see also Dołega 2017)

Here we will argue that this recent attempt to rescue the representational paradigm is not up to the job. Furthermore, we will argue that any explanatory advantages PP might have,

can be retained without representations. First, we will discuss general objections against representationalism. Second, we will discuss Gładziejewski's (2016) account of representation and analyze whether it solves those issues. Third, we will offer a sketch of an alternative approach to PP, to show that we can move forward without mental representations. Finally, we will conclude that PP is best understood in a non-representational way.

2.2 General objections against representationalism

The account of representation endorsed by philosophers and cognitive scientists alike is one according to which representing involves describing, characterizing, picturing "things as being thus and so—where, for all that, things need not be that way" (Travis 2004, 58). In other ways, representation involves specifying in a way which can be true or false, accurate or inaccurate, or something akin. Something that is specified in some way can be compared to the world, and it can either be correct or not correct, aligned or misaligned. Moreover, representation requires some form or medium of specification, that is description, characterization or picture—something which has "saying power". As this characterization of representation resonates with the way representation is widely understood philosophically, we will take it as offering legitimate criteria that any candidate representation in PP should minimally meet. Further, due to the causal/explanatory context in which representations are invoked in PP, defenders of representational PP need to establish that representations play a role *qua* representation: contents need to be causally efficacious.

Let's call "structural accounts of representation", all those that construe representation as a relation between some domain of reality and some distinct structured entity, where representation is construed in terms of some mapping, causal or other, between elements and or relations in the domain of reality, and elements or relations in the structure. A minimalistic structural account would count causal covariation between some wordly entity and some other entity as representational, while a cartographical map of a terrain would be an example of a more complicated form of structural representing. Pure structural accounts of representation face some well-known difficulties for meeting the criteria of our characterization (see for example Oliveira 2018 for more, and further references). A prominent issue is that they over-generate representations, as structural similarities are widespread. This applies not just to the most minimal, correlational, reading of similarity, but also to more complex readings. That is, there seems to be no principled limit on what can be seen as structurally similar to something else—and there seems to be no principled distinction between what can be seen to be

structurally similar, and what genuinely *is* structurally similar. Moreover, and independent of this, structural accounts of representation, should provide an explication of why the structural relations amount to being representational relations, over and above being *just* structural relations.

A way to accommodate these two concerns is to add a functional component to a structural basis. Such has become standard practice (see again Oliveira 2018). That is, most accounts of representation, while including a structural component, require that the structural elements and/or relations play some role in the functioning of the system of which are part. They should be a "*fuel for success*" (Miłkowski and Gładziejewski 2017, p. 339; Godfrey-Smith 1996; Shea 2007 2014). Adding function constrains the class of relations that could count as representational with respect to a pure structural account—representation is only to be found where there is function. Moreover, by introducing function, causal roles come into play, to potentially offer a solution to demand that representations be causally efficacious.

In order to assess whether structural-plus-functional theories of representation can successfully meet the challenge of showing that candidate representations meet the appropriate criteria for being a representation, the question should first be answered what is doing the representing in such accounts. Or: where are the candidate representations located in such an account?

Of course, the representation cannot just be identified with the structures, because then the allegedly different combined structural/functional view inherits all the problems of a pure structural account. That is, if the structural elements represent by themselves, all the questions raised with respect to structural representations can be raised again. So the function has to play a constitutive role for representation in a combined account. This leads to two problems, however. First, the need for adding function arises because structural similarities are cheap, and ubiquitous. Let us see whether a function plus structural similarities account sufficiently constrains the notion of representation.

Consider some examples. There are many structural similarities that will, *prima facie*, be excluded with this requirement of functionality. Think for example of the structural similarities between the dimensions of a countertop and that of a cardboard box. The relation is certainly not representational, and indeed, without further context the structural similarities are not a fuel for success of an action. But what if we decide to place the box on the countertop? The structural similarities are now a *fuel for success* in the action of placing the box. It is because of the structural similarities between the countertop and the box that the action is successful. Had the countertop been convex, the cardboard box most likely would have slid

off. In much the same way the structural similarities are a fuel for success in socks and our sliding them on our feet, door knobs and our turning of them to open doors, and our sitting on chairs. Yet none of these exploitable structural similarities are representations.

This overgeneration of representation extends further and encompasses any sort of adaptive behaviour. Any behavior that is adapted to the world contains structural similarities to whatever aspects of the world it is adapted to. These structural similarities, further, are, prima facie, a fuel for success. We are able to walk so easily due at least in part to the structural similarities that hold between the shape of our feet and the surface of the world. The same story can be told for any adaptive behavior.

Second, any combined structural/functional account will struggle with ascribing a causal role to representation. The move to include function, or use, in the definition of representation is *prima facie* convincing: if a representation is only the type of relation that enables usage, it has to have causal effect. There are now, however, two ways in which 'use' is employed in representations. This invites a circle that is difficult to escape from. A representation is defined by its *use*. Only when a composition of driftwood is *used* as a map, can the composition count as a representation for the structure of the terrain it is used for (Ramsey 2007). However, as representations figure in an explanatory context, the *usefulness* of a representation, in turn, is defined in terms of the representational relation it bears onto the target. In other words, the way the relations are used are what make them representational, and their representational status is what makes them useful. Put differently,

(A) a structural similarity is used, *because* of
(B) its representational relation to the world.

Further, a structural similarity is said to have

(B) a representational relation to the world, *because* of
(A) the way it is used.

This clearly shows the circularity of the argument. We ground (A) in (B), and, conversely, ground (B) in (A). The representational status of the structural similarities is thus grounded in our use, and our use of the structural similarities is grounded in their representational status. If representational status is to be grounded in use, then use cannot in turn be grounded in its representational status. This is viciously circular.

If the use of a relation is what grants the relation representational status, then there are two options. Either the use of the representation is based in something non-representational, or it is based in something representational. The representationalist thus faces a dilemma. She may attempt to ground the use of representations into something non-representational. One

option would be to ground its usefulness in, say, covariance (or some other characteristic of the relation). Yet this means that it is not the representational relation, but the covariational relation that actually figures in the causal explanation. The representational status of the relation has become causally irrelevant. There would be no point to ascribe representational status to a relation in which only its covariation is causally relevant. The same goes for any other characteristic invoked here—structural similarity instead of covariance, for example. The representationalist may then opt for the second horn, and attempt to ground the use of representations in their representational relation to the world. This ensures a position in the explanatory causal chain for the representation, but now the account of representation can no longer be use-based. If it remains use-based, this comes at the cost of confusing explanandum and explanans: the representational status and use now serve to explain one another circularly.[3] This does not mean that we argue representational status cannot ever be grounded in use. On the contrary, public representations attain their representational status by way of social norms under which particular objects are used as stand-ins for others, and are thus grounded in our practices, our use (Hutto and Myin 2017; Van Fraassen 2008). Yet the cognitive system's reliance on representations cannot be grounded in social norms in the same way (Tonneau 2012, p. 339). Mental representations in PP are thought to precede social activity, instead of being a product thereof. There is thus a clear disanalogy between a cognitive agent's activity and our personal level use of public representations. Thus for our current purposes, grounding representations in use is insufficient to qualify for full representational status.

2.3 Can PP Have its Representations?

Standardly, PP is viewed as a representational theory of cognition (Hohwy 2013; Clark 2016). Though most theorists working on PP take this as a background assumption, few people have attempted to unearth the foundations of this supposed representational character. There has, however, been one proposal that seems to have gained traction in the field (Kiefer and Hohwy 2017). This is Gładziejewski's (2016) proposal, in which he uses Ramsey's (2007) *compare-to-prototype* strategy to attempt to answer the job description challenge.[4] His prototype of choice is a cartographic map, to which he claims PP's models are analogous to. These models

[3] A similar objection is present in Di Paolo et al., (2017, p. 25). Oliveira (2018) argues with the same argumentative structure applied to the debate on the representational status of models in science.

[4] Lee (2018) argues that a structural account of representation can naturalize content. In this, he relies heavily on the notion of representation from Gładziejewski (2016) discussed here. Because of this, our criticism on Gładziejewski's proposal extends to Lee's reliance on it.

recapitulate the causal-probabilistic structure of the world. He argues that these models are structural representations that relate to what they represent by way of structural similarity, a broader category than, say, pictorial similarity. He acknowledges that structural similarity in itself is not enough to warrant representational status, and therefore gives a set of four distinct conditions for anything to count as a structural representation. Conjoinedly, he argues these conditions are what make cartographic maps representational, and by the compare-to-prototype strategy, the same should go for PP's models (2016, p. 570). Summing them up: structural representations (1) represent by way of structural similarity, "(2) guide the action of their users, (3) do so in a detachable way, and (4) allow their users to detect representational errors" (Gładziejewski 2016, p. 566). To cover all bases, we will discuss the four conditions separately and analyze whether they solve the problems mentioned above, or work towards a solution. Consequently, we can conclude whether the proposal has succeeded in naturalizing representations, thus countering the general concerns regarding the kind of representationalist account endorsed by Gładziejewski, or not.

The first condition is that structural representations represent their target by way of structural similarity. In particular, the structure that the brain trades in is thought to be the causal-probabilistic structure of the world (Gładziejewski 2016). This is intended to define the type of representations relevant here. As we have seen above, rather than excluding unwanted types of representation, however, structural similarities are 'cheap'; they are ubiquitous. This is acknowledged by Gładziejewski (2016) and the further conditions are intended to constrain the class of relevant relations.

The second condition is that the representation guides the actions of the user (Gładziejewski 2016). As we have seen above, the role of the addition of function is twofold. The first role is to emphasize the role of a representation's use, as opposed to being merely passive "mirrors of nature" (Gładziejewski and Miłkowski 2017, p. 338). The second role is to avoid with the problem of over-generating representations, by limiting "the class of representation-relevant similarities to *exploitable* similarities" (Gładziejewski and Miłkowski 2017, p. 338, emphasis in original). To recapitulate our earlier discussion, function does not fulfill either role. Action guidance does little to constrain "the class of representation-relevant similarities", because exploitable similarities are still nearly everywhere. It also cannot serve to ensure the causal efficacy of representations. This produces a vicious circle in which the representational status of a similarity is grounded in the similarity being used, and the use of the similarity is grounded in its representational status. This means that the remaining two

conditions, detachability and misperception, need to do all the necessary lifting to secure representational status.

The third condition to qualify for representational status is detachability (Gładziejewski 2016). This covers the notion that representations represent their target, regardless of the presence of that target. A cartographic map, for example, is thought to represent the terrain it is a map of regardless of whether this terrain is actually present to the user or not. Put differently, a representation is required to be usable *off-line*.

Presence is however not a simple unitary phenomenon. In any given room, the people that are in that room can be said to be present. Anyone counting as absent would be a person that isn't in the room. Some of these people may stand behind others, leaving some people obscured from a particular perspective. From that perspective, then, the stimuli pertaining to that person are absent. Presence in PP is of the latter sort: the presence of stimuli pertaining to the target. Because of this, according to PP, engagement with the obscured person requires the use of an internal representation to stand in for the lacking stimuli. Thus, for anything to be a representation, it needs to be capable of being used in the absence (of the stimuli) of the target.

The requirement to be able to use the representation off-line is intended to be a mark of distinction between representations and mere detectors or causal mediators. After all, a detector can only detect that which is *present*, whereas an internal representation should be available even in the absence of the target. The thought is thus as follows. Any behavior that is thought to require off-line cognition so that the stimuli of the target are absent, requires representation. Any representation, in turn, involves the possibility of off-line use. This creates a circularity, so that off-line cognition is defined in terms of requiring representation, and representation is defined in terms of being available for off-line cognition. Perhaps this problem could be lived with if there were additional justification for the representationalist position. However, the only justification given seems to be an argument by default, resting on the idea that representationalism is *the only game in town*. Non-representationalism is thought not to be able to account for off-line cognition.

Though not a strong argumentative strategy, it is reasonable to accept the stipulative definition above if representationalism is indeed the only possible explanation. There have however been a few non-representational proposals for what is termed off-line cognition that warrant discussion (for example Bruineberg et al. 2016; Kirchhoff 2018; Bruineberg and Rietveld 2014; Di Paolo et al. 2017, ch. 8; Van Dijk & Withagen 2015; see also Degenaar & Myin 2014). At the very least, the existence of non-representationalist alternatives requires the representationalist to put in more work to explain why the alternatives fall short and why

representations only fit the bill. Due to this, the circular stipulation defining representations in terms of off-line availability and off-line availability in terms of requiring representation lacks the further justification that is needed. This means that the condition of off-line use, detachability, without further support, does not solve nor work towards a solution for the arguments against representationalism.

The fourth condition is that representation can go wrong; in particular, that such an error should be detectable. In the case of cartographic maps, for example, an error will be detected when a pathway planned with the map does not lead to the intended destination. According to Gładziejewski, error detection in PP comes in the form of practical use (2016, p. 577-579). If an action is recruited on the basis of a particular prediction of the brain, yet the predicted outcome of the action *increases* prediction error rather than minimizes it, an error is detected. If one, for example, moves one's head about to get a cup of tea behind a laptop in view, and the cup of tea remains missing, the prediction that the cup was behind the laptop is detected to be erroneous.

The possibility of misrepresentation is crucial to the notion of representation. However, anything that meets only the first three conditions: structural similarity, action guidance and detachability can be described in non-representational terms (Gładziejewski 2016, p. 570). And there is nothing intrinsically representational about error detection, unless the error is supposed to be representational error from the start. There is nothing intrinsically representational about failure to grasp a cup, nor is there anything intrinsically representational about detecting that failure. A non-representational relation indeed does not necessarily become representational if one is to add in error detection as a fourth ingredient. Indeed, it is unclear from Gładziejewski's proposal when and how error detection, a sensitivity for reacting to misalignment, turns into mis*representation*.

More work has been done, however, on misrepresentation in PP. Kiefer and Hohwy (2017) recently argued for a naturalistic measure of misrepresentation in PP. This measure has been analyzed and shown to bottom out in Shannon-informational covariance by Kirchhoff and Robertson (2018). Shannon-information is not inherently representational, and covariance is insufficient to ground content (Hutto and Myin 2013). The discussion cited is rather technical, leaving a full exposition out of the scope of this paper. Essentially, it is as follows. Kiefer and Hohwy (2017) argue that the brain's continuous effort to minimize long term prediction error can be cast in models of Bayesian statistics. The brain approximates a probability distribution of the possible states of the world. The extent to which the brain's *approximation* of the prior probability distribution *diverges* from the *actual* posterior probability distribution is captured

by the Kullback-Leibler divergence. This is thus the degree to which the brain's *approximate* model diverges from the *actual* posterior model, and is thought to stand for the degree of misrepresentation. Kirchhoff and Robertson (2018) display that this divergence in itself is but a measure of the divergence of two covariational systems. In this sense, if A reliably covaries with system B, then information about system A will be information about system B. The divergence, in short, is "a measure of informational covariance", but falls short of providing a measure of misrepresentation (Kirchhoff and Robertson 2018, p. 277).

It thus seems that none of the four conditions actually help the representationalist to secure the status of representation. Yet there are two remaining questions to be answered. First, Gładziejewski's argument was not that the conditions separately would suffice for representational status, but only conjoinedly. As we have shown above, however, not a single condition seems to actually do the work it is intended to do by Gładziejewski (2016). It is reasonable to assume that, without additional argumentation to the contrary, if all the parts of a system do not work as intended, the whole will not work as intended either. Crucially, no reason is given in Gładziejewski (2016) *how* precisely the conjoining of the four conditions representational status engenders representational status where there avowedly is none before the conjoining. In justifying representational status, all the weight is carried by the analogy to maps. But maps are *found to be* be representational. The representational status of maps is taken as a given. This bring us to our second point: what about the compare-to-prototype strategy employed by Gładziejewski? Should we subscribe to this strategy, and agree on the analysis of the features of a cartographic map that make it representational? If so, how can it be that these features do not seem to secure representational status for models in PP while they do for maps? Is this not incoherent?

No it isn't, because not all use is equal. Some sociocultural practices are representational, because the aim of these practices is to describe the world in a certain way, that is evaluable for truth or falsity, accuracy or inaccuracy. Truth-telling practices, practices in which people offer description which can literally be true or false, are the paradigm of such. There is no point in justifying whether moves in truth-telling practices have truth conditions, because by definition, they have. Other social practices, such as those for making maps whose use can be shared, are analogous to truth-telling practices. A map-making practice prescribes how representers should go about to represent whatever can be represented in that practice, and how end-users should act based on what is represented. Because of a shared representational practice, bus stops will be represented by *this* item on the map, and distances in *that* way.

Because the practice determines ways of representing, users will be able to get to the bus stop from a distance by relying the map.

Clearly there are different uses of representations within a representational practice. Some uses create, or change representations, and other uses rely on existing representational practices, for example when deciding which route to take on the basis of for reading a maps. Crucially, however, both of these are uses *within* a representational practice. Even some uses within a practice are not representation-creating, but some are. Both uses, however are representational, and what makes them representational is their embeddedness in sort a practice that is a representational one

Crucially, the concept of functional "use" in the literature on representation in the way of Gładziejewski (2006) is a different one than use-within-a-representational practice. Even if it is analogous, even if it can be said to share features with use within a practice, this doesn't mean that it thereby shares all its features. Such is the nature of analogy. To begin with, there's a serious issue whether the kind of within-practice use the functional notion of use is most analogous to, is appropriate, because that kind of "consuming" use relies on there being existing representations and representational norms, rather than engendering such. More importantly, the reason why use in a representational practice constitutes representation, is because the practice is a representational one. The existence of such representational practices give us the idea of representation, just like the existence of truth telling-practices are what gives us the idea of truth evaluability.

Analogies might be drawn between functional use and use in representational practice, and between biological norms of adaptivity and sociocultural norms of accurate representation, yet there remains "a root mismatch between representational error and failure of biological function" Burge 2010, p. 301). Nothing in Gładziejewski (2016) shows that this idea of a mismatch is mistaken. To illustrate: because of a shared representational practice, bus stops will be represented by *this* item on the map, and distances in *that* way. Genuine misrepresentation, over and above malfunction becomes possible, as when one misrepresents a crossing as a bus stop. Mislocating a bus stop is a representational mistake, if it goes against the norms of the practice. In contrast, tampering with the neural antecedents of action might or might not lead lead to malfunction, but it doesn't violate representational norms—note that an unusual patterns of neural activation, one that goes against the "structural similarity" one can project on previous activations, might *add* functionality. Also, drawing the bus stop at the wrong place will be a misrepresentation from the moment it is drawn, while it will have to be found out whether an anomalous pattern of activation is functional or its opposite.

In short, the representational status of any object is attained in virtue of its use by agents, embedded in a social practice of using particular sorts of objects as representations for particular targets (Hutto and Myin 2017; Van Fraassen 2018; Tonneau 2012). If social practices are what lend objects their representational status, it is clear neural models do not qualify. After all, the neural models are thought to precede the practices that representations are a product of. This marks a clear difference between cartographic maps and neural models, and it is due to this disanalogy that cartographic maps may be said to represent a target terrain when being used in the relevant social practices, yet neural models may not lay claim to the same representational status.

The challenge for Gładziejewski (2016) was to provide a notion of representation that could solve the known issues with representations described in the previous section. These issues were that structural similarity is too broad, and if use is to be the grounding for representations, representations cannot also serve to ground use. We have seen that the first two conditions fall victim to exactly these objections. Gładziejewski (2016) anticipated this, and prepared two more conditions to cover the missing ground. We have shown above that neither detachability nor misrepresentation are up to the task that had been laid out. This means that, not only is PP's current notion of representation insufficient to warrant representational status, it also has not done well to solve the problems that were already known.

2.4. Does PP Need Representations Anyway?

Despite the aforementioned worries, we need to take care not to throw the baby out with the bathwater. The question is whether we can keep the explanatory appeal of PP whilst resisting the representational pull, the enticement to invoke representations (Di Paolo et al. 2017). If we believe the bulk of the PP literature, this issue is insurmountable (Clark 2015 2016; Seth 2013; Hohwy 2016). There has been, however, an effort to bring PP and non-representationalism closer together than initially seemed possible (Anderson 2017; Fabry 2017; Kirchhoff 2018a 2018b; Kirchhoff and Robertson 2018).

The primary explanatory appeal of PP is that it can account for our sensitivity to statistical regularities in the environment. The ongoing barrage of stimuli unfolds rapidly and dynamically, yet we seem to navigate the environment fluidly and skillfully, even when things don't always go as could be expected. PP's main explanatory resource is the internal model in which the statistical regularities are encoded. However, we have shown that the appeal to representations is unwarranted. This means that internal models and their purported

explanatory power meet a similar fate. How can we account for our sensitivity to such regularities if not by invoking internally encoded representations?

One option is to explain our sensitivity to statistical regularities in the environment by appealing to the environment that the organism is, and its ancestors have been, embedded in. Orlandi (2014 2012 2013) proposes the Embedded View (EV) to explain visual perception in particular, though it can reasonably be extended to cover other sensory modalities. EV explains visual processing by relying on the facts of the environment instead of internal representations of those facts (see also Myin and Degenaar 2014). There is a caveat: Orlandi (2014) argues we still need to invoke representations further down the line. We will argue that this is unnecessary with the explanatory tools EV offers. This puts us on the right track to understand how to continue without internal representation.

EV is explains the visual system's sensitivity to statistical regularities by relying on the facts of the environment themselves without encoding them internally. Orlandi states: "[r]elying on a fact means acting in accordance with a fact, and with a corresponding principle, without representing either" (Orlandi 2014, p. 3). This means that the organism can rely on particular statistical regularities, say, the co-occurrence of particular patterns of stimulation and a particular object, without representing this. An example she gives is the detection of edges in the environment. It is a fact that particular strong changes in light intensity are typically caused by edges. Historically, these changes in light intensity have been encountered so often in co-occurrence with the presence of edges that "it would be surprising if evolution did not take care of this". She continues, "[i]n this explanation, external facts account for why we pick up on edges rather than on something else. No additional representational resources are needed" (Orlandi 2014, p. 154). Thus, instead of representations, we appeal to the rich ancestral interactional history with the world to pick up on edges when confronted with particular changes in light intensity. We thus appeal to the environment the organism is embedded in to explain the organism's sensitivity to statistical regularities, rather than merely to a subpersonal part of the organism-environment system.

This may serve well to explain evolutionarily developed sensitivities, but PP accounts for regularities that the organism can become sensitive to ontogenetically even within an hour (Mole and Zhao 2016). Indeed, one of the features of modeling in PP is continuous updating to reflect newfound evidence, allowing for a flexible approach to a dynamically unfolding world. The model is thus highly malleable. As we will show, EV is well equipped to account for the flexibility of our sensitivity to statistical regularities.

The central idea, developed in Orlandi (2014) is that the system is *wired* to be sensitive. This means that the perceptual system has many features that "developed, and continue to develop under evolutionary and environmental pressure" (Orlandi 2014, p. 3). In particular, Orlandi likens the wiring of the perceptual system to the manner in which a connectionist network is wired. She states

> We can imagine (...) a connectionist network trained to detect something by being repeatedly exposed to it. Such training causes the network to display characteristic patterns of activation where low-level configurations are associated with high-level ones and *vice versa*. We can then think of a high-level state as 'checking' the pattern of activation at the lower, sensory level where this ultimately just means that the high-level state activates in a way that is more or less compatible with the lower-level pattern of activation. (Orlandi 2014, p. 88)

This means that a connectionist network can be trained by repeated exposure to become sensitive to particular patterns of activation. The 'high-level states' come to co-vary reliably with lower-level patterns of activation, such that it reliably detects particular features. We can see that in this case, the interactional history of the system allows it to develop a sensitivity to particular patterns, without encoding these sensitivities internally (Orlandi 2014; Ramsey 2007). Yet the system does require a training signal in the first place. Yet much like PP, for a connectionist network "the world itself (...) provide[s] the 'training signal' you need" (Clark 2016, p. Contrary to PP, it should be noted, is that the connectionist network's transition between lower-level and high-level states is not regulated by a representation. It is a transition afforded by a strong connection between the two levels (Orlandi 2014, p. 153).

It is important here to distinguish a wired system from a *hardwired* system. A hardwired system would be wired for a particular task without allowing for this particular configuration to be *re*wired, adapting to a new situation. There is nothing in the notion of a wired system however, that necessarily constrains the system's malleability. In fact, connectionist networks seem to be a particularly good example of how this could work. Orlandi states that connection networks "come to be sensitive to the presence of mine echoes by simply adjusting their connections. Similarly, pigeons may stop pecking on a key as a result of extinction or counterconditioning without following any rules" (Orlandi 2014, p. 147). What this means is that *repeated exposure* to the training signal is how the system comes to detect the relevant features of the environment in the first place. It also shows that the network's sensitivities are 'updated', or rather, that they adapt to a dynamically unfolding world so that the 'high-level states' that detect, say, eatability, do not co-activate anymore in the presence of the patterns of stimulation pertaining to a set of keys. In the perceptual system, this means that the organism's

interactional history with the world itself can account for its sensitivity to the encountered regularities. These regularities, again, need not be encoded internally (similar ideas have been expressed by many authors disagreeing with representationalism, including Gibson (1979) and Skinner (1953)).

We can now also see how it can be that the perceptual system can sometimes be misaligned with the world. This happens when a 'high-level state' co-activates with a particular pattern of stimulation that is not caused by the same feature in the world it typically co-activates with. Think of the occurrence of edges, for example, a robust statistical regularity. Realistically drawn scenes in street art, for example, may be drawn so as to give the illusion of depth from a particular perspective. What happens here is that the pattern of stimulation has the changes in light intensity that typically co-occur with edges, causing the 'high-level states' that track edges to activate. Upon closer inspection, by moving about, changing our perspective or touching the surface on which it is painted, we notice that these new patterns of stimulation are actually quite different from those pertaining to edges. Here we detect the error in our perception without requiring to match an internal representation with a world that is out of reach. It is mis*perception* without mis*representation*.

In sum, by relying on the facts of the world we can explain how we become sensitive to particular statistical regularities in the world, both phylo- and ontogenetically as well as how sometimes we are misaligned with respect to the environment. Nonetheless, Orlandi argues that we still need to invoke representations. Specifically, representations are needed to account for two distinct features of the visual system: 1) completion, in which an organism interacts with a whole object despite only its (sometimes partially occluded) perspectival front side impinging on its senses (Orlandi 2014, p. 127), and 2) taking as, in which we *take* the light reflected off a cow that impinges on our senses *as* coming from a cow (Orlandi 2014, p. 150).

To account for completion, Orlandi argues, there are parts of the object that are not currently present to the organism, that the system needs to represent internally (2014, p. 127). The idea is that, because these parts of the object are *absent*, there is no feature of the system that could *detect* those parts of the object for the relevant high-level state to co-activate. However, as we have argued elsewhere this need not be an issue (Hutto & Myin 2017, Chapter 7, van Es (2019)). The signal with which the system was trained to co-activate with was always a pattern of stimulation that only ever involved a perspectival front side of an object, and oftentimes these objects were partially occluded. The interactional history with an object that allows for the sensitivity to the particular regularities is based on widely varying sorts of interactions with the particular object, so that a 'high-level state' that detects, say, cups, co-

activates with patterns of stimulation that are caused by mere front sides of cups, partially occluded cups, upside down cups etc. This means that we can rely on the fact that we are only confronted with a perspectival front side of 3D objects and that regularly these objects are partially occluded to explain why, nonetheless, the relevant 'high-level state' still picks up on the presence of the object. Just as we do not need to encode the rule that 'specific changes in light intensity pertain to edges', we also do not need to encode a rule that 'partially occluded front sides pertain to full objects'. Instead, we can rely on the facts of the environment that the rigidity of objects is statistically robust.[5]

Our capacity to take a particular pattern of stimulation as coming from, say, a cow, also requires representation, according to Orlandi (2014, p. 150). The idea is that, for this to occur, there needs to be an internal representation of COWness with which the pattern of stimulation can be matched so as to count as a token of that type. Yet here too our interactional history with the world seems to cover all that is needed to account for what happens in the perceptual system. For whenever we have encountered patterns of stimulation that pertained to cows, they co-occurred with the presence of cows, or images of cows. This allows us to perceive a cow, and interact with the cow as we have interacted with previously met cows. All of this does not, however, cover the categorization of individual cows as tokens of a type. But, we argue, though such ability to apply a category is representational, it is not part of the visual system. Instead, this representational ability is part of a larger socio-cultural and lingual practice in which particular objects, features and scenes are given particular names or labels. This means that we do not need to invoke internal representations for our categorizing a cow as a particular instantiation of COWness, we need only invoke public, lingual representational practices, similar to cartographic maps.

In sum, we provide a non-representational alternative to account for our sensitivity to statistical regularities. We have shown that this can account for malleability, misperception, completion and 'taking as' to the extent that it involves the perceptual system. This means that we do not need to invoke representational models. However, there is a sense in which an organism would *be* a model, if it is responsive to statistical regularities in the way described. Locating models in PP at the organism level is not uncommon (Friston 2013, p. 213). But this claim is only distinctive if it is asserted that the organism *embodies* a model, in contrast to it

[5] Interestingly, Orlandi (2012) argues that the rigidity of objects is one example of the visual system's reliance on statistical regularities, which would presumably make invoking representations redundant. It seems that if we can rely on objects typically being rigid, we can also rely on the front sides we are confronted with to be caused by a rigid object: it pertains to the same feature of the visual system.

containing a model that the organism uses, or that in some way guides its actions. Rather, in the unfolding of action, the state of the world could be read off, or inferred, by an external observer. In fact, such inferring of how the world is could probably take place by observing aspects of the organism's brain activity. Hutto describes such an external perspective in which the brain can *be* a model through the lens of PP as follows:

> "the brain might be used as [a] model by scientists in the sense that they could use brain activity to make reliable predictions and claims about how things stand with the world on the basis of their background knowledge. But if REC is right that is not what the brain itself does in supporting basic cognition." (Hutto 2017, p. 13)[6]

There might be more to invoking this perspective than making a philosophical point about the relation between organism and model. Focusing on which variables in the environment are systematically tracked by the organism, where such tracking is enabled by systematic correlation between environmental variables and brain processes, can teach observers a lot about what the relevant environmental properties are for the organism, and about the phylogenetic and ontogenetic causes of its current activity. If some property is systematically tracked, it might have been selected to be tracked, or a tendency to track properties of that kind might have been selected, for example. By investigating the correlational or tracking structure of the environment/brain relations, it might be possible to separate what genuinely matters for this organism from what only looks to matter.

It is further possible to interpret some of the most celebrated work in cognitive science as at least partly clinging to this methodology. Marr's (1982) theory of vision, for example, might be given a correlational reading, so that the various stages of visual processing it describes are not about building up more complex representations, but about tracking more complex properties, much like the connectionist networks we have seen above (Orlandi 2014, Note 3). The search for environmental variables organisms are sensitive to is of course an integral part of much ecological psychology already—paying increased attention to the brain as enabling this sensitivity is complementary, not contrary to such research (Gibson 1979). Though Marr's approach and ecological psychology are standardly seen in contrastive terms, the proposed perspective forms a unifying umbrella covering both of them. That is, abandoning the commitment to universal representationalism holds integrative potential.

Rosa Cao has argued against a common interpretation of single neuron activity in the following way:

[6] REC is a radically non-representational approach to basic cognition expounded in Hutto and Myin (2013 2017).

> "It seems more reasonable to … give up the common interpretation that a single neuron is doing something like representing 'a very small piece of the world outside the organism' though indeed its activity may be well correlated with the structure of that small piece of the world. Instead, the role of the single neuron is more like that of the man inside Searle's Chinese Room—taking inputs and systematically producing outputs in total ignorance of their meaning and of the world outside. The single cell responds to local regularities and rewards, which as a result of evolution have become coordinated (in some complicated fashion) with external regularities and distal rewards for the whole organism. (p. 66)"[7]

What is said here regarding single cells applies to collections of cells, and to neural processes on multiple scales. They too are in the business of enabling coordination between what happens at various levels of coordination with "external regularities and distal awards for the whole organism". In fact, depending on contingencies of empirical support, a modified, that is non-representational, PP seems fit to allow us to learn much about how such multi-scale organism-environment coordination is possible.

2.5. Conclusion

In this chapter we have reviewed whether PP's use of representations is warranted. We have first outlined general objections against representationalism, focusing particularly on the structure plus function account of representation. Second, we analyzed Gładziejewski's (2016) account of representations in PP as it has gained significant traction in the literature. Here we argued that his account does not do the necessary explanatory lifting to combat the general objections. In particular, the first two conditions for representation, structural similarity and action guidance struggled with the general difficulties, and the further two conditions, off-line use and misrepresentation were not found to suffice to carry the additional weight necessary for representation. Finally, we argued that the pivotal analogy with an allegedly prototypical representation, cartographic maps, goes awry. It leaves out the social environment in which representational practices come to be and give them representational status: a dimension not available to justify the representational status of neural models. In the final section we have outlined a positive alternative, with particular focus on why PP does not *need* its representations. Drawing on Orlandi's work on the Embedded View (2014 2012 2013) we argue that sensitivity to statistical regularities can be explained by relying on the environmental facts without encoding them internally. Further, we have argued that there may be a place for models in a non-representational approach to cognition, but only as used by scientists, external

[7] This passage is cited in a related context in Hutto and Myin (2017, p. 238-239).

observers, in their investigation of the organism. Finally, we have we proposed that stepping away from representationalism offers potential for unifying approaches to perception.

References

Anderson, M. (2017) Of Bayes and bullets : an embodied, situated, targeting-based account of predictive processing in Metzinger, T. and Wiese, W. (eds.). *Philosophy and predictive processing*. Frankfurt am Main : MIND Group.

Bruineberg, J. and Rietveld, E. (2014) Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience 8* (599), pp. 1-14.

Bruineberg, J., Kiverstein, J.D. & Rietveld, E. (2016) The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese*. doi:10.1007/s11229-016-1239-1

Clark, A. (2015). Predicting peace: The end of the representation wars. *OPEN-MIND, 7*, 1–7

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

Degenaar, J., & Myin, E. (2014). Representation-hunger reconsidered. *Synthese*, *191*(15), 3639-3648.

Di Paolo, E., Buhrmann, T., Barandiaran, X., (2017) *Sensorimotor Life: An enactive proposal*, Oxford University Press.

Dołega, K. (2017). Moderate Predictive Processing. In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:10*. Frankfurt am Main: MIND Group

Fabry, R. (2017) Transcending the evidentiary boundary: Prediction error minimization, embodied interaction, and explanatory pluralism, *Philosophical Psychology, 30*:4, 395-414

Frigg, R. and Hartmann, S., (2018) Models in Science in *The Stanford Encyclopedia of Philosophy (Summer 2018 Edition)*, Edward N. Zalta (ed.)

Friston, K. (2013). Active inference and free energy. *Behavioral and Brain Sciences*, *36*, 212–213.

Gibson, J;J. (1979), *The Perception of the Visual World*. Lawrence Erlbaum.

Godfrey-Smith P (1996) *Complexity and the function of mind in nature*. Cambridge University Press

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

Hohwy, J. (2016) The Self-evidencing brain. *Noûs, 50*(2), 259-285

Hutto, D.D. (2017). Getting into predictive processing's great guessing game: bootstrap heaven or hell? *Synthese*, 1–14. doi: 10.1007/s11229-017-1385-0

Kiefer, A. and Hohwy, J. (2018) Content and misrepresentation in hierarchical generative models, *Synthese* 195: 2387-2415.

Kirchhoff, M., (2018a) 'The body in action: Predictive processing and the embodiment thesis' in Newen, A., De Bruin, L., and Gallagher, S. (eds) *Oxford Handbook of Cognition: Embodied, Extended and Enactive*. Oxford University Press.

Kirchhoff, M., (2018b) Predictive processing, perceiving and imagining: Is to perceive to imagine, or something close to it? *Philos Stud 175*:751–767

Kirchhoff, M. and Robertson, I. (2018) Enactivism and predictive processing: a non-representational view, *Philosophical Explorations, 21*:2, 264-281

Lee (2018) Structural representation and the two problems of content. *Mind and language,* 1-21

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information.* Freeman.

Myin, E., & Degenaar*, J*. (2014). Enactive vision. In *The Routledge handbook of embodied cognition/*Shapiro, Lawrence [edit.] (pp. 90-98).

Nunn, T. P. (1909-1910) Are secondary qualities independent of perception? *Proceedings of the Aristotelian Society, 10,* 191-218.

Oliveira, G. S. (2018) Representationalism is a dead end. *Synthese*, 1-27

Orlandi, N. (2012) Embedded seeing-as: Multi-stable visual perception without interpretation. *Philosophical Psychology, 25*:4, 555-573

Orlandi, N. (2013) Embedded Seeing: Vision in the Natural World. *Noûs (47)*4 727–747

Orlandi, N. (2014) *The Innocent Eye: Vision is not a cognitive process*. Oxford University Publishing

Pezzulo, G., Donnarumma, F., Iodice, P., Maisto, D. and Stoianov, I. (2017) Model-Based Approaches to Active Perception and Control. *Entropy, 19*(6), 266

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

Rao and Ballard, (1999) Predictive Coding in the Visual Cortex: a Functional Interpretation of Some Extra-classical Receptive-field Effects. Nature Neuroscience 2(1):79-87

Seth, A. (2013) Interoceptive inference, emotion and the embodied self. *Trends Cogn Sci. (11)*:565-73

Shea N (2007) Consumers need information: supplementing teleosemantics with an input condition. *Philos Phenomenol Res 75*:404–435. doi:10.1111/j.1933-1592.2007.00082.x

Shea N (2014) Exploitable isomorphism and structural representation. *Proc Aristot Soc XIV:*77–92. doi:10.1111/j.1467-9264.2014.00367.x

Skinner, B. F. (1953). *Science and human behavior* (No. 92904). Simon and Schuster.

Travis, C. 2004. The silence of the senses. *Mind 113* (449): 57–94.

van Dijk, L., & Withagen, R. (2016). Temporalizing agency: Moving beyond on-and offline cognition. *Theory & Psychology*, *26*(1), 5-26.

van Es (2019) The Embedded View, its critics, and a radically non-representational solution. *Synthese*.

Wiese, W. (2017) What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences 16*: 4, pp 715–736

Wiese, W. and Metzinger T. (2017). Vanilla PP for Philosophers: A Primer on Predictive Processing. In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing: 1*. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573024

Williams, D. & Colling, L. (2018) From symbols to icons: the return of resemblance in the cognitive neuroscience revolution, *Synthese 195*:5, 1941-1967.

3 Minimizing prediction errors in predictive processing: from inconsistency to non-representationalism[8]

Author

Thomas van Es 1

1 Centre for Philosophical Psychology, Department of Philosophy, Universiteit Antwerpen, Belgium

Abstract

Predictive processing is an increasingly popular approach to cognition, perception and action. It says that the brain is essentially a hierarchical prediction machine. It is typically construed In a representationalist and inferentialist fashion so that the brain makes contentful inferences on the basis of representational models. In this paper, I argue that the predictive processing framework is inconsistent with this epistemic position. In particular, I argue that the combination of hierarchical modeling, contentful inferentialism and representationalism entail an internal inconsistency. Specifically, for a particular set of states, there will be both a representation requirement and not. Yet a system cannot both be required to represent a certain set of states and not be required to represent those states. Due to this contradiction, I propose to reject the standard view. I suggest that predictive processing is best interpreted in terms of reliable covariation instead, entailing an instrumentalist approach to the statistical machinery.

3.1 Introduction

Predictive processing (PP) as an approach to explaining cognitive phenomena has seen a steady rise since its original formulation as predictive coding in 1999 (Rao and Ballard).[9] According to PP, the brain is essentially a prediction machine, and its only task is to best minimize prediction error in the long run (Hohwy 2013). Its appeal stems from its promise to explain perception, action and attention under the single banner of prediction error minimization, as well as from offering a mathematically elegant approach to the brain's computational activity. There has been quite some work in establishing empirical verification, and the prospects seem positive (Hohwy 2016, p. 260).

PP is at heart a hierarchical computational theory of neural processing, that is often situated as the neural component within the larger story of the free energy principle (FEP)

[8] The text of this chapter is published as van Es, T. (2019). Minimizing prediction errors in predictive processing: from inconsistency to non-representationalism. *Phenomenology and the Cognitive Sciences*, 1-21. https://doi.org/10.1007/s11097-019-09649-y

[9] The term predictive processing was first introduced in Clark (2013) to distinguish the action-oriented approach defended there from the more passive perceptual approach linked to the term predictive coding.

(Friston 2012; Friston and Siebel 2009). Free energy here is analogous to the notion found in thermodynamics, but is conceived in an information-theoretic sense. According to the second law of thermodynamics, any closed system's entropy will increase over time. The resistance to this increase in entropy, or minimization of free energy, is what characterizes living systems according to the FEP. In resisting entropy, living systems remain within livable bounds. In the long run, of course, living systems will die, after which their entropy will increase. Hohwy (2016) has argued that inherent to the central tenets of PP and the FEP is a seclusionist, representationalist, and inferential view[10]: only vicariously (through the senses) can the mind come to know anything of the world external to it by way of representation. Representation is a widely contested term. In this paper, I shall use it to mean something that, minimally, has content with truth, accuracy or veridicality conditions. That is, minimally, anything that represents a target system must do so in a way that the target system may not be so (Travis 2004). In this paper I analyze an implication of Hohwy's setup of PP. I argue that the hierarchical conceptualization of PP does not mix well with Hohwy's take on inferentialism and representationalism as construed in Hohwy (2016). We find an internal inconsistency so that there both is and is not a representation requirement for certain sets of states. This calls some of the central theoretical posits of Hohwy's view into question.

If Hohwy is right, and PP is inherently bound to these theoretical commitments, then PP, due to being internally inconsistent, is doomed to fail. Hohwy's construal of PP has been challenged, however, and I will show there is reason to believe we can use the PP framework within the FEP consistently (Anderson 2017; Fabry 2017; Bruineberg et al. 2016). I will briefly sketch an alternative view on PP, hinting that there may be a way out. This requires us to rid the framework of representations, alter the notion of inference, and, for clarity purposes, I argue to take an instrumentalist position on the statistical machinery of FEP. Before this, I shall give a brief introduction into the central notions of PP in general, focussing on a few of the notions that are vital to my current discussion of Hohwy's take on PP in particular. I will then analyze this view and expose the inconsistency, that primarily depends on Hohwy's epistemological position. I will finish with a suggestion of an alternative view on PP, and explain why this view fares better. We will see that the FEP can do without representation and without contentful inference as construed by Hohwy (2016).

[10] The formulation may seem redundant. It may seem that if a system is inferential, it is necessarily also representational. We will see below that, in the PP and FEP literature, this is not always the case due to a particular technical notion of inference.

3.2 The Basics of Predictive Processing

Predictive processing (PP) has since its pioneering formulation by Rao and Ballard (1999) been reformulated and reshaped in a wide variety of ways. In this paper, I will target specifically Hohwy's view as proposed in (2016; in line with 2013). As such, my explanation of PP will remain close to Hohwy's interpretation of PP. In overcoming the objection I raise for Hohwy's view, I shall also discuss the Free Energy Principle as discussed in a few recent publications (Kirchhoff and Kiverstein 2019; Hesp et al. 2019 and Ramstead et al. 2018, see also Bruineberg and Rietveld 2014; Bruineberg et al. 2016; Ramstead et al. 2019; Ramstead et al. 2016; Kirchhoff and Robertson 2018; Kirchhoff et al. 2018).[11]

3.2.1 On predictions

To understand Hohwy's view on PP, it is useful to understand it as a solution to the supposed underdetermination problem for perception. This problem is usually described as follows. Take any particular retinal image. This particular image is consistent with an infinitude of possible actual states of the world. The retinal image I have of my office is also consistent with the actual objects being mere holograms, realistic cardboard cutouts etc. Yet the world is only in one particular state, and that is the one we actually perceive. The question that arises is then, as Hohwy states it, "[h]ow does a system such as the brain manage to use its sensory input to represent the states of affairs in the world" (Hohwy, 2016, p. 259)? A traditional answer is by way of inference. The brain has encoded certain representations of the world, and will use those in some way to get at the actual states of affairs. Many established views argue that this process works bottom-up: the signal entering our system via our sensory channels is the primary source of our perception. In our processing of, say, a visual signal, we may detect particular features, and build step by step a 3D representation of the world that has caused this signal (Marr 1982 describes visual perception roughly along these lines).

PP turns the picture upside down, placing a strong emphasis on top-down influences. This means essentially that our (often unconscious) notions and ideas of how the world works play an essential role in our perceptual processing. Indeed, sensory processing is "an active and highly selective process. … the processing of stimuli is controlled by top-down influences"

[11] Since some of the most influential works in PP published in 2013 (Friston 2013; Hohwy 2013; Clark 2013), the literature has expanded rapidly. I will discuss Hohwy's seclusionist approach and touch on a more recent approach proposed by Kirchhoff, Ramstead and colleagues (see references in text) that rejects neurocentrism, seclusionism, and arguably representationalism (more on this in Section 4) as well as a radical enactive proposal by Gallagher and Allen (2018) that suggests rebranding the approach to *predictive engagement*.

(Engel et al. 2001). In practice, it is the cascade of top-down predictions and bottom-up correction signals stemming from the world the system is embedded in that marks the PP approach (Clark 2013).

More specifically, Hohwy suggests that the brain engages in a Bayesian probabilistic form of inference to the best explanation, akin to what a statistical scientist does (2016, p. 261-263). In solving the underdetermination problem, the brain forms hypotheses about the possible causes for the retinal image. These hypotheses are informed by the brain's knowledge of the world. This knowledge is captured in a *generative model*, which is a statistical mapping over probability distributions that pertain to possible states of the extraneural world and how they relate causally. This model is thought to be representational, and held, updated and manipulated by the brain, in Hohwy's view. The model represents the causal-probabilistic structure of the world by way of structural resemblance (Gładziejewski 2016; Kiefer and Hohwy 2017). Relying on this model, the brain is able to pick what, given the current situation, seems most likely to be the cause of its inputs. This winning hypothesis, the prediction, based on the knowledge gained through earlier interactions, give the generative model what is called *temporal thickness*, as it's both directed to the future and the past in making sense of the present (Friston 2013). This prediction will then be tested against the actual input entering the system via the senses (both extero- and interoceptively, Seth 2013). In practice, the brain is likely to predict certain aspects correctly, other parts wrongly. This constitutes a prediction error.

Prediction errors play a central role in PP, and drive the system. The central, perhaps even only, goal of the brain is to reduce these prediction errors. Hohwy suggests there are two ways to reduce prediction errors. A personal-level example can help explain these two ways, though keep in mind that prediction error minimization is thought to be a strictly subpersonal, neural process. Imagine you walk into your office, expecting to see a cup of tea on your desk. Your best guess at what's out there, will then involve a cup of tea. If you, to your surprise, notice there is no such thing, there are two ways to react: either you can accept that there is no cup of tea, which relates to *perceptual inference*, or you can move about to see if the cup was occluded by, say, your laptop that was in the way, or even move to the kitchen to get yourself a cup of tea, which relates to *active inference*. In Hohwy's view, these are two sides of the same prediction error minimizing coin.[12] In perceptual inference, the brain will *infer* there is no cup

[12] It is important to point out that this is a vital point of disagreement between later FEP theorists and Hohwy. Though Hohwy concedes that active inference is vital to maintain the organism in the long run, and perceptual inference can help alleviate prediction error short term, he regards neither as superior to the other (Hohwy 2016, p. 280; Hohwy 2017). Later FEP theorists will argue that perceptual inference is only a step in the process of

of tea, and update its model of the world accordingly. The brain's next best guess, hypothesis or prediction will simply not involve a cup of tea in the office. In active inference, the brain will infer bodily movement, that is, make a prediction that is conditional on bodily movement. This allows the brain to sample different aspects of the environment to reduce prediction error. As Hohwy says,

> This prediction is sent to classic reflex arcs, which initiate movement until the proprioceptive prediction is fulfilled and the arm is in the predicted position (Friston, Daunizeau et al. 2010, Friston, Samothrakis et al. 2012). This is what action is: minimization of prediction error through changing the body's configuration and position. (Hohwy 2016, p. 279)

Active inference thus plays a vital role in the system. Without it, the brain would not have a mode of action in the world, and would not be able to sustain itself when interoceptive hunger signals turn into massive prediction errors of starvation and eventually death (Hohwy 2013; 2017.

In Bayesian terms, the initial predictions constitutes the *prior* probability distribution, and the update after having reduced the error is called the *posterior* probability distribution.

Prediction error minimization is realized hierarchically on multiple levels. This means that, in visual processing for example, on lower levels the brain may predict the incoming signal's colours, surfaces and edges, whereas higher levels may involve predictions regarding particular objects, like teacups or laptops, or objects-in-a-larger-context like tea-cups-at-a-tea-party or tea-cups-in-an-office-room. This computational hierarchy corresponds with cortical layers in a neurophysiological sense (Hohwy 2013, p. 24; 2016, p. 273). This hierarchical layering will play a large role in the argumentation below.

3.2.2 On inferences and seclusion

Though not exclusive to, but nonetheless particular of Hohwy's proposal of PP is environmental seclusion (Hohwy 2016 2013). The metaphor Wiese and Metzinger choose to describe this feature is one borrowed from Dennett (2013; Wieze and Metzinger 2017). They liken the brain's situation to that of a person manning a giant robot inhabiting a dangerous world, and her success in piloting the robot will decide her life. The only way she is in touch with the external world, then, is via the robot's sensory displays. Switch the personal level of interaction of the human with the robot to subpersonal interaction of the brain with the body, and we have the type of seclusionism that Hohwy's version of PP endorses. The extended

active inference, which maintains the organism in healthy bounds for as long as possible (e.g. Kirchhoff and Kiverstein 2019 and Bruineberg et al. 2016).

implications of this include solipsism, Cartesian skepticism, and, of particular relevance for this paper, representationalism and inferentialism (Hohwy 2016, p. 259).[13]

The boundary by which we are secluded from the world Hohwy terms an *evidentiary boundary*. This evidentiary boundary relies on the notion of *self-evidencing*. This is a feature typical of inference to the best explanation, a statistical form of which the brain is thought to engage in (Hohwy 2013, 2016). Say there is some piece of evidence e_i that requires explanation, and a hypothesis h_i that best explains this. As h_i explains e_i, it provides evidence for itself. That is, the presence of e_i may be cited in support of h_i, and h_i may in turn be used to explain the occurrence of e_i. They make what Hohwy terms an explanatory-evidentiary circle (EE-circle) (Hohwy 2016, p 264). It is explanatory-evidentiary because the hypothesis *explains* the evidence, which in turn is *evidence* for the hypothesis. This may seem vicious at first, but it actually is not. Say the evidence is the missing cup of tea, and my hypothesis is that I must not have made any tea. To explain the missing tea, I will cite my hypothesis that I must not have made any, and in support of my hypothesis I will point at the lack of tea on my desk. Yet once these relations between evidence and hypothesis are changed, it may become vicious. If someone were to question the evidence of the missing tea because they saw it behind my laptop, I cannot cite my hypothesis that I hadn't made any in support of the evidence. After all, my hypothesis relied on the evidence that is currently in question (Hohwy, 2016).

In the neural story, the evidence will come in the form of the input in the senses as caused by the extraneural world, and the hypotheses are formed by the brain. As the brain minimizes prediction errors, the amount of evidence it explains will increase. This thus means that the brain finds more evidence for its model. As such, the brain is thought of as self-evidencing: by minimizing prediction error, the brain "maximizes evidence for itself" (Hohwy, 2016, p. 264). The EE-circle is formed at the boundaries of the sensory system, where the evidence appears, and includes the brain that forms the hypotheses. This creates an *evidentiary boundary* between the states the brain has access to (its own internal states as well as the evidence it encounters in the senses), and the states the brain will need to infer. This "boundary is strict, with only inference being done by the mind, and being done only on the inside". The term evidentiary boundary is then justified as follows: "It is evidentiary because it is defined by the occurrence of the evidence, and it is a boundary because causes beyond it must be

[13] This opposed to views of direct perception, in which the body is supposed to be in direct contact with the world with no space for an evil demon to intervene. In Hohwy (2017) (endearingly titled *How to Entrain your Evil Demon*) he explicates the 'wriggling space' of the evil demon more precisely. Roughly, by emphasizing the importance of active inference in sampling the world selectively, he minimizes the role the evil demon could play. This does not however affect the relevant principled commitments.

inferred—they can only be represented vicariously" (Hohwy 2016, p. 264). In this sense, it is what separates the inferential brain from the *causes* of input: the extraneural world.

The sum of prediction error over time is also known as 'free energy' as it appears in the free energy principle (FEP) advocated by Friston (2013; Friston and Stephan 2007; Hohwy 2016). Minimization of prediction error in the long run is thus equivalent to the minimization of free energy. The idea is that, on an organismic level, the same process appears as in the brain. By minimizing prediction error between the organism and the environment in the long run, the organism finds more evidence for its own existence. Put differently, by increasing the fit between the organism and the environment, the organism remains within livable bounds (Friston 2012).[14]

The evidentiary boundary in organisms is a division between mind and world, according to Hohwy (2016). Behind the boundary lurks the mind that, with access only to the evidence presented to it, attempts to minimize its errors in predicting what could have caused these inputs. Explicitly put:

> The location of the evidentiary boundary determines the relata of the mind-world relation: what is on the 'mind'-side and what is on the 'world'-side … the boundary should determine what is part of the representing mind and what is part of the represented world. (Hohwy 2016, p. 268)

This means that the boundary is essential in distinguishing what is on the 'mind'-side, and what is on the 'world'-side. As such, an evidentiary boundary in itself may not be a sufficient condition for agency, it is a *mark* of agency. Where there is an evidentiary boundary, there lies an agent within.

3.2.3 On the Markov blanket formalism

Hohwy's evidentiary boundary is equivalent to what in the FEP literature is typically referred to as the *Markov blanket* (Hohwy 2016, p. 274; Friston 2013).[15] A Markov blanket is a term that stems from machine learning, and is often used in the FEP literature as a mark of life or agency, but can be applied to many systems beyond those (Ramstead et al. 2018; Ramstead et al. 2016; Kirchhoff et al. 2018). More precisely, a Markov blanket is an information-theoretic

[14] This showcases the appeal of PP. A single computational pattern is thought to be found at all levels of self-organization, from single cells, via the neural architecture to full-blown organisms, and other times even niche construction, sociality and human cultural practices (see Hesp et al. 2019 for an overview). It is important to note that Hohwy typically prefers to remain neurocentric (2016).

[15] Relevant to the concept of a Markov blanket are many statistical equations. As this paper aims to make a purely conceptual point, I will not include a mathematical description of the formalism. See Friston (2010, 2013) and Friston and Stephan (2007) for a mathematically informed introduction into the formalism and its use in FEP.

formalism that is applicable to any system that resists the second law of thermodynamics over time (Friston 2013; Constant et al. 2018). This means the system has to remain within a particular bound of self-maintaining states, whilst not visiting others. Of course, living systems will answer to the second law of thermodynamics eventually. In the human case, for example, that means we visit homeostatic states instead of those states in which our parts randomly disperse, until we inevitably die (after which of course we stop resisting the second law of thermodynamics). By contrast, the contents of an open bottle of perfume will spread throughout the area in a fairly random sense, and the entropy increases lawfully. Humans thus have a Markov blanket, whereas perfume does not (Friston 2012).

The Markov blanket is a statistical formalism with which one can mark the separation between the system and its environment. Within this formalism, we divide states into *internal states* which capture all states within the Markov blanket, and *external states* which capture all states outside of the Markov blanket. The states of the blanket itself mark the only points of contact between the two sides, comprising of *active states* that are influenced by internal states and influence external states and *sensory states* that are influenced by external states and influence internal states. Importantly, this means that the internal states and external states are statistically conditionally independent. This means that, given the occurrence of the active and sensory states, knowing the internal states does not give us more information on the external states and vice versa.

An advantage of this formalism is that it instantly identifies the divide between a system and its environment. If we are interested in the system, the Markov blanket clearly distinguishes the states of interest from those external to the system. Further, the internal states can only be affected by the external states *via* the sensory states, and the external states can only be affected by the internal states *via* the active states. The internal states, by minimizing free energy relative to external states and their interaction conditional on the blanket states, come to reflect the external states in some sense (Friston 2013). This reflection, in Hohwy's view realized in the form of representations, is most easily explained in phylogenetic adaptation of an organism to its environment.[16] Think of the way a fish's phenotype reflects its environment: the presence of gills can be seen as a reflection of the fish having evolved in water. But ontogenetically the same occurs behaviourally. Our ways of acting increasingly reflect our surroundings, especially in designer environments. Think of the way in which the choice of vehicle and the infrastructure

[16] Whether or not to cash this out in terms of structural representations or not is currently a hot debate (Kirchhoff and Robertson 2018; Bruineberg et al. 2016; Gładziejewski 2016; Williams 2018) . I shall argue that, to avoid the objections raised in this paper, we will have to rid ourselves of representations.

shape the way you move (Clark 2016, ch. 9). The particular behaviors of the organism thus increasingly reflect their environment as to increase its fit (Friston 2013, Kirchhoff 2018b; Bruineberg et al. 2016). In this particular sense, the internal states of the system are sometimes said to represent the external environment (Friston 2013; Hohwy 2016). Interestingly, the Markov blanket formalism is not only applicable to living systems, such as single cells, organs and organisms, but tentatively is thought to apply to social situations such as two people in conversation, or even communities and sociocultural practices (Ramstead et al. 2016).

The translation of the Markov blanket formalism to Hohwy's evidentiary boundary is straightforward. A Markov blanket's division between internal and external states map onto neural and extraneural states respectively, with active and sensory states mapping onto action and perception respectively. The brain only has access to the internal, sensory and active states, but not the external states. The blanket is thus interpreted to be a sort of veil through which the brain is in contact with the world. Note that strictly speaking the Markov blanket is merely a statistical tool to divide sets of states and show statistical relations between the particular groups. In Hohwy's view, this division has received an epistemological interpretation in the form of an evidentiary boundary that shares the structure of the Markov blanket. I will come back to this in Section 4, when proposing the solution to the objection raised below.

3.3 Hierarchies, Markov blankets and inferentialism

At face value, PP seems like a theory with an impressively wide explanatory reach and a promising empirical backing (Hohwy 2017, p. 260). It describes elegantly what our brain does to cope with the external world, explaining a wide variety of cognitive tasks under a single banner. By Hohwy's lights, the Markov blanket's separation between internal and external states with the blanket itself as points of contact gains an epistemological flavour. This means that there are two ways in which the internal states represent external states. First, for an external observer, the internal states can be said to represent external states in the way that we can make inferences about the external states on the basis of our knowledge of the internal states (think about the fish's gills). Second, for the organism itself the model is reified so that the organism's internal states represent the external states as a tool for the organism to navigate the world. Here the statistical formalism of the Markov blanket thus entails an epistemological reading of the self-evidentiary boundary that implies internalism, solipsism, Cartesian skepticism, and of particular importance here: inferentialism and representationalism (Hohwy 2016, p. 259). When one unpacks this view, however, not everything adds up. To see this, one

only needs to apply the inferentialist implication of the evidentiary boundaries to PP's hierarchical layers. This evokes a worry that may be insurmountable. An internal inconsistency emerges that is not easily solved.

Before unpacking the implications of the view, it is helpful to explicate some of the features of Hohwy's take on PP. Whenever a system forms hypotheses to best explain or predict the occurrence of a piece of evidence, after which the evidence will count as support for the hypothesis, an evidentiary boundary is formed. Only states external to an evidentiary boundary afford prediction and modeling: the system has access to its internal and blanket states so that they need not be represented. This means that the presence of a prediction implies that the states being predicted are external to the states doing the predicting, with an evidentiary boundary between the two. This evidentiary boundary is grounded in a Markov blanket, and is interpreted by Hohwy (2016) to imply inferentialism and representationalism. Whenever there is a prediction, there is thus also an epistemological evidentiary boundary that is equated with a Markov blanket. Recall finally that PP's prediction mechanism is hierarchically layered so that different layers are attuned to evidence at different (time)scales (Hohwy 2013; 2016).

3.3.1 Hierarchical layering and evidentiary boundaries

In Hohwy's PP, each hierarchical layer has its own evidentiary boundary. We can see this in the following. Each layer predicts the activity of the layer below it (Anderson 2017, p. 5; Kirchhoff et al. 2018, p. 3). Hohwy says, for example:

> "Each level is only concerned with attenuating as best possible the input (i.e., explaining away the prediction error) from the level below and passing any unpredicted parts of this input up to the next level for it to explain away" (Hohwy, p. 262).

So each level is concerned with the minimization of prediction error, divided into two (sub-)activities: 1) it attenuates the best possible input from the level below, and 2) it passes any unpredicted parts of this input to the next level. Differently put, each layer is only concerned with predicting as well as possible, or inferring the best explanation for (the inverse of minimizing prediction error, colloquially put) the activity of the layer below.

For this to be the case, the input at the boundary of the layer will count as evidence. In Markov blanket terminology, the states from the upper layer are the internal states, the states from the layer below are the external states with the upper layer's boundary as points of contact where the evidence appears. This boundary is strict, and an object "outside the boundary (...) is represented by the inner states to the extent the system can infer its properties" (Hohwy 2016,

p. 269). From the perspective of the upper layer, then, the activity of the layer below is an external state that needs to be modeled, and represented, its properties only inferred. The central idea is that the neural system forms an EE-circle, bounded by an evidentiary boundary, a Markov blanket, that demarcates what is internal and accessible and what is external and to be inferred and represented. The story above means that within the brain, each layer has its own evidentiary boundary as well, for which states within the neural system but positioned at outermore layers are also external.

Hohwy concedes that his own view has this implication:

> if we focus on the entire hierarchy minus the bottom layer, we would still have a fully formed EE-circle with its own evidentiary boundary, modeling the external world, including the states of the bottom layer. The input to this system then becomes the evidence for a shrinked model, and thus for the existence of a new "agent". (Hohwy 2016, p. 273)

Indeed, not only the outermost ends of the neural system has one, but each layer has its own Markov blanket, its own evidentiary boundary. The entire neural hierarchy is thus viewed as having its own Markov blanket, but the entire system *minus the bottom layer* also has its own Markov blanket, and that second system *minus what is then the bottom layer* has its own Markov blanket and so on until we reach the minimally shrunk agent (Hohwy 2016, p. 274). This gives us a picture of the neural hierarchy as shown in Figure 1.[17]

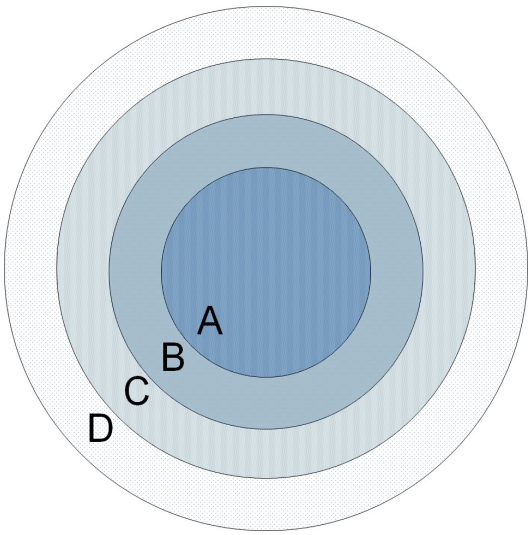

Figure 1 The view of the hierarchical layering of the brain as described by Hohwy (2016). In this view, A is the minimally shrunk system, and B, C, and D each form a

[17] This picture is close to what is being proposed in the FEP literature, where the world is described as consisting of Markov blankets of Markov blankets (Ramstead et al. 2016; Ramstead et al. 2018; Kirchhoff et al. 2018). We will delve more into this view below, and how it relates to Hohwy's view. We will see that a few important theoretical posits on Hohwy's side make the difference.

> further outer layer so that ABCD encompasses the entire neural hierarchy. The number of layers is chosen for didactical purposes and need not reflect the brain per se.

Hohwy admits such a view to be unappealing. It implies that parts will be

> beyond one evidentiary boundary, and within a further evidentiary boundary … It requires that we posit two overlapping yet intimately linked EE-circles with different evidentiary boundaries. If there are two EE-circles, then the input to each of the circles will be evidence for the existence of two distinct yet overlapping agents. This … [comes] at the unattractive cost of proliferating the number of agents centered on a particular organism.[18] (Hohwy 2016, p. 270)

An initial issue with this view is thus that, if each hierarchical layer implies the existence of an agent, there are many overlapping but distinct agents within one mind. Hohwy suggests two potential solutions to this issue. One option is to accept the multiplicity of agents as a given, and instead let our current inquiry decide which agent should be the focus of our investigation. Which of all circles is the *relevant* EE-circle is then dependent on our own interests. An issue with this, he states, is that it introduces ambiguity in the attribution of properties to these agents. Another solution he suggests,

> is to rank agents according to their overall, long-term prediction error minimization (or free-energy minimization): the agent worthy of explanatory focus is the system that in the long run is best at revisiting a limited (but not too small) set of states. (Hohwy 2016, p. 270)

This would mean that, again, we accept the proliferation of agents. However, rather than letting our current investigation mark the relevant agent, he suggests a ranking will show which agent is the only agent worthy of explanatory focus. In line with his neurocentric approach, he conjectures that

> such a minimal entropy system is constituted by the nervous system of what we normally identify as a biological organism: shrinked agents are not able to actively visit enough states, and extended agents do not maintain low entropy in the long run (cf. Friston 2013 for an argument along these lines, where EE-circles and self-evidencing are cast in terms of Markov blankets). (Hohwy 2016, p. 270)

Hohwy's neurocentric view thus implies a multiplicity of agents, each with their own evidentiary boundary, regardless of which solution is chosen. The issue he points out is that this proliferation obfuscates which agent of these is to be relevant. What I shall argue below is that this hierarchical layering of agents runs into a deeper problem still.

[18] The original context of this quote is Hohwy's discussion of the possibility of extended cognition in PP, in which the proliferation of agents is brought up as one of a few reasons for the notion of extended cognition to be unappealing under PP. Further on he remarks that this objection also holds for a neurocentric approach, after which he discusses some possible solutions I discuss in text below.

3.3.2 The internal inconsistency

In this section I will unpack an implication of Hohwy's particular interpretation of PP, which sees Markov blankets not merely as statistical formalisms, but as denoting particular boundaries with epistemological implications. The Markov blanket or the evidentiary boundary that divides the external states from the internal states implies seclusionism, inferentialism and representationalism, according to Hohwy (2016). We can only know of the world by means of a statistical inference to the best explanation with a representational model. Recall that Hohwy subscribes to the brain's hierarchical layering constituting evidentiary boundaries at each level, as we have seen above in Section 3.1 and Figure 1. This means that ABCD marks the full neural system, for which the extraneural world counts as the external states. ABC is smaller system for which D plus the extraneural world count as external states, AB is a system for which CD plus the extraneural world count as external states and finally, A is the minimally shrunk agent for which BCD plus the extraneural world count as external states. Each cortical layer can thus be viewed as having its own Markov blanket, and evokes inference and representation requirements in Hohwy's view. This means that agent A can only by way of inference and representation know of the states of B. AB is veiled off from C in the same way, and so forth. At each and every boundary seen in Figure 1, representation and inference are required. This on itself is a puzzling state of affairs. A single neural hierarchy is at multiple levels secluded from itself, and thus at multiple levels does not have access to its own states. This may be unappealing, but there is more to it.

From the perspective of AB, this is a peculiar situation. After all, the evidentiary boundary implies both that there is a representation and inference requirement for the states external to AB, which are CD plus the extraneural world, and the states internal to the Markov blanket are *not* to be represented or inferred. Both the states of A and the states of B are internal to AB, so for AB, there is no need for representation (or inference) for both the states of A and the states of B. For A however, there *is* need to represent the states of B, as it is external to its own evidentiary boundary. The states of B are thus both external to the states of A when A is viewed independently and thus requiring representation, and the states of B are cooperatively internal with the states of A as part of AB, thus *not* requiring representation by the states of A. Thus for A, depending on whether it is viewed independently or cooperatively as part of AB, the states of B need to be represented or not. The 'strict' boundary between A and B both does and does not exist at any one moment, seemingly dependent on our point of focus.

This is an issue that is a consequence of a combination of the proliferation of evidentiary boundaries and inferentialism as it occurs in Hohwy's take on PP. It shows an inconsistency. The boundaries cannot both be strict so that which lies beyond requires modeling and representation for the states internal to them, and appear and disappear depending on which system we are looking at. The states of A cannot both be required to have access to the states of B by way of modeling and representation and not have this access simultaneously. The states of B cannot both be (cooperatively) internal and external to the states of A.

There are three possible replies to this objection that I anticipate: perspectival relativism, shifting the the representation requirement, and adherence to the maths. Perspectival relativism is what it says on the tin: the relevant Markov blanket would be dependent on our perspective, or our current explanatory focus. The blanket then shifts from place to place depending on where we look. This is essentially one of the two solutions Hohwy provided for the proliferation of agents I discussed in Section 3.1. The implication of this is that it makes of the evidentiary boundary's entourage an affair dependent on our perspective as well. By shifting the boundary about, we change what counts as 'the world', as opposed to the internal states for the agent. With our own gaze, we would be able to change what for a particular agent requires modeling, representation and inference, and what does not. We would also be able to shrink or enlarge the agent at will, when including or excluding particular cortical layers or not. This solution does not square with Hohwy's view of PP's hierarchical inference to the best explanation that necessitates evidentiary boundaries at each layer. As such, for this reply to work, the entailments of the evidentiary boundary need reconsideration. It cannot be the case that the fundamental epistemic relation any system has to the world it is engaging with depends on the whim of our gaze.

The second possible reply is to shift the representation requirement. The idea would be that not just any Markov blanket denotes a representation requirement, but only, say, the outermost blanket, or, perhaps in line with Hohwy's solution to the proliferation of agents, the system that is best at minimizing prediction error in the long run. The issue with the outermost blanket option is that the Markov blanket formalism in FEP is now also applied to social cognition and human societies, as well as niche construction (Ramstead et al. 2016; Ramstead et al. 2018; Constant et al 2018; see also Hesp et al. 2019). 'Outermost' is not a clear notion anymore when we start applying Markov blankets at a social level. Ranking the systems will not be easy either, as it may not be so clear how to measure the prediction error minimization done by a society at large. Even if we do operationalize this, however, it is only a contingent fact that that particular Markov blanket wraps around a system that, in its current configuration

is the one that is best at minimizing prediction errors in the long run. Things can change at any moment when systems interact and social Markov blankets are not nearly as temporally stable as organismic ones (Hesp et al. 2019). Without a principled reason for one Markov blanket to imply a representation requirement and others not, this is merely a just-so solution. Furthermore, it makes the representation requirement seem tacked on. It would entail that plenty of evidentiary boundaries exist (as they do intra-neurally, in Hohwy's view), that do not demarcate the need for internal representation of external states. Only a few evidentiary boundaries would then be 'privileged' with a representation requirement, begging the question as to why.

The third possible reply I dubbed 'adherence to the maths'. The rationale behind this would be that it follows from the Markov blanket formalism, and, whether we like it or not, is what we need to accept. If the math has come out correctly, then this is what we need to accept, the thought may be. Yet consider this. The Markov blanket is a statistical formalism that separates internal from external states, and marks the blanket as points of contact. Internal states are conditionally independent from external states in a statistical sense. Further, due to the conditional independencies of the internal states and the external states in combination with the minimization of free energy, the internal states can be said to represent the external states. Due to this, we can make inferences about the external states if we know the internal states (Friston 2013). According to this formulation, the story would go, it is just a *matter of maths* that the internal states represent the external states. The states of A represent the states of B, and the states of A and B together as AB represent the states of C. If this is what the formalism tells us, then this is simply what we need to accept.

This approach may seem appealing. Note however that the statistical formalism does not actually denote epistemological relations. If a Markov blanket is applicable, this means that the internal states are conditionally independent statistically, the conditionality arising from the active and sensory states in a statistical sense. In the statistical formulation, representation is a representation from an external observer's, a scientist's, perspective. From that perspective, the internal states of A can be said to represent the states of B, and external observers can exploit this relation to make an inference about the states of B with their knowledge of the states of A. From this perspective, it is not problematic that A independently represents B, but when seen cooperatively with B, only the states of C (and beyond) are represented. If the states of AB count as internal, then the inferences external observers would like to make simply are about what is external to that system. This is not an epistemological claim about the system under scrutiny, but rather a pragmatic claim about possible methods of inquiry.

The use of representation in Hohwy (2016) concerns a representation requirement at work in the system's own cognitive workings. These are entirely different notions and it is not evident that what can be used as a representation by an external observer for a particular target also represents that target itself, to be used as a tool to navigate the target. It could in principle be that the system does represent that target.[19] Yet, that a particular system, a set of states, can be said to represent another set of states from an external observer's point of view does not imply that, from the system's perspective, those same states are represented. Put differently, the scientifically devised statistical model that describes a target system is not evidently what the system itself uses and manipulates to navigate the world. Though a proper explication of the relation between these two uses of representation and modeling in the FEP framework is outside the scope of this paper, it suffices that implications about representations from an external observer's point of view cannot without additional justification be employed for implications about representations from the system's point of view. In sum, reference to the mathematical structures underlying Hohwy's view is no way out. First, the mathematics are actually indifferent to epistemological readings such as the one Hohwy (2016) suggests. Second, there is a conflation of two notions of representation: a representation for an external observer and a mental representation to be used and manipulated by the organism. In the latter interpretation, Hohwy's preferred view, the inconsistency remains and is not solved.

PP has a wide explanatory reach and seems to have promising empirical backing, yet is internally inconsistent in Hohwy's (2016) reading. This is a problem that needs solving. I suggest that there are three features of PP that, conjoinedly, cause this inconsistency: hierarchical layering, a representation requirement, and Hohwy's view of inference in PP. In the section below I will suggest a possible solution to solve this inconsistency.

3.4 Towards an alternative approach to predictive processing

Above I have described an inconsistency in Hohwy's (2016) portrayal of PP. This appears due to the combination of hierarchical layering in PP and Hohwy's epistemological commitments to representationalism and his particular take on inferentialism. In this section, I will explore an alternative approach to PP that avoids these issues. I shall take hierarchical layering as essential to the PP framework, which means we will need a view that eschews Hohwy's form of inferentialism and representationalism. We shall see below that FEP can be construed non-

[19] I would argue that the system cannot and need not represent the target, but full exposition of the argumentation is outside of the scope of this paper.

representationally. Non-inferentialism is less obviously attained. There is, however, a notion of inference in the FEP literature that avoids this issue (Bruineberg and Rietveld 2014; Kirchhoff and Robertson 2018).[20]

Hohwy (2016) has argued however that representationalism is part and parcel of the PP framework. In particular, he argues that the evidentiary boundary, the Markov blanket, is what necessitates this perspective. Troubled though this view may be, this adherence needs to be deconstrued before we can engage with possible alternatives. We need to first separate the baby from the bathwater (at least conceptually), before we can go about carefully throwing out the bathwater, and keeping the baby. In broad strokes, Hohwy's claim has been challenged before (Anderson 2017; Fabry 2017; Bruineberg et al. 2016). For our current purposes, however, it may be fruitful to re-engage with the Markov blanket formalism, understand and reject why Hohwy thinks this necessitates seclusionism, but more importantly, what the formalism *does* tell us about the system under scrutiny, and what we can learn from that.

It has been emphasized throughout the paper, but it is important to remember that the Markov blanket is a formalism, a mathematical description of, roughly, (semi-)stable systems (Friston 2010, p. 1). The formalism is a descriptive tool, and by itself, cannot place constraints on the system under scrutiny. It is still a matter of the system being described whether it is one way or another, and the way the target system is places constraints on the formalism; not the other way around. To argue by analogy, our description of a visual scene is constrained in some sense by the visual scene itself: describing a summer scene as winter-like (whatever these terms may mean) means you are not describing the target scene. Our description of the visual scene "it is winter-like" does not place constraints on the target visual scene, as much as we may wish otherwise. In much the same way the Markov blanket is a description of the system, and if it ascribes particular structures to a system that it does not have, it fails at describing the system properly, as opposed to placing ontological or epistemological constraints on what the system can or cannot be like.

3.4.1 Representationalism and the free energy principle

A key feature of an abstract model is that it may lay bare features and aspects of the target system that may otherwise be difficult to pick apart. It seems indeed that it is not, then, the Markov blanket itself that places the constraint, but simply the features of the system that it

[20] Cartesian skepticism and seclusionism most likely stand or fall together with inferentialism and representationalism, but full exposition of these ideas is outside of the scope of this paper.

illuminates. The idea is, as we have seen above, that because of the conditional independencies the Markov blanket describes and the minimization of free energy, the internal states can be said to represent the external states so that inferences can be made about the external states with knowledge of the internal states. In Hohwy's view, this is taken to mean that the system under scrutiny itself represents the external states and needs to make such inferences. This conflates two uses of the term representation as described above, and does not justify a representationalist position or inferentialist position in itself.[21] This is exemplified when we consider that the blanket formalism is applicable to non-living systems, such as a pebble. It is one thing to claim that one can use the internal states to make inferences about the external states, but it is another thing entirely to state that the the pebble makes inferences about the external states on the basis of internal representations. We should be wary of reification, of turning the scientific model we use to describe life into a tool that the organism uses to navigate the world.

It may be fruitful now to look at what the Markov blanket *does* tell us about the target system. At the very least, the internal states are conditionally independent from the external states in a statistical sense as we have seen in Section 2.3. For the relatively stable internal states to come to reflect an intrinsically dynamic and uncertain external environment, there is a sense in which the internal and external states need to covary. In other words, the internal states need to change as the environment changes, the system needs to adapt. This covariational sense of embodied adaptation seems to be exactly what is at work when Friston talks about the organism *being* a model of its environment (Friston 2013; Bruineberg and Rietveld 2014; Kirchhoff and Robertson 2018, p. 4; see also Bruineberg et al. 2016). Standardly, this covariational relation is thought to be represented internally in the generative model under the name of structural similarity (Gładziejewski 2016). Indeed, covariance can feature as a central piece between a representational artifact and its target. Yet covariance in itself is insufficient to ground representations naturalistically (Hutto and Myin 2013; 2017). Minimally, representations have some sort of content, which at least means there are some sort of correctness conditions: the representation represents its target in such a way that it may not be so (Travis 2004). However, correctness conditions are not applicable in many forms of

[21] There may be other good reasons to prefer representationalism or Hohwy's take on inferentialism (I would argue there aren't, but that is outside of the scope of this paper). Here I restrict myself to what follows immediately out of the Markov blanket formalism and show that the formalism itself does not *entail* such an interpretation of the target system.

covariation. Consider the covariation between the shape of my hand and the shape of a door handle: in no sense does the shape of my hand *represent* the door handle due to this covariation.

A non-representational approach to the PP framework is being developed, and it shows promise regarding the direction we need to take if we are to avoid the issue discussed in the current paper (Bruineberg and Rietveld 2014; Bruineberg et al. 2016; Kirchhoff 2018a 2018b 2018c; Kirchhoff and Robertson 2018; Kirchhoff et al. 2018; Ramstead et al. 2019; Kirchhoff and Kiverstein 2019). A key feature of this approach is that it does away with representations in PP explanations. As we have seen in Section 2, however, representational models reside at the very core of PP. Kirchhoff and Robertson (2018) have shown that the generative model is representational in name only.[22]

To see how, we need to briefly discuss Kiefer and Hohwy's (2017) construal of *misrepresentation* in PP. Representations are marked by their ability to represent a target system in such a way that it may not be so. This means that *misrepresentation* is an essential feature to account for in a representational system. Kiefer and Hohwy argue that the Kullback-Leibler divergence is a measure of misrepresentation in FEP models (2017; see also Friston, 2013). They argue, roughly, that it is a measure of the difference between the brain's *prior* probability distribution (the brain's distribution of hypotheses *before* encountering the current situation) and the *actual* posterior probability distribution (the actual state of affairs in the world). This divergence between the brain's estimate and the actual state of the world is thought to measure misrepresentation. However, we can see that all it measures is the difference between two separate probability distributions. Only if we *assume* that the brain's model is a representation of the world's state of affairs, can we call the Kullback-Leibler divergence a measure of mis*representation*. As it stands, Kirchhoff and Robertson argue, all it measures is negative Shannon-informational covariance (2018). This is merely the Shannon-informational difference between the two probability distributions: the extent to which they *do not* covary.

As such, what was assumed to be the representational relation between the model and the environment bottoms out in Shannon-informational covariance. Yet mere covariance is insufficient to ground representations. Kirchhoff and Robertson (2018) thus show that PP describes a non-representational cognitive system that reliably covaries with its environment, rather than representing it, *despite appearances* (Kirchhoff and Robertson 2018; Hutto and Myin 2017).

[22] Technically, one may choose to invoke representations at higher hierarchical levels, yet this is not necessary. Moreover, the inconsistency may well reappear as long as there are still multiple hierarchical levels at which representation is required.

To ensure the solution is not worse than the problem it intended to solve, we should review whether covariation runs into the inconsistency. The crucial aspects of the representation requirement that cause the inconsistency are both the requirement of representation of anything external to the system, and the requirement of no representation of anything internal to the system. This asymmetrism is incompatible with a hierarchically layered view of the mind consisting of overlapping, but distinct agents as found in Hohwy (2016). Covariation is a relation that does not place the same constraints on a system. Covariance is a symmetrical relation. If the dynamics of system A covary with the dynamics of system B, then the dynamics of system B also covary with the dynamics of system A. A standard example of covariance is the relation between the rings in the trunk of a tree and the years the tree has lived. These two systems covary: as the tree lives for more subsequent years, more rings are added. It is thus perfectly consistent to say that the states of AB covary with the states of C, and simultaneously that the states of A covary with the states of B. If representation bottoms out in covariance, then there is no inconsistency anymore.

With this, we have tackled one aspect of the issue, yet if we don't also tackle inferentialism as Hohwy portrays it, the inconsistency will remain. Indeed, A will still both need to infer the states of B and not need to infer the states of B depending on whether it is taken by itself or as part of AB. And as we have seen in Section 3.1, A both has its own evidentiary boundary *and* is within a further evidentiary boundary (as AB) of the same neural system.

3.4.2 Not the inference you know

The PP framework may thus be rid of representations, but in Kirchhoff's view, too, the proposal seems ridden with inferentialist terminology. The goal is to find an approach to PP that does not fall victim to the same inconsistency. After all, a view of inference such as is used in Hohwy (2016) will, even without representations, remain inconsistent. The states of A cannot both need to infer and not need to infer the states of B. Yet what exactly could inference be without any sort of content that it is supposed to infer? A standard notion in which a conclusion is formed following some argumentative steps might be difficult without contents of the steps as well as the conclusion. What exactly could inference be, then, if it is not contentful? Kirchhoff et al. (2018) may offer some insight:

> Active inference, in its simplest formulation, describes the tendency of random dynamical systems to minimize (on average) their free energy, where free energy is an upper bound on (negative) marginal likelihood or evidence (i.e. the probability of finding the system in a particular state, given the system in

> question). This implies that the kind of self-organization of Markov blankets we consider results in processes that work entirely to optimize evidence, namely self-evidencing dynamics underlying the autonomous organization of life, as we know it. (p. 2)

Active inference is thus cashed out naturalistically in the ability to optimize "self-evidencing dynamics underlying the autonomous organization of life, as we know it". The dynamics underlying the autonomous organization of life in themselves say little about our epistemic position relative to the world. Consider also the following passage from Kirchhoff (2018b), where he, in part quoting Bruineberg and Rietveld (2014), advocates

> a 'minimal' notion of prediction (and also inference) by which it is coherent to say that for any two dynamical systems, A and B, it is coherent to say that A 'models (or 'infers') the ''hidden causes of its input (the dynamics of B) when it reliably covaries with the dynamics of B and it is robust to the noise inherent in the coupling.'' (Bruineberg and Rietveld 2014, p. 7)" (Kirchhoff 2018b, p. 762; see also Ramstead et al. 2019)

According to this 'minimal' notion, then, for any one system to infer or model the dynamics of another system, is simply to say that it covaries reliably with the dynamics of that other system. Inference, taken in this way, just as representation, bottoms out in the same symmetrical covariational relation we discussed above. To the extent relevant here, this deflated notion of inference thus does not encounter the inconsistency.

This is consistent with the statistical notion of the Markov blanket and an adherence to free energy minimization. The Markov blanket as we take it here is a statistical separation of the states under scrutiny. It divides the system into internal and external states, with the blanket states as points of contact. The internal and external states are then conditionally independent. If we take the internal states to minimize their free energy in relation to the external states, we will see that the internal states will come to covary with the external states. When the states of A then, are said to statistically 'infer' the states of B, this entails that they are statistically conditionally independent and that they covary.

This notion of statistical inference as denoting statistical conditional independence and covariation by way of free energy minimization gets us out of the inconsistency. It should be noted that this is a rather technical definition, and does not stroke with the general notion of inference, nor the notion of model that it is accompanied by. Though there is no clear consensus on what models are like, nearly all theorists consider them to be representational (Frigg and

Hartman 2018).[23] As such, for FEP to employ strictly non-representational, covariational models is unusual at least.

3.4.3 Scientific models and engaging organisms

In this section, I have suggested a non-representational, covariational approach to PP as a solution to the inconsistency. The notion of inference has seen a strong transformation relative to the standard notion of reaching a conclusion through (argumentative) steps. With this, the inconsistency no longer arises. In order to create clarity on the ontological status of the statistical machinery, however, and thus to avoid easily reinstating the inconsistency, I think it is important to elaborate on the distinction between the types of representation discussed above. The view I suggest is a strictly instrumentalist view of models and inferences in PP.

With this, I intend to be explicit about what is a scientific tool for description, and what is ascribed to the organism: what we think is accessible to the organism (both on a subpersonal and personal level). My view is captured rather well by Ramstead et al. (2019)'s *enactive inference* approach:

> The generative model is *enacted*; in the sense that adaptive behaviour brings forth the conditional dependences [sic] captured by the generative model, that is, keeping the organism within its phenotypic, characteristic states. (Ramstead et al. 2019, emphases in original)

In this sense, I would argue that the organism behaves adaptively in its environment in an ongoing attempt to increase its fit. In doing so, particular statistical relations between the organism and the environment, bringing forth the conditional (in)dependencies we have spoken of before. These particular relations between the organism and the environment can be captured in a statistical generative model by scientists that have been trained in our modeling practices. The model is thus a scientific construct that is inaccessible to the organism. The organism can thus not 'leverage' or 'use' features of the model to make a statistical inference in a usual sense of the word. The organism *can* however, exploit certain covariational relations between itself and the environment in the same way that we daily exploit such relations when we shape our hand to covary with the shape of the door handle, and in many other situations. It is important to note that this may deviate from the position advocated by Ramstead et al. (2019), who remain rather ambiguous about the ontological status of the model or its probability densities.

[23] De Oliveria (2018) has argued for a non-representational approach to models in science. This notion of modeling explains the origin and use of modeling in terms of its surrogacy for the target system and our scientific practices of modeling.

Instead, speaking with Gallagher (2017), I would suggest that "active inference is not 'inference' at all, it's a doing, an enactive adjustment, a worldly engagement (Bruineberg, et al. 2016; Gallagher and Allen 2018)". This 'enactive adjustment' is then no more than a tighter covariational relation between the organism and its environmental niche. In the view I suggest best helps us overcome the inconsistency, the statistical models we devise remain statistical models to be made, updated and manipulated by scientists, and do not become the tools our target systems use to navigate the world.

3.5 Conclusion

Hohwy argued that accepting PP means accepting inferentialism in terms of inference to the best explanation and representationalism. This is not only problematic in itself, as it may offer a misguided view of our epistemic position in the world (Anderson 2017), but it also invites an inconsistency. As seen above, the states of A need to infer the states of B when A is viewed as a separate system, yet, when cooperatively part of the system AB, the states of A need not infer the states of B. This is no reason to give up on PP however, as there is hope for an alternative approach to PP that can solve this issue. In this view, both representation and inference are cashed out in terms of reliable covariance. Explicitly, however, this has been taken to only rid them of representations, and allows them to keep inferences. However, these sorts of inferences are not the inferences you may know. Inferences as they appear here are cashed out in terms of covariance and denote particular statistical relations between internal and external states as divided by a Markov blanket. As covariance is a symmetrical relation, there is no inconsistency in the above situation. It is clear that, if inferences are thus construed, they avoid the inconsistency. I argue that if we take an instrumentalist stance on the statistical machinery of FEP, we gain in ontological clarity. This entails that models in FEP are scientific constructs, tools we use to *de*scribe systems, but do not *a*scribe them to the systems under scrutiny.

References

Anderson, M. (2017) Of Bayes and bullets : an embodied, situated, targeting-based account of predictive processing in Metzinger, T. and Wiese, W. (eds.). *Philosophy and predictive processing*. Frankfurt am Main : MIND Group.

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

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

Clark, A. (2013) Whatever Next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences 36*, 181–253

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

Constant, A., Ramstead, M. J., Veissière, S. P., Campbell, J. O., & Friston, K. (2018). A variational approach to niche construction. *Journal of the Royal Society Interface, 15*, 141.

De Oliveira, G. (2018) Representationalism is a dead end. *Synthese*, 1-27

Dennett, D. C. (2013). *Intuition pumps and other tools for thinking*. W.W. Norton & Company.

Engel, A. K., Fries, P. & Singer, W. (2001). Dynamic predictions: Oscillations and synchrony in top-down processing. *Nat Rev Neurosci 2*(10), 704–716.

Fabry, R. (2017) Transcending the evidentiary boundary: Prediction error minimization, embodied interaction, and explanatory pluralism, *Philosophical Psychology, 30*(4), 395-414,

Frigg, R. and Hartmann, S., (2018) Models in Science, in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy (Summer 2018 Edition)*, URL = <https://plato.stanford.edu/archives/sum2018/entries/models-science/>.

Friston, K. (2010). The Free-Energy Principle: A Unified Brain Theory? *Nat Rev Neurosci. 11*(2):127-38

Friston, K. (2012) A free energy principle for biological systems. *Entropy 14*, 2100–2121.

Friston, K. (2013). Life as we know it. *Journal of the Royal Society, Interface, 10*, 20130475.

Friston, K., J. Daunizeau, J. Kilner and S. Kiebel (2010). Action and behavior: A free–energy formulation. *Biological Cybernetics 102*(3): 227–260.

Friston, K., S. Samothrakis and R. Montague (2012). Active inference and agency: Optimal control without cost functions. *Biological Cybernetics 106*(8): 523–541.

Friston, K. and Siebel, S. (2009). Predictive Coding under the Free-Energy Principle. *Phil. Trans. R. Soc. B 364*, 1211-1221.

Gallagher, S. (2017) *Enactivist Interventions: Rethinking the Mind*. Oxford University Press

Gallagher, S. and Allen, M. (2018) Active inference, enactivism and the hermeneutics of social cognition. *Synthese 195*(6), 2627-2648

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese, 193*(2), 559–582.

Hesp, C., Ramstead, M. J., Constant, A., Badcock, P. B., Kirchhoff, M. D., & Friston, K. J. (2019). A multi-scale view of the emergent complexity of life: A free-energy proposal. In M. P. e. al. (Ed.), Evolution, Development, and Complexity: Multiscale Models in Complex Adaptive Systems: Springer.

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

Hohwy, J. (2016) The Self-evidencing brain. *Noûs, 50*(2), 259-285

Hohwy, J. (2017) How to Entrain Your Evil Demon In: Metzinger, T, and Wiese W., (eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Hutto, D. D. and Myin, E. (2013) *Radicalizing Enactivism: Basic Minds Without Content*. MIT Press

Hutto, D. D. and Myin, E. (2017) *Evolving enactivism: Basic minds meet content*. MIT Press

Kirchhoff (2018a) Autopoiesis, Free Energy And The Life Mind Continuity Thesis. Synthese 195, 2519–2540

Kirchhoff (2018b) - Predictive processing, perceiving and imagining: Is to perceive to imagine, or something close to it. Philosophical Studies 175:751–767

Kirchhoff (2018c) - The Body in Action: Predictive Processing and the Embodiment Thesis in Newen, A., De Bruin, L., and Gallagher, S. *Oxford Handbook of 4E Cognition: Embodied, Extended and Enactive*. Oxford: Oxford University Press.

Kirchhoff M, Parr T, Palacios E, Friston K, Kiverstein J. (2018) The Markov blankets of life: autonomy, active inference and the free energy principle. *J. R. Soc. Interface 15*, 20170792.

Kirchhoff, M. and Robertson, I. (2018) Enactivism and predictive processing: a non-representational view, *Philosophical Explorations 21*:2, 264-281

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information.* Freeman.

Ramstead, M. Badcock, P., Friston, K. (2018) Answering Schrödinger'squestion- A free-energy formulation. *Physics of Life Reviews 24*, 1–16

Ramstead, M. Kirchhoff, M. Friston, K. (2019) A tale of two densities: active inference is enactive inference. *Adaptive behavior*, 1-15

Ramstead, M., Veissière, S., and Kirmayer, L. (2016) Cultural Affordances-Scaffolding Local Worlds Through Shared Intentionality and Regimes of Attention. *Frontiers in Psychology 7* (1090)

Rao and Ballard, (1999) Predictive Coding in the Visual Cortex: a Functional Interpretation of Some Extra-classical Receptive-field Effects. Nature Neuroscience, 2(1):79-87

Seth, A. (2013) Interoceptive inference, emotion and the embodied self. *Trends Cogn Sci. (11)*:565-73

Travis, C. 2004. The silence of the senses. *Mind 113* (449): 57–94.

Wiese, W. and Metzinger, T. (2017) Vanilla PP for Philosophers: A Primer on Predictive Processing in Metzinger, T. and Wiese, W. (eds.). *Philosophy and predictive processing*. Frankfurt am Main : MIND Group.

Williams, D. (2018) Predictive Processing and the Representation Wars. *Minds & Machines 28*: 141.

4 The Embedded View, its critics, and a radically non-representational solution[24]

Author

Thomas van Es 1

1 Centre for Philosophical Psychology, Department of Philosophy, Universiteit Antwerpen, Belgium

Abstract

Whether perception involves the manipulation of representations is currently heavily debated. The Embedded View (EV) advanced by Nico Orlandi seeks a middle passage between representationalism and radical enactivism. In this paper I argue for a non-representational take on EV. I argue that this is the best way to resolve the objections EV has received from both representationalists and non-representationalists. I analyze this debate, and distinguish four sorts of objections: 1) the objection of the wrongfully cut middleman, 2) the argument against explanatory exclusionism, 3) the case for scientific benefits of representations, and 4) the charge of inconsistent ascription of representational status in EV. I argue that (1) the middleman was never cut in EV, and is controversial to boot, (2) *otherwise equal*, non-representational explanations have primacy over representational explanations, due to the lack of naturalistic grounds for representations and the unnecessarily ascribed cognitive load to the system. Further, I show that (3) puts the cart before the horse, and the arguments on offer are viciously circular. However, the final objection, (4) lays bare a deeper issue for EV. At the cost of giving up the middle position, however, the explanatory tools already available to EV can be shown to cover the work initially thought to require representation. I conclude that EV is best altered to be a non-representational theory of perception.

4.1 Introduction

Whether perception involves the manipulation of representational content is currently heavily debated. A classic Marrian view sees perception as built up step-by-step from a 2D retinal image into the full-blown 3D world we find ourselves in. Each individual step involves a separately produced representational image in which the brain detects relevant features to produce a more detailed, experientially rich picture (1982). There has been a large variety of theories that all regard representational content as central to perception (see Pitt 2017 for an overview).[25] Opposed to this, there is a tradition that rejects the idea that representational

[24] The text of this chapter is published as van Es, T. (2019). The Embedded View, its critics, and a radically non-representational solution. *Synthese*, 1-17. https://doi.org/10.1007/s11097-019-09649-

[25] The textbook interpretation of Marr is representational, but it is up for debate whether this is the only interpretation (Orlandi, 2014, note 3).

content plays any role in perception. Typically, non-representational approaches emphasize embodiment and environmental interaction, stressing that action and perception are intimately connected (Gibson 1979; Varela et al. 1991; Di Paolo et al. 2017; Hutto and Myin 2017).[26]

Recently, many theorists have tried to develop an approach that avoids either of these extremes. A popular strand here is prediction error minimization theory, which attempts to incorporate the intimate connections between action and perception into a fully representational framework (Hohwy, 2013; Clark, 2016). The Embedded View (EV) furthers this compromissory agenda by steering a middle course in the so-called 'representation wars'[27] (Orlandi 2014; see also 2012; 2013). Orlandi argues that perceptual processing is non-representational, but its perceptual products are representational. Yet EV has seen criticism from both sides for its limited use of representations. Where representationalists argue EV incorporates too few representations, non-representationalists argue there are too many.

In this paper, I argue that, to respond to the criticism from both representationalists and non-representationalists, EV is best off going fully non-representational. I will first set the stage by describing EV. Proceeding, I will then analyze the debate to show the perceived shortcomings from EV. I distinguish four such objections: 1) the argument that Orlandi leaves out the representational middleman; 2) the argument against *explanatory exclusionism*, according to which Orlandi seems to rule out representational readings prematurely; 3) the argument for the explanatory benefits of representations, according to which there may be benefits to invoking representations that have been overlooked, and 4) an inconsistency charge in EV's ascription of representational status. I argue that the first three objections can be deflected, and so pose no threat to EV. Yet the fourth objection, the inconsistency argument, is successful and thus problematic for EV. Despite this, I argue that the tools and the solution are found in Orlandi's own work (2014; and 2012 respectively), hence EV can easily be amended to incorporate this criticism. Indeed, I will argue that, by taking the explanatory tools available to EV more seriously, the range of applicability of EV can be shown to cover those aspects of perception that Orlandi thinks require representation. If we buy into this solution, EV ceases to be the compromissory view Orlandi (2014) put forward and becomes wholly non-representational. In turn, it benefits in having resolved the raised objections.

[26] In both instances there is an incredible variation of theories, and though some certainly find themselves more in one camp than another, a certain degree of fluidity between these camps exists.
[27] I think the name 'representation wars' stems from Clark (2015).

4.2 The Embedded View Explained

In our processing of visual stimuli, we are sensitive to statistical regularities in the world (Mole and Zhao 2016, p. 366). This means that certain regularities that we have encountered before influence the way we engage with particular stimuli; particular changes in light intensity may for example indicate edges. Traditionally, it is thought that the information of the statistical regularities needs to be encoded in our brain to explain this sensitivity. For example, Marr argues that the brain encodes very specific statistical information regarding the occurrences of edges (Marr and Hildreth 1980, p. 202; Marr 1982). More recent proposals have reimagined the manner in which this encoding works. Still, the necessity of such encoding remains.[28] For example, currently popular prediction error minimization theories state that any statistical regularity we may be sensitive to must be, in some way or some form, encoded in causal-probabilistic models of the world (Gładziejewski 2016; see also Hohwy 2013; Clark 2016).

EV is not at odds with this sensitivity to statistical regularities. However, it challenges the encoding requirement. Orlandi argues for EV by way of an inference to the best explanation. Here she relies on a widely shared assumption that representations need only be invoked "when appeal to environmental and other conditions fails, or is not illuminating" (Orlandi, 2014 p. 7, see also Note 15; Pylyshyn 1984, p. 26; Brooks 1991; Burge 2010; Fodor 1987; Segal 1989; Van Gelder 1995). The tactic is thus to provide a non-representational alternative to standard representational explanations with at least equal explanatory power. With this in place, invoking representations is unnecessary. According to Orlandi, our visual system *relies* on the world's statistical regularities themselves rather than needing to encode those regularities internally. "Relying on a fact", then, "means acting in accordance with a fact, and with a corresponding principle, without representing either" (Orlandi 2014, p. 3; Orlandi 2012, 2013; see also Ramsey 2007). In which case, perceivers can rely on the facts in the world without having to representing those facts. Orlandi states:

> We can imagine (...) a connectionist network trained to detect something by being repeatedly exposed to it. Such training causes the network to display characteristic patterns of activation where low-level configurations are associated with high-level ones and *vice versa*. We can then think of a high-level state as 'checking' the pattern of activation at the lower, sensory level where this ultimately just means that the high-level state activates in a way that is more or less compatible with the lower-level pattern of activation. (Orlandi 2014, p. 88)

[28] Predictive processing (see Hohwy 2013 and Clark 2016 for some good introductory monographs) is one such theory, in which the rules are thought to be represented in the form of generative models. Some recent takes on the theory however, advocate a non-representational reading (Kirchhoff and Robertson 2018; Bruineberg et al. 2018; Hutto 2017)

The idea here is that a connectionist network can become sensitive to a particular regularity. In more detail, the high-level states of the network become sensitive to the regularities at the lower, sensory level. This constitutes a certain bias in the system such that the higher level states will track the patterns of activation of lower level states, 'acting' in accordance with similar patterns, thus tracking, for example, edges. This, it should be noted, is a 'mere' tracking activity, and thus need not be thought of in representational terms. Indeed, there is nothing intrinsically representational about the reliable co-activation or covariation of two (parts of) systems (Hutto and Myin 2013). Further, Orlandi says:

> The transition between these two tracking states *is not regulated by an encoded assumption*. We can think of it as a mere associative transition. What is relevant is that there is a strong connection between states that track discontinuities in intensity and states that track edges. (2014 p. 153, emphasis added)

What this means is that the lower level states track the discontinuities in intensity, whereas the higher level states activate in accordance with particular patterns of lower level states, whilst tracking edges. Crucially, this associative transition does not require any encoding of the statistical regularities. "[T]he facts (...) explain why the device does what it does without being represented by the device" (Orlandi, 2012 p. 561). In explaining a particular behaviour, then, the facts themselves factor into our explanations, not representations of those facts.

Thus, we could certainly describe this particular network as *representing* the statistical regularities that it is sensitive to. *Prima facie*, at least, it is not conceptually incoherent. Adding this representational gloss, however, confers no explanatory advantage. Moreover, it comes at some cost. For it would ascribe more labor to our cognitive architecture than necessary. And it would lose some of the original ontological parsimony (Orlandi 2013, p. 739). Recall that, as we saw above, "representational notions are only needed when organic, functional, and environmental conditions are insufficient" (Orlandi 2014, p. 54). Here however, they suffice just fine.

One of the key issues with representational content is its origin (Hutto and Myin 2013; 2017). Broadly, the issue is as follows. Though representations have been defined in a multitude of ways, a core feature that distinguishes a representational relation from a merely covariational relation, is that they have content: they describe their target in a way that it may

not be so (Travis 2004).[29][30] The hard problem of content lies in finding a naturalistic origin for content: a task that is, as of yet, unfulfilled (Hutto and Myin 2013, 2017).

Ideally, an alternative proposal, like EV, should avoid the problems it is supposed to resolve. If so, we can ask: what is the origin of the reliance relation? Orlandi speaks of the visual system as being *wired*, having features that "developed, and continue to develop under evolutionary and environmental pressure" (2014 p. 3). She argues that it would be perhaps more surprising if evolution had *not* wired us to be sensitive to, say, the particular changes in light intensity corresponding to edges (Orlandi 2014, p. 154).[31] However, while it is likely that the wiring of our visual system is a consequence of evolutionary pressures, it seems more is needed if we are to account for the full range of statistical regularities we can become sensitive to. For example, a lot of these sensitivities emerge only later in life, and some at timescales that evolutionary adaptation could only dream of! How then does EV account for this malleability of the sensitivities of the visual system? Or in other words: how do we account for the features that develop not only under evolutionary, but also under environmental pressures?

Typically a wired system is conceived of as *hardwired* and incapable of adapting to the environment (Orlandi 2014, p. 146; see for example Rock 1983, p. 277). This need not be the case, however. Remember that a connectionist network is wired in the sense EV deems the visual processing system to be. Orlandi states:

> Networks are wired systems capable of adjusting to context. In the learning period, they come to process information in a new way without having any localizable structures that encode instructions. They come to be sensitive to the presence of mine echoes by simply adjusting their connections. Similarly, pigeons may stop pecking on a key as a result of extinction or counterconditioning without following any rules. (Orlandi 2014, p. 147)

She continues:

> The fact that they are adaptable in this way does not mean that they are rule-following. Rewiring due to inverting lenses, or to reinforcement and reward, amounts to changing the way a certain function is performed through a physical intervention. It constitutes an instance of malleability without being an instance of following (new) principles. (Orlandi 2014, p. 147)

[29] This is not to say that content is all it takes for a relation to be representational, only that it is minimally necessary.

[30] This excludes deflationary notions of representation. In this paper, representations are minimally required to have content, as to not over-generate representations or ascribe them too liberally (see also Ramsey, 2007).

[31] Similar arguments can be found in the work of Järvilehto (1998) and Warren (2005).

This means that adapting a function to better fit the current environment does not entail an exchange of less fit rules for more adaptive rules. It does not entail any form of rule-following whatsoever. A functional change can do the trick of malleability just as well.

Above I have given a brief overview of how Orlandi envisions visual processing to be possible non-representationally, whilst also retaining sensitivity to statistical regularities. Orlandi also covers typical issues for non-representational accounts of perception, such as misperception, illusions, stability and constancy (Orlandi 2014, ch. 4) and multi-stability (Orlandi 2012). However, although Orlandi argues strongly for non-representational visual processing, she does think of the visual product as being representational. It is this concession to representationalism that puts her in the crossfire of the 'representation wars'.

Yet, one would think that if representations are not needed to account for the wide range of cases covered above, then what still remains unexplained? I will answer this question by identifying what Orlandi deems necessary for something to count as a representation. She discerns three distinct features of a representation: 1) standing in, 2), informing, and 3) guiding performance (Orlandi 2014, p. 9). Thus, representations need to "*stand in* for something other than themselves and in so doing they act as *mediators* within a system or a process" (p. 9). Representations also need to inform and although she concedes that the term 'information' is heavily disputed in representation contexts, Orlandi thinks that representations inform, say or 'mean' something about the target (p. 9). Finally, representations need to be "causally active", "guiding the behavior of a system or organism, described in certain general terms, by standing in and informing" (p. 10).

My focus here will be on the first feature: standing in. For us to legitimately describe a system as containing representations that *stand in* for something, these representations need to be *decouplable* from what initially caused the representation. This means that the state should stand for its representational target (that which it represents) even in the *absence* of that target. If we take a portrait picture to be a representation, for example, the picture stands for the portrayed person even in the absence of the person. Of particular interest here is the term absence. For Orlandi, anything that does not directly impinge on our senses, should be deemed absent.

She uses two cases to exemplify this. First, when a rabbit is looking at a coyote, one may think that the coyote's reflected light is currently directly impinging on one's senses. Orlandi concedes that the *front* of the coyote is, but not a full coyote. For example, the coyote may even be partially occluded by a rock it is standing behind. Regardless, the rabbit reacts to the presence of a full coyote (Orlandi 2014, p. 127). Second, Orlandi claims that the same

phenomenon occurs whenever we see, say, a cow. For although only the image of the frontside of a cow impinges on our senses, we still react to it, judge it and take it to be a full-blown cow. She argues that "this *taking* constitutes an early representational capacity. It is a capacity that involves an abstraction" (Orlandi 2014, p. 150). Representational capacities are thus twofold. The perspectivally invisible (and thus absent) backsides of the objects we encounter require representation, since these aspects are abstracted from the direct stimuli. Further, to *take* the light reflected off a cow *as* a full-blown 3D cow also requires representation. The representation meets all three conditions: it *stands in* for the currently invisible backside, it *informs* about the cow, and it *guides our actions* in the sense that we judge it to be a full-blown cow (and, under this construction, this judgment guides us so that we may approach it to pet it etc.).

In sum, EV pulls visual processing apart from the visual percepts it produces. If EV is right, rather than needing to encode particular rules, we instead rely on the facts in the world. There is thus no need for the visual system to represent the statistical regularities it relies on. However, in the visual product there is the need to invoke representations. For despite being only exposed to a single side of an object at a time, we nonetheless take the stimulus to come from a full-blown object. Further, we take the light reflected off of particular objects to be caused by objects of particular types. If so, then the visual product does require representational capacities.

4.3 The Embedded View Under Fire

Not loyal to either the representational or non-representational sides, EV has attracted criticism from both camps, creating an interesting bipartisan agreement that halfway is no way at all. The agreement extends beyond this point however. I distinguish four distinct objections put forward by opponents of EV, the first three explicitly in favour of a more thoroughly representationalist approach, the final amenable to theorists across the representational-non-representational divide. The first objection is that of the relevant middleman, and is specific to Bayesian brain theories. In Bayesian brain theories, priors are often identified with representations (Hohwy 2013). The argument is that priors requiring further explanation does not mean they do not have any explanatory power so that they need to be excluded *tout court*. The second objection is against explanatory exclusionism: this argues against the notion that, if there are non-representationalist explanations, then their representationalist counterparts become obsolete. The third objection states that representations hold explanatory benefits for

our scientific endeavours, and thus need to be included. The final, unifying objection is that Orlandi's application of representations is inconsistent. Below I will discuss these objections in more detail.

4.3.1 The relevant middleman

Rescorla argues that Orlandi ignores the middleman. He says: "[i]f we explain X by citing Y and then explain Y by citing Z, it hardly follows that Y is explanatorily irrelevant". He explains this with an example: "a physicist can at least partially explain the acceleration of some planet by citing the planet's mass and the net force acting on the planet, even if she does not explain why that net force arises" (Rescorla 2015). Put differently, even if Y requires further explanation, it may still serve in explaining X. Y's requiring further explanation does not diminish its explanatory power.

To be precise, Rescorla's point concerns Bayesian modeling theories of perception and cognition. An issue Orlandi brings up is that it is unclear where the priors, that is, the brain's anticipation of what it is about to encounter, come from (Orlandi 2014, p. 91-93). An elaborate discussion of the issue of the origin of priors in Bayesian theories of perception is outside of the scope of this paper, but it is roughly as follows. According to Bayesian theories of perception, the brain attempts to reduce the uncertainty concerning the stimuli entering the system (X in Rescorla's example). An issue here is that, for any one stimulus, there is conceptually an infinite amount of possible causes (the underdetermination problem). To limit the hypothesis space, then, priors are introduced. These are (basic) assumptions about the way the world works that restrict the variety of possible causes of the stimuli (Y in Rescorla's example). Standardly, Bayesian brain theories are considered representational. The representations are thought to be instantiated in the form of priors (Hohwy 2013; Gładziejewski 2016).[32]

In this sense, the question about the origin of priors in Bayesian brain theories, though technically separate from, can be cast as a particular instantiation of the question about the origin of representations, specifically in the context of Bayesian brain theories. Indeed, if priors are thought to be representational in Bayesian brain theories, then, in that context, a discussion of the origin of priors in Bayesian brain theories is also a discussion of the origin of representations in Bayesian brain theories. Rescorla thinks that, due to this controversy, Orlandi

[32] Discussion of Bayesian brain theories is outside the scope of this paper. Hohwy (2013) and Clark (2016) are two excellent monographs for an introduction into the most popular form of Bayesian brain theories.

excludes the priors (Y) because her environmental account (Z) would be needed to explain our visual system (X) anyway. Rescorla treats it as a form of cutting out the middleman.[33] Instead, he argues that the explanatory value of Y is not impacted by it requiring further explanation. Even more, he argues that "one can explain X by citing Y, without in turn explaining Y" (Rescorla, 2015). Representations thus need not be explained for them to be explanatorily useful, according to Rescorla. Put differently, even if the priors (and thus the representations) are to be explained further with Orlandi's embedded explanation, the priors themselves still have explanatory value.

4.3.2 Against explanatory exclusionism

EV is also accused of committing a form of *explanatory exclusionism* (Polger 2015, p. 345). Polger considers Orlandi to be choosing between competing explanations. He wonders: "[w]hy should we embrace an explanatory exclusion principle to the effect that the availability of functional or mechanistic explanations rules out the legitimacy of representational explanations"? (Polger 2015, p. 345). Polger ponders rhetorically why representational explanations should be less preferable to functional or mechanistic explanations. Non-representational explanations are not, Polger thinks, by default preferable to representational explanations. The idea is that Orlandi would have to do more to show that her non-representational explanation is preferable to competing representational explanations. Indeed, there is no good reason to throw representational explanations out simply because there are non-representational accounts.

4.3.3 Representations with benefits

Another objection against EV is that invoking representations in our explanations of perceptual phenomena offers explanatory benefits. Leaving our explanation to merely "organic, functional, and environmental" conditions, so it is claimed, is insufficient. The addition of representations is supposed to make up for the difference. Rescorla argues that "Orlandi's analysis entrains a significant loss in explanatory power. If we adopt a realist perspective on priors", he thinks, "we can explain why various changes in environmental conditions yield various changes in the mapping from sensory stimulations to percepts: namely, because the priors change a certain way" (Rescorla 2015).

[33] As we shall see below, I think this portrayal of Orlandi's position is incorrect.

This again concerns Bayesian modeling theories of perception and cognition, where the priors refer to the brain's anticipation of what it is about to encounter. Orlandi *de-representationalized* this notion, thinking of such anticipation in functional terms, the training of a neural network. Rescorla then argues that priors add explanatory utility by exchanging a step in Orlandi's story. Rather than the environment influencing the neural network's training, it influences the priors. Both of these are thought to be the explanation for the change in our percepts in their respective theories. Rescorla thinks that this exchange, the realist position on the priors, adds explanatory value.

Polger (2015) puts forward a similar point. He thinks that "[r]epresentationalist explanations might be justified by their utility for explaining how those mechanisms of biases and constraints came to be, or what those systems have in common with other human perceptual systems and with physiologically distinctive visual systems in other creatures" (Polger 2015, p. 345). The addition of representations is thought to help both evolutionary and environmental explanations of the functional features described by Orlandi. The idea seems to be that, if we posit representational contents as that which is present in all our perceptual channels (both our vision and hearing can be imagined as being representational) and also for such perceptual channels in non-human creatures, the shared representational contents can serve as a starting point for explaining what we have in common.

4.3.4 Inconsistency

The third and final objection has managed to unify both representationalists and non-representationalists alike. It is an inconsistency objection concerning Orlandi's ascriptions of representations in EV. As we have seen in Section 2, Orlandi understands visual processing as non-representational, but the perceptual product does require representations. However, if we were to consistently apply Orlandi's conditions for what counts as a representation, so it is objected, then a lot more features she has termed non-representational, would in fact count as representations. She thus applies her own conditions inconsistently. Put differently, her conditions for representational content actually lead to a proliferation of such representations.

Recall that Orlandi argues that representations are used in abstracting the image of a cow from mere electromagnetic radiation that impinges on our senses. Polger argues that

> if this basic sort of abstraction is all that is required to count as a representation, then contrary to Orlandi, representation occurs at the earliest stages of visual processing … It seems to me that Orlandi has made it too easy for early visual states to count as representations. (Polger 2015, p. 344)

Despite Orlandi's care in the ascription of representations in visual processing, Polger argues that this frugality is not mirrored in explaining the visual product. Mole and Zhao (2016) argue along the same lines. In particular, they discuss one experiment that displays features of our visual system that, if we are to follow Orlandi's conditions, would be representational. Yet, when analyzed, all of the examples they offer are exactly of the sort that EV is supposed to de-representationalize, that is, they are examples in which the subject becomes sensitive to a newfound statistical regularity. Mole and Zhao portray these as counterexamples she may not have been aware of.[34] However, I claim these examples may be better understood as demonstrating the shortcomings in Orlandi's application of the stated conditions for representational content.

Hutto and Myin raise a similar point. They state:

> Consider that Orlandi argues, as does REC[35], that positing sensitivities to statistical patterns of variation is all that is needed in order to explain perceptual processing and that there is no need or advantage in positing contentful representations in such explanations. Yet such statistical patterns are just as unavailable to directly inform perceptual processing as are the backsides of objects are to inform the final products of visual experiences. (Hutto and Myin 2017, p. 162)

Absence, or unavailability, is a central term in Orlandi's usage of representations. This is intended to cover the *decouplability* requirement as well as the *standing in* requirement for representational content. What is present and currently available in the environment, need not be represented internally. The 'original' may be used without making a 'copy'. Yet if this is all that it takes, Hutto and Myin argue, then what are we to make of the statistical regularities themselves? Surely, if the image of a cow requires representational abstraction from electromagnetic radiation, then how do the statistical regularities present themselves directly? Statistical regularities, of course, are never witnessed directly. Hutto and Myin thus argue that the representational content of statistical regularities stands or falls together with that of the backsides of objects. This too points out Orlandi's inconsistency in the ascription of representations. In sum, Orlandi's inconsistency in the ascription of representational content is

[34] Mole and Zhao (2016) argue roughly that the subjects in the experiments have become sensitive to newfound statistical regularities in the environment, so that they could not have been hardwired. With a hidden assumption that systems can only become sensitive to new statistical regularities with representations, they conclude that representations thus *must* in some cases be used in visual processing. This bypasses Orlandi's explanation of malleability.

[35] REC stands for Radically Enactive Cognition, and is a non-representational approach to cognition proposed by Hutto and Myin (2017; see also 2013).

consistently picked out by people from a wide variety of theoretical colors. As we will see below, it also is the most pertinent issue for EV.

4.4 The Embedded Fire Extinguisher and the Ember

In this section I will give a response to the above objections from the standpoint of EV. In the cases of answering the objections mentioned in Sections 3.1, 3.2, and 3.3, I will attempt to offer a rebuttal. I argue that the inconsistency issue of Section 3.4 proves to be trickier, although the explanatory tools available to EV provide the solution.

4.4.1 The controversial middleman

Rescorla objects that Orlandi cuts out a relevant explanatory mediator. Specifically, he argues we can explain X by citing Y, which in turn can be explained by citing Z. The appearance and necessity of Z, he argues, need not impair the explanatory utility of Y. He makes an analogy with a physicist, who explains the movement of celestial bodies (X) with net force (Y). Net force itself requires further explanation, perhaps by citing the mass of all physical bodies in the system (in the form of Z). Nonetheless, it can be used to explain X. Though I think the example is solid, it need not generalize to other cases. In particular, it need not apply to the use of representations to explain perceptual phenomena.

There are two relevant differences here. 1) Orlandi could be understood as cutting out Rescorla's imagined middleman, that is, if we can explain X by citing Z directly, then we do not need to first explain X by citing Y before finally turning to Z. For example, if we can explain perception by citing functional, organic mechanisms, then we need not explain perception by first citing priors, which can then be explained by citing functional mechanisms. If this is correct, then contrary to Rescorla, it is not the case that Y becomes explanatorily irrelevant because it is explained by Z. For the functional mechanisms do not explain the priors, but rather the phenomena themselves. In which case, Y and Z are thus *competing* explanations both intended to explain X. Y's representational explanation has thus never started her job as a middleman in EV; she thus could not have been fired as Rescorla seems to argue.

2) Recall the widely shared assumption that representational explanations are only necessary if we lack simpler, ontologically and computationally more parsimonious, explanations that we have seen in Section 2 (Orlandi 2013, 2014; Pylyshyn 1984, p. 26; Brooks 1991; Burge 2010; Fodor 1987; Segal 1989; Van Gelder 1995). It is thus not thought to be conceptually incoherent, or otherwise impossible for there to be a representational explanation

in tandem with a functional, organic or environmental explanation. Instead, it is thought to be *unnecessary* for us to appeal to a term as ontologically and cognitively heavy as representations if we can show simpler explanations are sufficient. These considerations follow from Orlandi's inference to the best explanation strategy. She does not aim to show that representational explanations are incoherent or incompetent. Instead, she offers a better explanation, which at least in part relies on being a simpler explanation.

Further, at least to my knowledge, net force is an uncontroversial term in physics. This status allows it to be useful in explaining celestial movements. However, if net force were to turn out to be a fundamentally questionable concept, that is, one for which there is no naturalistic foundation (as of yet), then using it to explain celestial movement would not be warranted. Consider using the touch of Zeus to explain celestial movement. At least one of the reasons this explanation is not accepted is because the touch of Zeus has no naturalistic foundation. I propose that the same goes for representational contents. Consider that the project to naturalize content has so far proved unsuccessful, and increasingly more people are becoming pessimistic towards a resolution (Rosenberg 2015 and Hutto and Myin 2013; 2017 among others). Rescorla acknowledges this difficulty (2015), but fails to see why the lack of even a prospect of a naturalistic explanation makes the use of representational explanations inherently unappealing.

Note, however, that Rescorla's point only indirectly concerns representations. This is because the priors he discusses are typically conceptualized as representations in Bayesian brain theories (Hohwy 2013; Gładziejewski 2016). To recapitulate, to limit the hypothesis space for the brain's estimate of what stimuli it is about to encounter, priors are introduced. Here priors are continuously updated assumptions about the structure of the world. It is still unclear however, where priors originate from. Nonetheless, he argues, priors need not be explained in order to have explanatory force. Yet this seems highly problematic. We explain how the brain reduces uncertainty by reference to priors, of which the origins are unknown. It seems like we explain a mystery by a further mystery, merely moving the goalpost without actually cashing out an explanation. Indeed, it echoes the problem of invoking representations without explaining them. In both cases we attempt to use unexplained, quite possibly unexplainable concepts to explain the phenomena.[36]

[36] See Clark (2016) for a proposed solution to the origin of priors, and see Hutto (2017) and Hutto and Myin (2017) for an objection to this.

4.4.2 Explanatory frugality

Polger (2015) wonders why we should exclude representational explanations merely due to the availability of functional or mechanistic alternatives. On the surface, this exclusion seems to presuppose the superiority of non-representational contenders from the offset. But remember that, as of yet, the naturalization of representational contents remains unsettled. Further, even among those that use representational explanations it is commonly thought that the ascription of representations is only necessary in the absence of simpler alternatives (Orlandi 2013, 2014; Pylyshyn 1984, p. 26; Burge 2010; Fodor 1987; Segal 1989). The only way out, some think, is to simply accept content as a basic metaphysical particle of existence (Shapiro 2014 p. 218). If non-representational explanations are explanatorily equal to representational explanations, then we should prefer the metaphysically lighter option (using Occam's time-tested razor), much like we do when rejecting the touch of Zeus as an explanation. Until the naturalization of content becomes a more promising area of research, equally or more powerful non-representational explanations should be preferred over representational ones. In this particular sense, non-representational explanations, otherwise equal, are superior to representational explanations. This, again, does not mean it is incoherent to add representations to a non-representational explanation. It is simply that doing so does not add explanatory value (as the non-representational explanation is at least equal in explanatory power to the representational explanation). Moreover, it unnecessarily adds a metaphysically unclear and cognitively heavy component to an otherwise simpler explanation. In short, it is not incoherent. But, it is not wise either.

4.4.3 Begging questions on explanatory benefits

Rescorla argued that we could gain explanatory benefits by adopting a realist stance on priors. This would allow us to explain *why* our percepts have changed. A realist's priors are, in a sense, quantifications of the neural network-like activity described by Orlandi (2014). Not only does the neural network change its wirings, but these wirings represent numerical statistical values distributed over a variety of possible outcomes. But does this offer more of an explanation than the functional approach? It is easy to see that EV's wiring explanation does the same work, by adapting Rescorla's sentence: "we can explain why various changes in environmental

conditions yield various changes in the mapping from sensory stimulations to percepts: namely, because the [wirings] change a certain way" (2015, edited by the author).[37]

There are two reasons why, in itself, a realist position on priors does not add any explanatory benefit. First, it is unclear what quantification of a biological process adds beyond numbers. These numbers may feature well in models of biofunctional mechanisms. But there is an important difference between using these numbers and models to describe or predict animal behavior and actually *ascribing* these numbers and models to the animal itself. Doing the latter prompts the question of how these models are instantiated in the brain, in what way the numerical probabilistic values are 'real', and even whether this requires a realist position about, say, abstract objects (and the problems that ensue). None of these questions are easily answered.

Second, even if we were to allow that there is an explanatory benefit, it is unclear how introducing a feature that is in itself fundamentally contested would help our explanatory efforts. This draws back on our earlier 'touch of Zeus'-style explanations. Until we have, at the very least, a promise of how the 'touch of Zeus' can be naturalized, naturalistic but otherwise equal explanations should take precedence. The same should go for representational contents.

Polger argues that representations can explain how our functional mechanisms came to be, and expose the similarities between similar systems within human organisms (think of different perceptual channels), but also similar perceptual channels of different organisms (such as a dog's vision). Yet it is unclear why representational explanations would do a better job at explaining how particular functions came to be than organic and functional explanations: the type more typically used in evolution-theoretic explanations. More importantly however, it seems circular. We first posit that our and non-humans' sensory channels share the use of representations, before using this posit to explain those similarities. Finally, we claim this supports a realist position of our earlier theoretical posits. With this we have come full circle, and viciously so.

[37] This is not to discredit the empirical research programme of Bayesian brain theories of perception and cognition, which seems promising (see Hohwy, 2013; Clark, 2016). This is a point about how to interpret the priors used in this research. There is a difference between using Bayesian models as scientists to model and predict behaviour of animals, and a realist position concerning the models used as existing in the animal's head (or body). Rescorla (2015) favors an interpretation of the scientific literature that involves a realist position of these priors. Here I argue that doing so does not add any explanatory value.

4.4.4 The inconsistency and the non-representational solution

Polger (2015), Mole and Zhao (2016), Myin (2016) and Hutto and Myin (2017) have all argued that Orlandi's ascription of representations is inconsistent. Each of them has argued that at least one other feature of the visual processing system fulfills EV's conditions. In this section, I will focus in particular on Hutto and Myin's (2017) suggestion that the backsides of objects and the perception of abstract properties such as COWness warrant the same representational status as the statistical regularities, but the solution extends to other cases.

Recall that Orlandi identifies three conditions for anything to be a representation: 1) standing in, 2), informing, and 3) guiding performance (Orlandi 2014, p. 9). A possible objection to Hutto and Myin's suggestion is then that our 'perception' of backsides of objects legitimately covers condition (1), whereas getting attuned to statistical regularities does not. The backsides are (to some extent) present in our visual percept (Orlandi 2014, p. 125). This requires, according to Orlandi, a representation to stand in for a backside that is not technically present, but does appear in our percept. The statistical regularities however, do not appear in the same way. There is only the brain's interactional attunement to regularities encountered in the world. There is thus nothing internally that 'stands in' for them. This would cause the statistical regularities to not meet condition (1). The backsides of objects do meet that condition.

The issue is whether this distinction between statistical regularities and the backsides of objects as being non-representational and representational respectively is justified. Both the statistical regularities and the backsides of objects are stimulus-free; they are absent in the sense discussed in Section 2. At this point, the ascription of representational status seems stipulative. Orlandi concedes that representational explanations for sensitivity to statistical regularities are possible. The reason why this feature of visual processing is de-representationalized, is because she provides a functional explanation that does not require representations. This is a metaphysically and cognitively less taxing way to explain the same phenomena, and should thus be preferred, given all that has been said above. Statistical regularities then do not have a stand-in, thus do not meet condition (1), and so are non-representational.

Orlandi (2014) does not discuss non-representational options for 'seeing the backsides', and thus makes it seem like the representational interpretation is the only route (perhaps much like how the sensitivity to statistical regularities may have seemed before her account). This, in the context of Orlandi (2014), makes it appear that she consistently applied her own conditions of favouring non-representational explanations when present. Interestingly, Orlandi

(2012) has previously offered a non-representational solution to this issue. Here, our interaction with the rigidity of objects (and those objects thus having backsides) is explained through the same mechanism as other statistical regularities in the world, like the co-occurrence of changes in light intensity and edges. After all, there is a robust statistical regularity so that the patterns of light pertaining to objects co-occur with these objects' rigidity. Indeed, it is a regular occurrence that the objects we encounter are 3D objects with front- and backsides. She says:

> "Objects in the world are typically rigid and the visual system can rely on this fact to produce a representation of objects directly from the retinal stimulation. Because the causes of retinal stimuli are typically rigid, we end up seeing the world that way" (Orlandi 2012, p. 560)

The rigidity of objects is thus explained by way of the exact same non-representational process as, for example, edges are. We do not have to invoke internal representations of the backsides of objects to explain our perception of rigid objects. Instead, we can rely on the robust statistical regularity that the objects we encounter are typically rigid. Any object that is rigid has a backside (even if they are not always the backsides that we have become attuned to, like a realistic cardboard cutout of a cow)[38], so if we see rigid objects, we 'see' the backsides to the extent necessary. Positing that we need representational content for the backsides of objects is superfluous.[39]

The further representational capacity concerns the abstraction involved in *taking* a particular pattern of electromagnetic radiation impinging on our senses as an image of a cow. It is now not difficult to see how her own proposed mechanism can be employed to explain this capacity to some extent. There is an important distinction that needs to be explicated here. I suggest that what is captured in the *taking* capacity by Orlandi is two distinct capacities. In saying that we *take* a particular pattern to be a cow, 1) there is a judgment that goes beyond perception in itself in which we *judge* the pattern to be a cow, and 2) there is the capacity to, by being exposed to that particular pattern, interact with the cow in the world we encounter it. (1), the judgment, is to be read in the same way we *take* a sarcastic comment to mean the opposite of what it explicitly says. This *taking* capacity is a lingual, socioculturally developed

[38] The possibility to make mistakes is not reason to doubt this explanation. Instead, it speaks in its favor. After all, it is because we are wired to see rigid objects when encountering particular patterns of stimulation, that the system sometimes fails and we become susceptible to illusions. Complete exposition of this is outside the scope of this paper, but see Orlandi (2014, ch. 4) for a more elaborate explanation, as well as similar non-representational, embedded explanations of misperception in general and multi-stability.

[39] See also Di Paolo et al. (2017) that explain the 'seeing of backsides' in terms of interactional relevance. We approach a particular object as having a backside when the backside is relevant to our current activity. A keyboard is not interacted with as having a backside currently not visible to the eye, unless we, say, pick it up. A teacup on the other hand, typically affords picking up and its backside will thus be 'seen' more regularly.

capacity that goes beyond the actual hearing of the comment itself. As such, this aspect of the *taking* capacity is not an explanandum in the category of perception, and with that it is outside the scope of EV, and this paper. Roughly, the idea is that language, or content-involving practice, is an extension of action in a sociocultural environment (further exposure of this idea is outside the scope of this paper, but see Moyal-Sharrock 2019; Hutto and Satne 2015; Hutto and Myin 2017 for positions along these lines).

Part (2) of the *taking* capacity, the capacity to interact with a cow in the world, by being exposed to a particular pattern of stimulation *is* part of the explanandum of EV, and can also be explained. The perception of a cow relies on the statistical regularities involved with the particular patterns of stimulation caused by cows, and the presence of cows. The perceptual system can become sensitive to these statistical regularities, and may rely on them in our perception of cows. These sensitivities go on to influence our interactional behaviour and perception much in the same way our sensitivity to edges does. Indeed, this explains our perception of cows in the same way the statistical regularities and interactional history explain edge detection or our perception of rigid objects. A primary difference is that a sensitivity to edges can plausibly be thought of as having developed by evolutionary pressures, whereas a sensitivity to cows is more likely to have developed ontogenetically due to environmental pressures. This allows EV to circumvent the inconsistency objection raised in Section 3.3. Given Orlandi's idea that representational explanations should only be instantiated if functional, organic or environmental explanations are insufficient, EV should thus adopt this non-representational approach.

What, then, is left of the compromissory position EV intended to take in the representation wars as put forward in Orlandi (2014)? EV started out as a view that held onto non-representationalism for visual processing, but argued for a representational view of the visual product. The solution for the inconsistency problem offered above de-representationalizes the two central features of vision that Orlandi thought required the ascription of representational capacities to be explained. As a consequence, EV has now gone fully non-representational and loses its status of a representationally compromissory view. Indeed, the view is now much closer to a fully non-representational view as defended by the likes of Hutto and Myin (2017).

Buying into this solution comes with a few further advantages. EV in its original form would still have to answer to the hard problem of content for visual products, that is, the problem of finding a naturalistic grounding for representational content, commonly agreed to be minimally necessary for representation (Hutto and Myin 2013). Non-representational EV

circumvents this hard problem of content by not positing representations in either perceptual processing nor perceptual products. This further increases the ascribed frugality of the brain's role as it need not trade in cognitively heavy representations. It also has ontological frugality. After all, the non-naturalistic status of representational contents is unlikely to change anytime soon (Rosenberg 2015; Shapiro 2014, p. 218; Hutto and Myin 2013, 2017). Further, not trading anymore in such unexplained posits allows this to remain metaphysically parsimonious.

Finally, it is worth noting that no new additions or similar *ad hoc* reasonings are added. Orlandi's body of work on EV already held the tools needed to fix the issue of inconsistency, while also broadening the explanatory range of its primary feature, namely, the outsourcing of cognitive tasks onto the environment.

4.5 Conclusion

In this paper I have assessed the debate around Orlandi's proposed Embedded View of vision. This view steers a middle course between representationalist and non-representationalist views by arguing that visual processing is non-representational, whereas its products are representational. Both representationalists and non-representationalists alike have challenged this view. I described four such challenges. I then offered rebuttals of three of those challenges, while also acknowledging that the fourth raises a significant problem. This problem concerned the inconsistency in Orlandi's ascription of representational contents. However, I showed that Orlandi has the tools necessary to fix this problem (and perhaps even part of the solution (see Orlandi 2012)). I conclude then that EV is best interpreted non-representational all the way down. As such, EV becomes a metaphysically and cognitively more frugal view, which, when placed within the context of the representation wars, is to its advantage.

References

Anderson, M. L. (2017) Of Bayes and Bullets: An Embodied, Situated, Targeting-Based Account of Predictive Processing. In Metzinger, T. and Wiese, W. (Eds.). *Philosophy and Predictive Processing: 4*. MIND Group

Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence, 47*, 139–159.

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*(5), 2417-2444

Burge, T. (2010). *Origins of objectivity*. Oxford University Press.

Clark, A. (2015). Predicting peace: The end of the representation wars. In Metzinger, T. and Windt, J. (Eds.). *Philosophy and Predictive Processing: 7*. MIND Group, 1–7

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

Di Paolo, E., Buhrmann, T., Barandiaran, X., (2017) *Sensorimotor Life: An enactive proposal*, Oxford University Press.

Fodor, J. A. (1987). *Psychosemantics*. MIT Press.

Gibson, J. J. 1979. *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Gładziejewski, P. (2016) Predictive Coding and Representationalism. *Synthese 193*: 559–582

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

Hutto, D. D. (2017) Getting into predictive processing's great guessing game: Bootstrap heaven or hell? *Synthese, 195*(6), 2445-2458

Hutto, D. D. and Myin, E. (2013) *Radicalizing Enactivism: Basic Minds Without Content.* MIT Press.

Hutto, D. D. and Myin, E. (2017) *Evolving Enactivism: Basic Minds Meet Content*. MIT Press.

Hutto, D. D. and Satne, G. (2015) The Natural Origins of Content. *Philosophia 43*(3), 521-536

Järvilehto, T. (1998) The Theory of the Organism-Environment: System: I. Description of the Theory, *Integrative Physiological and Behavioral Science, 33*(4), 321-334.

Kirchhoff, M. D., and Robertson, I. (2018) Enactivism and predictive processing: a non-representational view, *Philosophical Explorations, 21*:2, 264-281

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information.* Freeman.

Marr, D. and Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London. Series B; Biological Sciences, 207*(1167): 187–217.

Mole, C. and Zhao, J. (2016) Vision and abstraction: an empirical refutation of Nico Orlandi's non-cognitivism. *Philosophical Psychology, 29*:3, 365-373

Moyal-Sharrock, D. (2019) From deed to word: gapless and kink-free enactivism. *Synthese,* 1-21

Myin, E. (2016) Perception as something we do. *Journal of Consciousness Studies* 23(5–6): 80–104.

Noë, A. (2004) *Action in Perception*. MIT Press

Noë, A. (2006) Experience of the World in Time. *Analysis* 66 (1), 26-32.

O'Regan, J. K., and Noë, A. (2001) A Sensorimotor Account of Vision and Visual Consciousness. *Behavioural and Brain Sciences, 24*, 939 –1031

Orlandi, N. (2012) Embedded seeing-as: Multi-stable visual perception without interpretation. *Philosophical Psychology, 25*(4), 555-573

Orlandi, N. (2013) Embedded Seeing: Vision in the Natural World. *Noûs 47*(4) 727–747

Orlandi, N. (2014) *The Innocent Eye: Vision is not a cognitive process*. Oxford University Publishing

Pitt, D. (2017) Mental Representation. In Zalta, E. N. (Ed.) *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition).

Polger, T. (2015) The Innocent Eye: Why Vision Is Not a Cognitive Process, by Nico Orlandi. [Review] *Analysis Reviews 75*(2), 343-345

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

Rock, I. (1983). *The logic of perception*. MIT Press.

Rosenberg, A. (2015) The Genealogy of Content or the Future of an Illusion, *Philosophia 43*(3), 537-547.

Segal, G. (1989). Seeing what is not there. *The Philosophical Review, 98*(2):189–214.

Shapiro, L. (2014) Radicalizing Enactivism: Basic Minds without Content, by Daniel D. Hutto and Erik Myin. [Review.] *Mind 123*(489), 213–220.

Travis, C. 2004. The silence of the senses. *Mind 113*(449): 57–94.

Van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy, 92 (*7), 345–381.

Warren, W. (2005). Direct perception: The view from here. *Philosophical Topics, 33* (1), 335-361.

Part II

5 Living models or life modeled? On the use of models in the free energy principle[40]

Author

Thomas van Es 1

1 Centre for Philosophical Psychology, Department of Philosophy, Universiteit Antwerpen, Belgium

Abstract

The free energy principle (FEP) is an information-theoretic approach to living systems. FEP characterizes life by living systems' resistance to the second law of thermodynamics: living systems do not randomly visit the possible states, but actively work to remain within a set of viable states. In FEP, this is modeled mathematically. Yet the status of these models is typically unclear: are these models employed by organisms or strictly scientific tools of understanding? In this paper, I argue for an instrumentalist take on models in FEP. I shall argue that models used as instruments for knowledge by scientists and models as implemented by organisms to navigate the world are being conflated, which leads to erroneous conclusions. I further argue that a realist position is unwarranted. First, it overgenerates models and thus trivializes the notion of modeling. Second, even when the mathematical mechanisms described by FEP are implemented in an organism, they do not constitute a model. They are covariational, not representational in nature, and precede the social practices that have shaped our scientific modeling practice. I finally argue that the above arguments do not affect the instrumentalist position. An instrumentalist approach can further add to conceptual clarity in the FEP literature.

5.1 Introduction

In recent years an increasingly popular view on life and cognition has taken shape, which is based on the free energy principle (FEP). FEP, in this context, started as a mathematical tool to understand the workings of the brain (Friston, 2002, 2003, 2005, 2011). Yet since then, it's been applied to a wide variety of biological, psychological and social phenomena (Friston, 2013; Ramstead et al., 2016; Ramstead et al., 2018; Hesp et al., 2019). FEP is proposed to be a 'research heuristic' by which systems under investigation are modeled as free energy minimization systems (Ramstead et al., 2019). In some formulations, the system is also thought to implement this model either internally or in an embodied (or extended) sense so as to make statistical inferences about the external world (Friston, 2012; Hohwy, 2013; Clark, 2016; Bruineberg et al, 2016; Kirchhoff and Kiverstein, 2019). These different uses of FEP can be

[40] The text of this chapter has been published as van Es, T. (2020). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior*. https://doi.org/10.1177/1059712320918678

related to a difference between models *of* a target system made by scientists, and models *used by or embodied by* a target system, the latter of which is further divided into neurocentric and embodied/extended approaches.[41] In this paper, I argue that we ought to disambiguate these two uses of modeling. Further, I argue that only the former, instrumentalist understanding of free energy minimization in biology is justified, whereas the realist view that models are biologically implemented in organisms is not. This does not change the scope of what we can investigate with FEP, but it does constrain the conclusions that can be drawn from free energy models.

I will first give a brief overview of the FEP, explain its core theoretical posits and briefly discuss some of its applications. The distinction between models as mathematical constructs from models as they are ascribed to organisms and brains will stand central in my explanation. This deviates from the usual manner in which the FEP is introduced (in which, as I will argue here, the two are typically conflated), but it is required by the goals of this paper. Once FEP is on the stage, I shall disentangle the two uses of model, and show that they are conceptually distinct. After this I will argue that we should eschew a realist interpretation of these models. The first argument is the charge of overgeneration and trivialization of models. That is, if we accept a realist position on models in FEP, many non-living systems will be cast as modelers, entailing a trivialization of the notion of modeling. The second argument relies on Kirchhoff and Robertson's (2018) analysis of FEP as essentially covariational, not representationally contentful. Because models, both in representationalist and non-representationalist accounts, require more than mere covariation relations, a realist position on models in FEP is off the table. Finally, within scientific practice, covariational relations can be exploited to create models in the proper sense of the word, vindicating an instrumentalist approach to models in FEP. Before concluding, I will discuss some possible objections to my proposal.

5.2 The free energy principle explained: Brain and life

5.2.1 A model of life

FEP is a principle-first approach to living systems (Friston, 2012). According to the second law of thermodynamics, any closed system's entropy will increase over time. Consider, for

[41] In this paper, I will refer to such embodied/extended approaches as 'the embodied approach'. This captures multiple views, with plenty of differences among them. Whereas Bruineberg et al. (2016) focus on an embodied approach, Kirchhoff and Kiverstein (2019) argue for a more flexible, context-sensitive approach, ranging from extension into the world to merely the brain (see Clark, 2017; Hesp et al., 2019; Ramstead, Badcock, Friston, 2018, 2019 for similar positions).

example, a bottle of perfume in a closed room. As soon as you open the bottle, the perfume will slowly spread throughout the entire room: its entropy will increase. Living systems, however, put in a lot of effort to not disperse randomly, and, relative to all states they possibly could visit, remain within the subset of viable states (Friston and Stephan, 2007). In part, this is because they are semi-open systems: they regulate their exchanges with the environment to maintain homeostasis. Variational free energy is formally an upper bound on entropy, and thus, living systems minimize their free energy to stay alive (Frison, 2012, 2013). A research goal of FEP is to unearth the minimal characteristics that are necessary for any system to maintain its boundary with the environment. This is done with the idea that free energy minimization is an imperative of life (Friston, 2012). Via mathematical modeling, FEP can bring to light the relations between living systems at multiple spatiotemporal scales (Kirchhoff et al, 2018; Clark, 2017).

If we suppose that every organism has a clearly defined set of states, *characteristic* states, in which the organism remains alive, there is a boundary of states within which the organism is alive, outside of which the organism is dead (Friston, 2012). A state, here, corresponds to a particular value for a particular variable in a model. The characteristic states of a living system, then, are a collection of values for particular variables in which the organism is alive. The characteristic states differ greatly per phenotype: a fish stays alive in rather different circumstances than we humans (Friston, 2012; Bruineberg et al., 2016; Kirchhoff et al., 2018). For any system at any particular point in time, we could make a probability distribution in which probabilities are assigned for each possible collection of states the system could possibly be in. If we assume such a distribution to be applicable to an organism (which is debatable, see Colombo and Wright, 2018; see also Longo et al., 2012), then there will be high probabilities for those states in which it is alive, low probabilities for fatal states. Conversely, viable states will have a low uncertainty, and fatal states will have a high uncertainty: they will be rather surprising. As such, the organism would do best to avoid those surprising states. Moreover, "surprise" in the long run, as it is used here, is considered formally equivalent to entropy, which in turn means that variational free energy is an upper bound on surprise (Friston and Stephan, 2007). This means that, by minimizing free energy, the organism will indirectly also minimize surprise in the long run.

Uncertainty plays a central role in FEP. An organism's primary imperative is to stay alive: it is supposed to steer clear from surprising states, or in other words, it is supposed to minimize the uncertainty or surprise as it is calculated in the model. In the long run, when the organism remains within surprising states, the imperative is thus to resolve this situation. That

is, to make it less surprising. To do so, the organism needs to act. FEP theorists understand action to occur in terms of *active inference*. Active inference assumes we have a probability distribution over possible action policies of the organism, which in a particular context translate to particular actions to be taken (Kirchhoff and Kiverstein, 2019; Friston, 2012). It is an *inference* because it is an inferred prediction about a future state of the organism, and it is *active* because this inference is *conditional* on endogenous, organism-created action. That is, the prediction can be brought about by action of the organism (Friston, 2012). Put differently, the movement that the prediction is conditional on, is the action the organism ought to undertake in order to make the predicted state reality, to remain in unsurprising (viable, high probability) states. This means that, given the state of the organism and a probability distribution of action policies, we can predict a new probability distribution of action policies that will show which actions best minimize surprise in the current situation. That is, what the organism is best off doing to stay alive.

An alternative to using active inference to minimize free energy is to simply *update the probability distribution* so that previously surprising states are not so anymore. This is called *perceptual inference* (Friston and Kiebel, 2009; Friston, 2005; Hohwy, 2013). This will not help if the surprising states are directly fatal, but it may be helpful, say, if a human for the first time learns to swim. Being in the water is a highly surprising state, which, if applying solely active inference, would have to be steered clear of. Yet if the human learns to swim, this particular state is viable, in which case the probability distribution of states of the organism needs to be updated.

Important technical terms surrounding FEP are the *Markov blanket* and the *generative model*. A *Markov blanket* is a statistical tool with which a system is divided between internal and external states with blanket states between them. Blanket states are the points of contact between internal and external states, and are comprised of sensory states through which external states influence internal states, and active states through which internal states influence external states. The Markov blanket itself, thus strictly seen comprises only those blanket states between internal and external states. Further, in this formalism, the internal and external states are statistically *conditionally independent* (Friston, 2013). That is to say that, given (conditional on) the blanket states, knowing more about internal states does not offer further insights into external states and vice versa. The sensory and active states are *conditionally dependent* on the external and internal states respectively. This means that the sensory states are dependent on the external states, and the active states are dependent on the internal states. In the context of the organism, the internal states map onto the biophysical organism, the external states map

onto its environment, and the active and sensory states map onto action and perception respectively.

Intuitively, the environment, the external states, have a much wider range of variability than the internal states. This is because an organism can be in the ocean, on top of a volcano, deep underground in a cave or in the middle of an urban city, yet the organism itself, the internal states, remain semi-stable. This means that inferring the external states directly is computationally intractable (Friston, 2010). Yet due to the conditional dependencies, one only needs to know the internal states and sensory states to be able to calculate an *approximate* prediction concerning the external states. For active inference, such an approximate inference conditioned on action will suggest a pathway that minimizes free energy or surprise. Note that when the organism moves in accordance with active inference, this constitutes an influence on the external states via active states, by the internal states. For perceptual inference the same knowledge is needed, yet here the direction of influence is the other way around: the external states impose a particular influence on the internal states, *via* the sensory states.

Generative model is the term used for the probability distribution of the internal states (Friston, 2012). In the context of an organism these are the states of the organism itself. It is these states that are essential to have access to in applying both active and perceptual inference. As we apply both active and perceptual inference in order to keep the modeled organism in viable states, minimizing its free energy, the generative model will recapitulate, *adapt to*, or covary with the structure of the external states, also called the *generative process* (Bruineberg et al., 2018, Ramstead, Kirchhoff, Friston, 2019; Kirchhoff and Kiverstein, 2019). This relation with the generative process unfolds on ontogenetic and phylogenetic timescales. Consider how a fish's phenotype (the gills, a scaled body) shows covariation with the water it lives in, and how organisms during their lifetimes familiarize themselves with and adapt to new environments. This means that, from an outsider's perspective, the internal states hold *predictive value* over the external states, and in this sense we can say that, from an outsider's perspective, the internal states can represent the external states (Hesp et al., 2019, p. 10, 26; Friston, 2012).

5.2.2 Living models

So far I have approached the FEP strictly as invoked in practices of scientific modeling. Standardly, however, these models identified in FEP theories are thought to be employed by, instantiated or encoded in living systems one way or another, or even in multiple ways on multiple hierarchical scales (Kirchhoff and Kiverstein, 2019; Hesp et al., 2019; Ramstead,

Badcock and Friston, 2018, 2019; Clark, 2017). Roughly, there is a neurocentric and an embodied variant. The former corresponds to *predictive processing* and is usually explained in terms of the brain *having* a model (Hohwy, 2013; Clark, 2016). The latter is influenced by enactive proposals, typically under the banner of FEP (though not always, see Kirchhoff and Robertson, 2018, Kirchhoff and Kiverstein, 2019) and is usually explained in terms of the system *embodying* a model, or *simply being* a model.

In neurocentric predictive processing, or predictive processing, the brain is thought to encode the generative model—the heart of the predictive machinery—explicitly. In vision, for example, the brain employs the model to resolve the underdetermination problem in perception (Hohwy, 2013).[42] Also known as the poverty of the stimulus problem, this derives from the consideration that any particular retinal image is coherent with an infinitude of possible actual external worlds. Yet, on a daily basis, we perceive a constant world: how does the brain achieve this? Neurocentric predictive processing's answer is that the brain is a Bayesian prediction machine that attempts to anticipate the incoming signals (Hohwy, 2013; Clark, 2013; Clark, 2016). The generative model is thought to be a representational recapitulation of the causal-probabilistic structure of the external world, which will guide the predictions (Gładziejewski, 2016). As explained in Section 2.2, an exact prediction of the external states is computationally intractable. This is why, according to predictive processing, the brain engages in *approximate inference*. The inference using internal and sensory states, which the brain has access to, is thought to be computationally tractable. The probabilistic inference the brain engages in, thus *approximates* the direct inference of the external states. Whenever the brain doesn't get it right, prediction errors occur. The brain's imperative is to minimize such errors by engaging in either perceptual or active inference (Hohwy, 2013; Clark, 2013, 2016). A familiar example elucidates this. If I, say, walk into my office and predict the presence of a cup of tea that is not actually present in the stimulus entering my system, there are two ways to respond. I can apply perceptual inference and update the generative model of the world to not predict a cup of tea there anymore. I can also apply active inference, a prediction conditional on movement, and

[42] Note that the underdetermination problem depends on how the situation is set up. When we take a single snapshot of the world, there is strong underdetermination. In an ecological situation, however, we do not take snapshots of the world, but action and perception unfold simultaneously over time. If one approaches perception as a doing (thus with duration, extended through time), the poverty of the stimulus disappears (Myin, 2016; Gallagher, 2017). If one approaches perception not like a film as a sequence of snapshots, but instead as a diachronically constituted process with duration, our attunement to sensorimotor contingencies shaped by a phylo- and ontogenetic interactional history, much of the uncertainty of the external world dissipates (O'Regan and Noë, 2001; Kirchhoff, 2015; Di Paolo et al., 2017). Indeed, if we note that we do not passively take in the world, but actively explore it, the poverty of the stimulus ceases to exist as such. Exploration of this idea is outside the scope of this paper, but the references in here offer accessible introductions.

move about to see if the cup was occluded by other objects or even go to the kitchen to get myself a cup of tea. Interestingly, the *free energy* mentioned in Section 2.1 is formally equivalent to prediction error in the long run in predictive processing (Friston and Stephan, 2007; Hohwy, 2016). Under the FEP, free energy needs to be minimized, and under predictive processing, the formally equivalent value of prediction error needs to be minimized. This is, roughly, why predictive processing is considered a neural implementation of the free energy principle.

According to the neurocentric approach to predictive processing, "prediction error minimization is the only principle for the activity of the brain" (Hohwy, 2016, p. 260). Any neural activity, the thought is, conforms to the principle of prediction error (or free energy) minimization. Yet we figure in many more situations than looking for missing cups of tea, and often these situations have multiple layers of complexity. Consider a simple activity such as walking to university, which entails the longer term aspect of maintaining a job, but also getting to work today, manoeuvring the current traffic situation, dodging potholes or uneven roads and even lifting and putting down legs one by one to walk. All of this is thought to be controlled and computed by the brain. The brain's prediction machinery is thought to be hierarchically layered, mapping onto neurophysiological cortical layers (Kiebel et al., 2008, 2010. These layers deal with different levels of complexity, ranging from colours, surfaces and edges, objects, to objects and, further, objects-in-context (Hohwy, 2013; de Bruin and Michael, 2017).

Further, the blanket states are considered to appear at the ends of the nervous center, so that the extra-neural body also figures as part of the external states (Hohwy, 2016). This means that the brain not only attempts to attenuate the external world best as possible, but also the body it is in. Inherited through evolution, there are certain predictions that remain constant: those that pertain to the very essential states of the system to maintain homeostasis (Clark, 2013; Friston, 2010; Friston et al., 2009). Cases like hunger, then, appear as prediction errors to be alleviated by applying active inference, a prediction conditional on the action of, say, preparing a meal. In this way, predictive processing intends prediction error minimization to cover all bases.

In what I have called the embodied approach, the organism is thought at least to embody the model in its dynamics, and the model is (typically) considered to be non-representational (Kirchhoff and Robertson, 2018; Bruineberg et al., 2016). According to this view, the organism does not *have* a model, it *is* a model (Kirchhoff, Parr, Palacios, Friston, Kiverstein, 2018, p. 4; Friston, 2013). Ramstead and colleagues phrase it as follows:

> generative models are *not explicitly encoded* by physical states. That is, they are *not encoded by states of the brain*. Rather, it is the adaptive behaviour of the system that implements or instantiates a generative model … adaptive behaviour brings forth the conditional dependences [sic] captured by the generative model, that is, keeping the organism within its phenotypic, characteristic states. (Ramstead, Kirchhoff, Friston, 2019, p. 7, emphases in original)

This means that, as opposed to being encoded in neural activity, the model is implicit in the adaptive behaviour of the agent. Initially, this makes the generative model sound epiphenomenal. It seems that what actually does the work is the adaptive behaviour and the conditional dependencies, and that these relations can merely be *captured* in a generative model. This seems to imply that the model is merely a scientific construct that *captures* real statistical relations in the world. Yet the organism is also thought to "leverage" the generative model together with the *recognition density* (which, roughly, is a measure of the divergence between the prediction and the actual *sensory state* encountered by the system) (Ramstead, Kirchhoff and Friston, 2019, p. 7; Friston, 2012)). This means that the organism uses a calculated result (the recognition density) to minimize its free energy. In some sense these probabilistic densities must thus be accessible to the organism. This seems to imply that the generative model is thus *not* epiphenomenal or a mere scientific construct. This reading is further consistent with the claim that it has a 'causal bite' by playing a vital role in action policy selection (Ramstead, Kirchhoff, Friston, 2019, p. 9).

There are thus two readings open to the embodied approach, both of which are consistent with at least some of the writing: a model-instrumentalist and a model-realist reading (Ramstead, Kirchhoff, Friston, 2019; Kirchhoff, Parr, Palacios, Friston, Kiverstein, 2017; Bruineberg et al., 2016). I will discuss this more elaborately in Section 3.1, but roughly it is as follows. The two readings map onto two different interpretations for what it means to *embody* or *simply be* a model, as well as for *approximate inference*. In the *model-instrumentalist* reading, the embodied model is a scientific construct. In the adaptive behavior of the organism, particular statistical or covariational relations appear between the internal and the external states (e.g. a fish's gills on an evolutionary scale, a hand's covariation with the shape of the door knob on the scale of organismic activity; Ramstead, Kirchhoff, Friston, 2019, p. 7). These statistical relations are *real* as the relations between the rings of a tree and the years it's been alive are real. The model that these statistical relations 'embody' and captures them is a scientific construct. Approximate inference here means that the organism behaves so that the statistical relations that are brought forth can be *cast* as conforming to the norms of probabilistic inference. In this sense the organism does not engage in *any* form of inference itself, but it

behaves so that the probabilistic relations *approximate* probabilistic, computational inferences. This reading will be defended in this paper.

In the *model-realist* reading, the model embodied by the organism exists independent of our scientific modeling practices. The organism *literally is* a model in an objective sense. Under this reading, the organism can 'leverage' particular computational results from the model that it embodies in navigating the environment, without conscious access (Ramstead, Kirchhoff and Friston, 2019, p. 7; Hesp et al., 2019). Approximate inference here is much the same as it is in predictive processing: directly inferring the external states is intractable, so the organism *approximates* this inference by way of computing over the internal and sensory states instead. In this reading, thus, the organism does engage in inference itself. However, the model that is leveraged is embodied by the organism, not encoded by the brain.[43]

There is a third route that takes both options, in which

> the brain *does not just contain* a hierarchical generative model of the world, its dynamics *also instantiate* one – its form and function reflect a physical transcription of causal regularities in the environment that has been optimised by evolution within and across nested spatiotemporal scales. (Badcock, Friston, Ramstead, 2019, p. 6, emphases added)

In addition, "different organisms instantiate unique 'embodied models' of their specific biological needs and eco-niches" (Badcock, Friston, Ramstead, 2019, p. 7; referencing Ramstead, Badcock, Friston, 2018, Friston, 2011, Allen and Friston, 2016, Gallagher and Allen, 2017). In short, this is a representational interpretation that takes a generative model to be *encoded* in the brain like neurocentric predictive processing as we have seen above, and *instantiated* both in the brain and in the body conform the embodied approach (Badcock, Friston, Ramstead, 2019, p. 7).

An interesting result that has come from this modeling practice is that, by applying free energy minimization to a simulated primordial soup, self-organizing patterns appear naturally (Friston, 2012). This can hint at the broad dynamics that must've been in place for the actual ontogeny of life, and is an interesting model of what a minimal life form could require. The mathematical model is also applicable to single cells, organs, organisms and even extends into

[43] It is not clear *how* the organism infers on the basis of the model that it embodies. In neurocentric predictive processing, one can imagine the brain, cast as an agent, manipulating and making inferences on the basis of a statistical model, analogous to the way we do in our modeling practices. There's a, fairly obvious, mereological fallacy here (see Hohwy, 2016), and a few other issues pertaining to invoking representations independent of human practices reappear (Tonneau, 2012; Hutto and Myin, 2013). Nonetheless, there is a clear *modus operandi* for the exploitation of the model. In the embodied sense, it is much less clear. What agent engages in inference? What does it mean for the organism to *implicitly* infer the external states over and above the organism's behavior to instantiate particular probabilistic relations that *can be captured* in a model?

social situations, cultures, niche construction and evolution. FEP is typically said to unify theories of brain activity (Friston, 2010), but is now also argued to unify studies of life at multiple spatiotemporal scales (Hesp et al., 2019, Ramstead et al., 2018, Ramstead et al., 2019; Bruineberg et al., 2018). Under the FEP, a single modeling practice seems capable of capturing the dynamics of how the brain makes sense of the world as it does in predictive processing, how the organism copes with its environment as seen in the embodied approach, and so on. Indeed, Friston considers it to be "a theory of every 'thing'" (Friston, Under Consideration). This supposed unifying ability is a central attractor for the FEP.

5.3 A Model of Life and Life's Model

The wide range of applicability of the Markov blanket formalism is a double-edged sword. On one hand, for many theorists a unified 'theory of everything' is appealing. On the other hand, it requires extra care. We may be able to model any two distinct phenomena using the same tools, yet this does not mean we can extrapolate findings in one application to the other. Nonetheless, I suggest that this may be happening in the FEP literature carelessly. In this section, I shall first show that there are two distinct applications of the FEP model. One use is consistent with an instrumentalist position on the model in which the model is a scientific tool used to study a system. The other use is strictly compatible with a realist position in which the model is implemented so that the system under scrutiny actively uses or literally embodies the model to make statistical inferences. Second, I will show that these two are often conflated, leading to certain invalid claims that overstate the FEP's accomplishments.

5.3.1 Model entanglement in FEP

In Section 2, I separated scientific modeling practices from ascriptions to organisms and/or brains. In the FEP literature, these two notions typically seem to co-exist peacefully, and intertwine naturally as though one is a seamless continuation of the other. I suggest that these two are to be distinguished by the agent involved, the modeler. In one take on modeling, the scientist models a self-organizing system (and its environment) in order to study it by way of a surrogate. This is the scientist's model, and is consistent with instrumentalism. In another take on modeling, the self-organizing system models its environment or 'simply is' a model of its environment, and exploits this model in order to navigate the world, and maintain its dynamics and physical integrity. This is compatible only with a realist position on models. Another way to pick them apart is that one is a model *of* life: a scientist's construal of relevant

relations of the target system, and the other is a model *used by* life: purportedly a statistical model the organism uses unconsciously to navigate the world, life as a modeler. In this section, I will argue that, in the FEP literature, there is a conflation of models as they appear in science and models as ascribed to organisms.

The scientist's model often plays an important role in FEP literature. One may read that the models in FEP are "*representations* of dynamical systems", and "may provide a *metaphor* for behaviour with different timescales and biological substrates" (Friston, 2013, p. 1, emphases added); that is, rather than an objective part of nature, the model is human-made: "an information-theoretic *construct*" (Constant et al., 2018, p. 5, emphasis added). Further, what makes it interesting, is that "it connects probabilistic *descriptions* of the states occupied by biological systems to probabilistic modelling or inference as described by Bayesian probability and information theory" (Friston, 2012, p. 2101, emphasis added; Korbak, 2019, p. 3).[44]

Representations or metaphors of a system are typically not to be taken literally: a portrait of your colleague is distinct from your actual colleague and the paper the portrait is made out of does not constrain the material your actual colleague is made out of (see also Di Paolo and Thompson, 2014; Tonneau, 2012). When Friston states that probabilistic descriptions connect to 'Bayesian probability and information theory', we do not seem to depart from the level of description. Moreover, both Bayesian probability and information theory are human-made theoretical constructs that are deeply embedded in multiple interrelated practices such as scientific, mathematical, probability-theoretic, and modeling practices. Each of these practices have evolved in intersubjective engagement and require teaching and practice for participation: they were only formed in very specific social contexts. Such formulations seem to indicate that FEP models are *descriptive* of actual biological dynamics, just like meteorological models of actual weather dynamics are used to make statistical inferences about future states of the weather.

Yet, as I have alluded to in Section 2, there are several ways that the mathematical machinery of FEP is ascribed to the organism under scrutiny. The representationalist, neurocentric approach lays their cards out on the table most clearly. Friston's earlier work is focused on representational learning, and explicitly is aimed at unearthing the model that the brain encodes for this (Friston, 2002, 2003, 2005). Later too, Friston writes that "an agent *must have* an implicit generative model of how causes conspire to produce sensory data" (2010, p.

[44] There are plenty more of such phrasings in the FEP literature. Friston writes that "any system that exists will *appear* to minimize free energy" (Friston, 2013, p. 11). FEP's success makes it an "important *metaphor* for neuronal processing in the brain" (Friston, 2012, p. 2101, emphasis added).

129). More specifically, he discusses "the form of the generative model and how it *manifests* in the brain" (*Ibid*). This indicates a realist take on models as encoded, exploited and manipulated in the neural architecture. This is also the approach taken up by Hohwy (2013, 2016), Gładziejewski (2016), and Clark (2013, 2016) most notably, and has seen widespread further influence. Their position is, roughly, that the brain encodes a generative model of the extra-neural world that, in some sense, recapitulates the causal-probabilistic structure of the world so as to maintain homeostasis (via active inference) and infer the causes of sensory inputs (via perceptual inference).

There is also an embodied approach to the FEP. Where the neurocentric approach clearly commits to a realist position, we have seen in Section 2.2 that the embodied approach remains ambiguous, mirroring Friston's pioneering writing (Friston, 2012, 2013). This is conspicuously expressed in the descriptive language employed by FEP theorists. The embodied model finds a basis in the claim that "biological systems can distil structural regularities from environmental fluctuations (...) and embody them in their form and internal dynamics" (Friston, 2012, p. 2101). This is contrasted with the notion that an organism *has* a model of its environment. Kirchhoff (2018) writes:

> by 'model' it does not follow that an organism has an internal, representational model of its niche and that it is this model that does all the cognitive work (if you like). Instead, an organism *is* a model, viz., the causal and statistical regularities reflected in some environment are reflected in some phenotype, i.e., model. (p. 761; see also Kirchhoff et al., 2018, p. 4).

An example of this is "the physiological make-up of a fish, say, as a model of the fluid dynamics and other elements that constitute its aquatic environment—its internal dynamics depend on the dynamics of the niche" (Kirchhoff et al., 2018, p. 4; Bruineberg et al., 2016).

Taken in this sense, in the embodied approach, the FEP seems like a *definition* of life, or more specifically, "a mathematical *formulation* of how adaptive systems (that is, biological agents, like animals or brains) resist a natural tendency to disorder" (Friston, 2010, p. 1, emphasis added), so that "any system that avoids surprising exchanges with the world (i.e., surprising sensory states) *will look as if* it is predicting, tracking, and minimising a quantity called variational free energy, on average and over time" (Ramstead, Kirchhoff, Friston, 2019, p. 4). Implied, but not explicit, is that the system *does not actually* predict, infer, track or minimise a quantity called variational free energy, but *merely looks as if*. The probabilistic model merely tracks certain real statistical relations in the organism-environment system. In this sense the organism is thought to *approximate* inference. It seems again that, with regards to the status of the model, we take the scientist's perspective. When we take a fish's phenotype

as an example, and we look at the shape of the fish, the scales and the gills, they can, for an external observer, represent or hold predictive value with regards to the water it lives in. We, as external observers, can then make inferences on the basis of what are here marked as the internal states (the phenotype) about the external states (the environmental niche), and in this sense we can use the fish's physiology as a model for the external states.[45]

The instrumentalist reading seems encouraged as more realist-leaning terminology is often accompanied by 'scare quotes' or termed *implicit*, implying it should be read in a non-standard way. In Bruineberg et al.'s (2018) model of niche construction under the FEP we read that the "effect of the agent on the environment can be understood as the environment 'learning' about the agent", and in this sense "the agent and the environment 'get to know each other'" (Bruineberg et al., 2018, p. 162). In Badcock, Ramstead and Friston's (2019) attempt at a "free-energy formulation of the human psyche" we read that "brain dynamics (i.e., the general 'behaviour' or 'ensemble dynamics' of neural mechanisms) can be described as realising an implicit hierarchical generative model: a Bayesian hierarchy of 'hypotheses' or 'best guesses' about the hidden causes of our sensory states" (p. 5).

Recall that, according to Ramstead, Kirchhoff and Friston, "it is the adaptive behaviour of the system that implements or instantiates a generative model … adaptive behaviour brings forth the conditional dependences [sic] captured by the generative model" (2019, p. 7). The model is thus "realised" through a specific sort of action of the organism (p. 9). This could be taken to mean that what is *real* is adaptive behaviour of the organism, and the particular statistical relations that exist between the organism and its environment that can be *described* as a generative model. In this sense, when the organism "leverages"' the recognition density (a measure of the match between the prediction of the generative model and the actually encountered input), perhaps what the organism exploits there is *not* the recognition density itself, but merely the conditional dependencies between the organism and its environment. These conditional dependencies could be thought of as covariational in nature (Bruineberg and Rietveld, 2014). In practice, this happens when I exploit the covariation of the shape of my hand and the door knob in opening the door, or the structural similarities between a cardboard box and a table top when I place the former on the latter. As such, the embodied approach seems to afford an instrumentalist reading.

[45] As an anonymous reviewer pointed out, it is important that this is not what models in the FEP are used for in science. Instead, they are descriptive models that allow for a mathematical formulation of the organism-environment dynamics, but they do not afford new predictions, *per se*. This is only meant to point out the manner in which talk of inference makes sense in the FEP framework.

Hesp et al. (2019) also emphasize the *implicit* status of the model in FEP, they argue "it means that these Bayesian concepts do not require the system itself to be "conscious" of inferences in any way or that these inferences need to be "explicit" and couched in propositional or linguistic terms" (p. 231). "Implicit", thus minimally means that no conscious access is required, and that the inferences made need not be propositional or linguistic in nature. This purely negative description of what implicit inference means rules out the caricature of an organism consciously engaging in advanced statistical computations whenever it moves about. This is consistent with an instrumentalist reading in which the organism does not engage in inference at all, it merely enacts certain statistical relations that can be captured in a model. Yet it is also consistent with a predictive processing reading in which the organism, subpersonally and unconsciously, infers the world in strictly probabilistic, non-propositional and non-linguistic terms (Wiese, 2017).

Indeed, it is also said that the organism directly uses (or "leverages") the generative model that it embodies, or conversely, that the generative model actively controls the organism's coupling to the environment. For example, Bruineberg et al. assume "the agent *uses its generative model*" (2018, p. 164, emphasis added), and the "generative model functions to regulate and control the agent's coupling to the environment" (Kirchhoff and Kiverstein, 2019, p. 59). More explicitly, Kirchhoff and Kiverstein state that "*generative models* coupled in active inference to generative processes are uniquely equipped to *make use of* the properties of an organism's embodiment and associated species-typical environment" (2019, p. 59). Here the generative model *makes use* of the properties of an organism's embodiment and associated environment. As such, it cannot simply be that the generative model is merely a "mathematical formulation" of a complex interplay between an organism and its environmental niche. It is seen as separate from the organism's embodiment, insofar it can *make use of* properties thereof. Indeed, in each of these citations, it seems the generative model is ascribed active force in the causal web of an organism: it has 'causal bite' (Ramstead, Kirchhoff and Friston, 2019). Under an instrumentalist reading, a model cannot have a direct *causal bite*: a description of a particular behaviour does not feature in causing the target behaviour (barring self-referential descriptions). If it were merely a description, it would describe features of and (statistical) relations between the system and its environment (or the agent-environment system) that have causal bite. The model, on itself, remains on a descriptive level.

Despite their theoretical differences, as far as the status of models in FEP is concerned, the embodied approach in which the organism simply is a model is thus equivalent to neurocentric predictive processing. In both approaches, the model exists independent of human

practice, computational results are either encoded and manipulated, or embodied and leveraged. This is also seen in the manner in which inference can be said to be approximated. Contrary to a model-instrumentalist reading, approximate inference in both neurocentric predictive processing and a model-realist embodied approach is cashed out in terms of an inference over internal and sensory states, as opposed to a direct inference of the external states (which is intractable).

For both neurocentric and embodied approaches to the FEP framework, it thus seems that there is a conflation of models that are created in a scientific endeavour and a model that is used, leveraged or employed by, or instantiated in (the dynamics of) an organism. With no sense of direction, FEP theorists seem to veer from scientific models of life to living models.

5.3.2 The error in conflation

"So what?" one may think. There are two possible modes of interpretation, and one may, for various reasons, prefer one over the other. The issue is that these two interpretations are often muddled and few take care to distinguish the two (Colombo and Wright, 2018). Consider how Friston (2013) sums up exactly why free energy minimization and inferentialism are central to life:

> [1] Under ergodic assumptions, the long-term average of surprise is entropy. [2] This means that minimizing free energy—through selectively sampling sensory input—places an upper bound on the entropy or dispersion of sensory states. [3] This enables biological systems to resist the second law of thermodynamics—or more exactly the fluctuation theorem that applies to open systems far from equilibrium. [4] However, because negative surprise is also Bayesian model evidence, systems that minimize free energy also maximize a lower bound on the evidence for an implicit model of how their sensory samples were generated. [5] In statistics and machine learning, this is known as approximate Bayesian inference and provides a normative theory for the Bayesian brain hypothesis. [6] In short, biological systems act on the world to place an upper bound on the dispersion of their sensed states, while using those sensations to infer external states of the world. (Friston, 2013, p. 2, numbering added)

Let us dissect this quote sentence per sentence. In [1], Friston points out an equivalence in a formal model of long-term average of surprise and entropy under specific assumptions . In [2], we see a formal implication of [1]. In [3], the terms in the model are linked to their target system: biological systems. In [4], the terms in the model (average of surprise and entropy) are also linked to Bayesian model evidence. Here he thus introduces a further formal equivalence. So far, Friston stated a mathematical formulation of organisms' resistance to the second law of thermodynamics and offered formal equivalences of terms in that formulation. Yet the line gets

blurred when Friston introduces "the implicit model of how the organism's sensory samples were generated" in [4]. The aforementioned model is now 'implicitly' implemented in the system. In [5-6] this quickly gets expanded with further equivalences to entail that "biological systems … us[e] those sensations to infer external states of the world" (emphasis added). The organism here uses the model to make statistical inferences about the external states of the world. Simplified a little, it seems that the argument in the above quote can be put as follows.

1) Biological systems φ with *X*.
2) Formally, *X* is represented by *x*.
3) Formally, *x* is equivalent to *y*.
4) Formally, *y* is equivalent to *z*.
5) Thus, biological systems φ with *z*.

The issue is, I argue, that a formal representational relation between two systems is conflated with one of natural identity. Put differently, the equivalence of term *x* that represents target feature *X* in the world, with a different term *y* does *not* warrant the interchangeability of target feature *X* with *y*. *X* concerns a real world feature, whereas *x* and *y* are both *mathematical descriptions* of such a real life feature. What it *does* warrant, however, is the following. If *X* can be mathematically described as *x*, and *x* is equivalent with *y* and *z*, then *X* can be mathematically described as either *x*, *y*, or *z*. *Within* the realm of mathematical description, these terms are equivalent. In the above quote by Friston (2013, p. 2), there is a hidden argument in which a mathematical formulation of behaviour is turned into an underlying mechanism for that behaviour.

This is unjustified, yet gives off the pretense that the battle on how to interpret modeling under FEP is already won, hidden behind complex mathematical and statistical jargon. It suggests that for a biological system to model and infer the external states of the world is a tautology (Friston, 2013; Friston and Buzsáki, 2016). We see this being picked up in the literature too. Constant, Bervoets et al. (2018) for example, say: "To be a living system then *means* dynamically modeling oneself in relation to one's body, and one's environment" (Constant, Bervoets et al., 2018, p. 3). To a lesser extent, Hohwy (2016) makes a similar claim. In this, the representational relation between internal and external states within the Markov blanket formalism is reified so that the internal states in themselves represent the external states, irrespective of there being an external observer. This is taken to hold for the target system as a matter of course once one accepts the basic description of life under the FEP (Hohwy, 2016).

The above discussion of model-realist versus model-instrumentalist readings of the embodied approach in FEP also seems indicative of the issue. In one reading, there is no conflation: there is a realist attitude with regards to the statistical relations enacted by an organism's adaptive behavior, and an instrumentalist attitude with regards to the scientific model that captures these relations in a mathematical formalism. Approximate inference is understood in terms of the relations brought forth by the behavior approximating inference on the basis of the model as we could do as modelers. Yet in another reading of the same literature, there is a conflation. Simply being alive, then, means that the organism will model its environment because of the statistical relations that it is accompanied by. The existence of statistical relations, as we have seen, does not warrant the (science-independent) existence of a model in itself: the rings in a tree trunk covary with the years the tree's been alive, but this does not entail that the rings model the environment in any sense independent of human practice. This conflation seems to permeate the very basis of the FEP approach to cognition. As such, it is imperative to distinguish models *of* a target system from models *used by* a target system.

5.4 Anti-realism: against the life's model interpretation

I have argued that in the FEP literature, two approaches to models, instrumentalist and realist, are present without being properly distinguished. Blurring the lines between these two takes on models can cause theoretical mishaps. Here I will argue that, because of these errors, only the instrumentalist approach to FEP models is warranted.

5.4.1 Overgeneration of models

The applicability of Markov blankets extends far beyond only biological systems. Friston even says that "*any system that exists* will (...) engage in active inference" (2013, p. 2). Though this formulation seems to draw no boundary at all, even weaker ways of understanding FEP related notions such as Markov blankets have very broad applicability. At least every single biological system has a Markov blanket, ranging from single cells to macroscopic organisms (Kirchhoff et al., 2018). Certain non-living systems such as Huygens pendulums also have a Markov blanket, and engage in active inference (Friston, 2013; Kirchhoff et al., 2018). "[C]arefully chosen nodes of the World Wide Web surrounding a particular province" may constitute a Markov blanket (Ramstead et al., 2018). The Markov blanket formalism is also applicable to a process called 'niche construction', according to which organisms and their environments co-evolve and mutually influence each other. Here, the organism models the environment, and the

environment models the organism, because, from the environment's perspective, the organism is external (Constant et al., 2018). Ramstead et al. (2016) further suggest it is applicable to human cultures and may be able to elucidate how cultural practices take shape due to social free energy minimization.

This wide range of applicability of the formalism constrains the range of applicability of a realist approach. Though we may be able to imagine a statistical model implemented in a neural architecture, it is harder to imagine a bacterium subpersonally engaging in advanced statistics (Hohwy, 2013, 2016). Stranger still is when 'the environment' is thought to model the organism. It is unclear what the physiological substrate of this model could be, as the 'environment' is composed of a wide variety of physiologically diverse systems such as different and different sorts of trees, plants, animals, etc. This would require a sort of hivemind of systems that may have a hard time communicating. The 'environment' is also defined relative to a particular agent, and, as the agent moves, the constituents of the 'environment' are also in constant flux. The World Wide Web, or human cultural evolution all do not easily afford the autonomy of *implementing* models and *making statistical inferences*.

The overgeneration of Markov blankets is not new in the literature. A similar issue is dealt with in Kirchhoff et al. (2018), in which an extra property is suggested to be added to distinguish non-autonomous active inference systems from autonomous active inference systems. Huygens pendulums, for example, engage in '*mere* active inference', whereas organisms engage in '*adaptive* active inference'. The distinction between the two is marked by whether one is "entirely 'enslaved' by its here-and-now—and, in particular, its precedents" or not, the latter entails that one is capable of "modulation of [one's] sensorimotor coupling to [one's] environment (Kirchhoff et al., 2018, p. 5). This is thought to constrain the ascription of autonomy according to FEP to systems that we would actually consider autonomous. Certainly, Huygens pendulums are cleared, but one may wonder whether an organism's environment, or human cultural evolution are only 'enslaved' by their here-and-now and their precedents any more than single living organisms. It seems they may allow a stronger sense of novelty or adaptability to their perspectivally determined external states than a Huygens pendulum. There is also the question to what extent an autonomous system is *not* 'enslaved' to its here-and-now and its precedents. We may be able to modulate our sensorimotor coupling to the environment, but one may wonder to what extent this is done in a way that is *not* enslaved by our environment and our interactional history (our precedents) with that environment. There is surely a sense in which our environment, our phylo- and ontogenetic interactional histories can be said to simply *determine* our sensorimotor coupling with that environment. Consider the purported open-

endedness in minimizing the prediction error in me expecting a cup of tea on my desk. Whether I will either 'update my model' and accept the cup isn't there (perceptual inference) or get myself a cup of tea (active inference) will depend on my current state and my interactional history. Whether I am thirsty or not, but also whether I, through my previous interactions, have learned that a cup of tea can be conducive to a productive afternoon and how to make a cup of tea, are essential in the determination of my actions. Indeed, when considering the whole organism-environment system, the issue may be less open-ended.

Moreover, though this novel distinction between *mere* and *adaptive* active inference touches on a *similar* issue, it does not affect overgeneration of modeling in itself. Indeed, the Markov blanket formalism is still applicable to Huygens pendulums, environments and cultural practices, even if we do not grant them autonomy, and they are still said to be modelers of their external states. This is an issue because for none of these things it is clear how they could implement a model, nor make statistical inferences. Proposing a solution by suggesting we take an instrumentalist approach to non-autonomous systems and a realist approach to autonomous systems, is *ad hoc*. It would require additional argumentation or a principled reason to show that of all systems that can be modeled in a particular way, only those that are also autonomous should be described as implementing and using that model, whereas the others merely *act as if*.

5.4.2 Covariation, no content, no model?

Let us entertain the possibility that the mathematical construct is implemented biologically, either neurally by having a model, or organismically by simply being a model, independent of modeling practices. It is still open whether the organism actually uses, or embodies and leverages, a *model*. In this section I will argue that a realist position on the *mathematical mechanisms* described by the FEP still does not warrant a realist position on the *models* as used by FEP theorists. This relates to the FEP debate on representations. The primary argument will be roughly as follows. In Section 4.2.1, I discuss that models in FEP as thought to be implemented in organisms are not representational, but covariational in nature (Kirchhoff and Robertson, 2018; Ramstead, Kirchhoff, Friston, 2019; Bruineberg and Rietveld, 2014; Bruineberg et al., 2016). Continuing, in Section 4.2.2 I consider what it takes to be a model. The literature on the epistemic and ontological grounds of models in science may comprise a wide variety of positions, but most agree that they are representational in some way, shape or form (Frigg and Hartman, 2018). This means that, in virtue of being non-representational when used by the organism, the mathematical machinery as thought to be instantiated in the

organism's dynamics cannot be a model. Recently, a non-representational position on scientific models has been proposed that could, *prima facie*, be considered a solution for the model-realist FEP position (de Oliveira, 2018). Yet this is to no avail, because the pragmatist appeal to human practices de Oliveira (2018) relies on is unavailable to FEP model-realists. As such, even if we entertain the possibility that the mathematical constructs are implemented biologically, these constructs are not models.

5.4.2.1 Covariation, no content

It will be fruitful to first describe Kirchhoff and Robertson's argument for a non-representationalist interpretation of FEP. Recall that according to FEP, an organism's primary imperative is to minimize its free energy. In predictive processing, the neural implementation of the FEP, this means that the brain continuously attempts to match its internally generated predictions with the signals that enter the system. To call this representational, it is crucial to show that *misrepresentation* is possible. Kiefer and Hohwy (2017) have attempted to do this in reference to the Kullback-Leibler divergence (see also Friston, 2013). Roughly, they argue that this divergence is a measure of the difference between the brain's estimate and the incoming signal.[46] This, they argue, means that it measures misrepresentation.

However, as Kirchhoff and Robertson (2018) argue, the Kullback-Leibler divergence only measures a *Shannon-informational* divergence. Shannon information is covariational information, in itself not representational (Godfrey-Smith and Sterelny, 2016). This means, roughly, that two particular systems co-vary: when one system changes, it is likely that the other changes in a similar fashion, broadly construed. In this sense, if a scientist knows that system A and system B covary, then knowledge of one system's state increases the information they have about the other system's state. A typical example of covariance is the rings on the trunk of the three that reliably covaries with the amount of years it's lived. Covariance is not necessarily a one-to-one relation, and systems typically covary more or less strongly. In this particular sense, Kirchhoff and Robertson argue that the internal states of the organism covary with the external states: reliably, but not per se one-to-one (2018). After all, we do make mistakes regularly and this betrays misalignment. This misalignment, however, is a form of negative covariance, not one of misrepresentation. As such, for system A to 'model' or 'infer'

[46] Technically, the Kullback-Leibler divergence measures the difference between the brain's recognition model and the actual posterior probability (Kiefer and Hohwy, 2017; Friston, 2013). This means that it measures the divergence of the brain's approximation of the prior probability and the "true state of affairs" (Kiefer and Hohwy, 2017, p. 23). See Friston (2013) and Kiefer and Hohwy (2017), and Kirchhoff and Robertson (2018) for a more technical discussion on the matter.

system B, the dynamics of system A must covary reliably with the dynamics of system B (Bruineberg and Rietveld, 2014, p. 7; Kirchhoff, 2018, p. 762). This relation is thus not representational. Both modeling and inference are captured by dynamical covariation (Friston, 2013).[47] Kirchhoff and Robertson (2018), unlike Bruineberg and Rietveld (2014) for example, do not argue to alter FEP by expanding it with a theory of affordances or fit. They merely show that the mechanisms proposed by FEP are, contrary to popular opinion, in themselves *not representational*.

Though the generative model is not representational, Ramstead, Kirchhoff and Friston emphasize it does use "exploitable structural similarities" (2019, p. 11). Such exploitable structural similarities lie at the foundation of the notion of structural representation popular in predictive processing (Gładziejewski, 2016). However, proponents concede mere exploitable structural similarity is insufficient to ground a notion of representation. Indeed, when I open a door, I exploit the structural similarities between the shape of my hand and the door knob. Yet my hand in no sense *represents* the knob. As such, the exploitable structural similarities do not allow for representations to be snuck back into the FEP.

5.4.2.2 No model?

What makes a model a model? Where do models get their epistemic import from? How can it be that from studying a model of a target system, we can learn more about the target system? These questions have been discussed elaborately (Frigg and Hartman, 2018). Despite a multiplicity of views, there is a broad consensus that models are inherently representational. The debate is, largely, about *how* models are representational. A few options that constitute this representational relation are isomorphism (van Fraassen 1980) or similarity (Giere 1988; see Frigg and Hartman, 2018 for an overview). The representational relation is then thought to be constituted solely by the extent of similarity the model displays to the target system. These views have been heavily criticized, and there is now a broadly shared consensus to include reference to *use*: a model is representational because it is used as such, which may include further reference to objective similarities between the model and the target system (de Oliveira, 2018, p. 12; see Suarez, 2003; van Fraassen, 2008; Giere, 2010 for examples). A more recent

[47] The equivalence of inference and covariation is also picked up in Badcock, Friston, Ramstead (2019). We read that "our actions will tend to infer or reflect the statistical structure of the environment to which they are coupled" (p. 6). Both "infer" and "reflect" are used as interchangeable, equivalent terms. Reflection here should be thought of in the way that the body of a fish reflects the environment it has evolved and lived in (Kirchhoff et al., 2018; Friston, 2013). This reflection, as we have seen here, is captured by a covariation relation between the organism and its environment, which in turn means that to infer is again seen as captured by a covariation relation.

non-representationalist account suggests that a model is a surrogate used for a target system within a particular scientific social practice of creating models "in terms of skill development and learning transfer" (de Oliveira, 2018, p. 23).

There are thus essentially two options: models are representational in some way, shape or form, or they are not, in which case their epistemic import is explained pragmatically by surrogacy and human practices of skill development and learning transfer. Above I have there is good reason to think that models in FEP are non-representational. That which is purported to measure misrepresentation, an essential feature of representation, instead measures negative Shannon-informational covariance. This means that, if we follow the broad consensus on models in science, 'models' as they are instantiated in organisms according to the FEP simply *do not count* as models. The non-representationalist view on models in science is of no help either. This has, roughly, two conditions: the purported model should 1) be used as a surrogate of a target system, and 2) be embedded in the scientific social practice of modeling (de Oliveira, 2018). Though the surrogacy condition could potentially be met, models in FEP cannot rely on human practices that are thought to be *products* of the social interaction that the models are thought to *precede*. As such, a realist position on the mathematical relations as described in FEP still does not warrant a realist position on these models: covariance is simply too weak a notion.

5.5 Model-instrumentalism, covariation-realism

Above I have argued that there are two issues with model-realism in FEP. The first is that the model is widely applicable, leading to over-generation of models and trivialization of the notion. The second is that the notion of modeling in FEP is not representational, but covariational in nature. Models as they are commonly understood in science, however, *are* representational. A non-representational approach exists, but to no avail for the FEP realist. De Oliveira's approach relies on a model's embeddedness in scientific social practices. FEP models are thought to figure at the very core of life itself: they should *precede* the sorts of social behaviour that form our practices. The mechanisms described by the FEP thus do not seem to fit the bill for a model anymore. One may now wonder whether this also affects the instrumentalist position on models in FEP. Perhaps, instead, we should not speak of models in FEP *tout court*.

I argue that this is not necessary. Overgeneration in itself is not an issue. As a modeler, one chooses to model a target system in a particular way. There being a multiplicity of potential

target systems that could be modeled using the same tools does not detract from the legitimacy of the model. Further, models as implemented in organisms in FEP cannot rely on a sociocultural practice of using models in science. An instrumentalist approach to modeling in FEP does not encounter the same obstacle. After all, FEP models are created by experts in mathematical modeling that have been trained in a scientific context. Indeed, it seems that models in the FEP as formed, used, and exploited by people in a scientific socio-cultural context are prime exemplars of models, whether we take a representationalist view or not. As such, the instrumentalist view quite clearly is free of aforementioned worries.

In Section 2.2 I discussed the embodied approach to the FEP, and in Section 3.1 I discussed more elaborately the extent to which the theory as proposed in the literature affords an instrumentalist or a realist reading. I argued that there is plenty of wiggle room, allowing for both a model-instrumentalist and a model-realist reading of the same literature. Relying on this wiggle room, we can see the broad contours of an instrumentalist approach. In this view, the *models* nor the *mathematical machinery* are taken to be *instantiated* in the organism. The adaptive behaviour, nor the relations between the organism and its environmental niche instantiate, encode or embody a generative model in any realist sense. There is *approximate* inference so that the organism's adaptive behaviour corresponds to probabilistic inference in a scientific model. Yet the organism does not engage in statistical inference in any realist sense. It is not for the organism, nor the brain, as it is for the scientist (contrary to Hohwy, 2016; 2013).

Instead, what in this view does warrant a *realist* position are the different statistical and covariational relations between the organism and its environment. This view takes a realist position on the notion that adaptive behaviour brings forth, or displays the conditional dependencies, that can then, by scientific modelers, be *captured* in a generative model (Ramstead, Kirchhoff, Friston, 2019, p. 7). This view also takes a realist approach to the notion, further, that we form increasingly tighter covariational relations with our environment both in a niche-construction, designer environment sense and in an active, ontogenetic adaptational sense, such as described in FEP terms in (Bruineberg et al., 2018; Clark, 2016, section 9.5; Constant et al., 2018; Hesp et al., 2019). The mathematical model created by Bruineberg et al. (2018) in particular is a good way to investigate minimal conditions for niche construction, as well as investigate certain statistical relations between the agent and the environment. That is to say, an instrumentalist view need not impose constraints on the FEP research programme itself. It does, however, impose constraints on the conclusions that can be drawn from the findings.

There are a few possible objections to constraining FEP models to an instrumentalist take. One possible objection is that models in FEP simply are quite unlike any model in science. Indeed, though scientific models are thought to be representational, these models are covariational, they are further thought to be implemented by organisms to minimize free energy or prediction error. System A can be said to model the dynamics of system B if the dynamics of A covary with the dynamics of B (Bruineberg and Rietveld, 2014; Kirchhoff and Robertson, 2018). This is simply a *special case* of modeling.

First, this does not match the consensus view on models as described in Section 4. Even a special case should be thought to be captured in a definition. If it is too special to be captured in the definition, this may be good reason to consider that, whatever it is, it may not be a model. Second, persisting it is a model results in trivialization of the notion, and it may not indicate to a reader what the author thinks it does. When we say that the organism *has* or *is* a model that the organism uses or leverages to make statistical inferences, it is difficult to imagine what this means if it is not contentful. According to Kirchhoff and Robertson (2018) and Bruineberg and Rietveld (2014), both to model and to infer are captured by dynamical covariation between two systems. A model is thus a covariation relation. A statistical inference is, too, a covariation relation. As both a model and an inference are now identical, it becomes difficult to see what it would mean for an organism to *use* its covariation relation with the dynamics of B (the model) to covary with the dynamics of B (to make a statistical inference). The errors shown in Section 3.2 are made here. By maintaining a realist position on models in FEP, even in this deflationary sense, we find that some theorists draw unjustified conclusions. This may be avoided if we use clearer terminology. If a FEP model unearths statistical (in)dependencies and covariation relations between internal and external states, we may as well call them by name: statistical relations and covariation.

Another objection is that the best explanation for why our predictive models work, is that the organism actually uses these models itself.[48] It is true that, if an organism were to employ the same models we use to predict its behaviour, that would explain our predictive success, but this does not work *vice versa*. We model weather dynamics, animal population rates as well as biodiversity in- or decreases, yet we are not inclined to take a realist position on these models, despite their predictive successes. For biological organisms to be the exception to this rule, we would need additional argumentation. Note that this is a distinct point from scientific realism versus instrumentalism *tout court*. The model-instrumentalist view is

[48] Rescorla (2016) and Korbak (2019, p. 16) appear to defend something akin to this position.

perfectly compatible with certain forms of realism with regards to the particular features of the target system that are picked out by the model. Recall the model-instrumentalist reading of the embodied approach. This is a realist position with regards to the patterns of covariation and the statistical relations, but an instrumentalist position with regards to the model that captures those relations. Put differently, the *scientifically devised* models of organism-environment systems (for example) pick out *real* statistical covariation relations.

5.6 Conclusion

In this paper I have argued that we should take an instrumentalist approach to models in FEP. I have shown that in the literature, instrumentalist models as created by scientists and realist models as thought to be implemented in and used by organisms are often conflated, and this can lead to erroneous conclusions. Further, a realist take on models in FEP is unwarranted. This is due to two reasons. First, a realist approach leads to overgeneration and trivialization of models. Second, a realist approach to the mathematical mechanisms described in FEP still does not warrant a realist take on models: the mechanisms described are covariational in nature, which falls short of meeting the conditions of both representational and non-representational takes on modeling in science. I have further shown that an instrumentalist approach to models in FEP is safe from these arguments, and is a justified use of modeling. This does not keep one from embracing the results of FEP modeling, and taking a realist approach to the particular dynamical relations between, say, the organism and its environment. I suggest that, in going forward, an instrumentalist position on models in FEP should be maintained. This signals clearly to the reader what is being meant, and may allow us to avoid conflation. Indeed, if a model of a target system unearths a covariational relation between internal and external states, we should take it as the covariational relation that it is.

References

Allen, M., and Friston, K. (2016) From cognitivism to autopoiesis: towards a computational framework for the embodied mind. Synthese, 1–24.

Badcock, Friston, Ramstead (2019) The hierarchically mechanistic mind: A free-energy formulation of the human psyche

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

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

Bruineberg, J., Rietveld, E., Parr, T., van Maanen, L., & Friston, K. J. (2018). Free-energy minimization in joint agent-environment systems: A niche construction perspective. Journal of Theoretical Biology, 455, 161–178. https://doi.org/10.1016/j.jtbi.2018.07.002

Clark, A. (2013) Whatever Next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and brain sciences 36, 181–253

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

Clark, A. (2017) How to Knit Your Own Markov Blanket: Resisting the Second Law with Metamorphic Minds in Metzinger, T. and Wiese, W. (eds.). Philosophy and predictive processing. Frankfurt am Main : MIND Group.

Constant, A., Ramstead, M. J. D., Veissiere, S. P. L., Campbell, J. O., & Friston, K. J. (2018). A variational approach to niche construction. Journal of the Royal Society Interface, 15(141). https://doi.org/10.1098/rsif.2017.0685

Colombo, M. Elkin, L. and Hartman, S. (2018) Being Realist about Bayes, and Predictive Processing. The British Journal for the Philosophy of Science, axy059, https://doi.org/10.1093/bjps/axy059

Colombo, M. and Wright, C. (2018) First principles in the life-sciences: the free-energy principle, organicism and mechanism. Synthese, 1-26

Constant, A., Ramstead, M., Veissière, S., Campbell, J., Friston, K. (2018) A variational approach to niche construction. J. R. Soc. Interface 15: 20170685. http://dx.doi.org/10.1098/rsif.2017.0685

Constant, A., Bervoets, Jo., Hens, K., Van de Cruys, S. (2018). Precise worlds for certain minds: An ecological perspective on the relational self in autism. Topoi. Advance online publication. https://doi.org/10.1007/s11245-018-9546-4

de Bruin, L and Michael, J. (2017) Prediction error minimization: Implications for Embodied Cognition and the Extended Mind Hypothesis. Brain and Cognition 112:58-63

de Oliveira, G.S. (2018) Representationalism is a dead end. Synthese, 1-21. https://doi.org/10.1007/s11229-018-01995-9

Di Paolo, E., & Thompson, E. (2014). The enactive approach. In Shapiro, L. (Ed.), Routledge handbook of embodied cognition. Routledge

Di Paolo, E., Burmann, T., Barandiaran, X. (2017) Sensorimotor Life: An Enactive Proposal. Oxford University Press

Frigg, R. and Hartmann, S., (2018) Models in Science, in Zalta, E. N. (ed.) The Stanford Encyclopedia of Philosophy (Summer 2018 Edition), URL = <https://plato.stanford.edu/archives/sum2018/entries/models-science/>.

Friston, K. (2002) Functional integration and inference in the brain. Progress in Neurobiology 590, 1–31

Friston, K. (2003) Learning and Inference in the Brain. Neural Networks 16, 1325–1352

Friston, K. (2005) A Theory of Cortical Response. Phil. Trans. R. Soc. B 360, 815–836. doi:10.1098/rstb.2005.1622

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

Friston, K. (2011) Embodied inference: or "I think therefore I am, if I am what I think". In: Tschacher, W., Bergomi C. (eds) The implications of embodiment: cognition and communication. Imprint Academic.

Friston, K. (2012) A Free Energy Principle For Biological Systems. Entropy 14, 2100-2121. doi:10.3390/e14112100

Friston, K. (2013). Life as we know it. Journal of the Royal Society, Interface, 10, 20130475.

Friston, K. (2019) A free energy principle for a particular physics. Unpublished manuscript

Friston, K. and Buzsáki, G. (2016) The functional anatomy of time: What and when in the brain. Trends in Cognitive Science, 20(7): 500-511.

Friston, K. J., Daunizeau, J. & Kiebel, S. J. (2009) Reinforcement learning or active inference? PLoS (Public Library of Science) One 4(7): e6421.

Friston, K. and Stephan, K. (2007). Free energy and the brain. Synthese 159(3): 417–458.

Friston, K. and Kiebel, S. (2009) Cortical circuits for perceptual inference. Neural Networks 22, 1093–1104

Gallagher, S, and Allen, M. (2018) Active inference, enactivism and the hermeneutics of social cognition. Synthese 195(6): 1–22.

Giere, R. (2010). An agent-based conception of models and scientific representation. Synthese, 172(2), 269–281.

Giere, R. N. (1988). Explaining science: A cognitive approach. Chicago: University of Chicago Press.

Godfrey-Smith, P., and Sterelny, K. (2016) Biological Information in Zalta, E. N. (ed) Stanford Encyclopedia of Philosophy (Summer 2016 Edition)

Gładziejewski, P. (2016) Predictive coding and representationalism. Synthese, 193(2), 559–582.

Hesp, C., Ramstead, M., Constant, A., Badcock, P., Kirchhoff, M., Friston, K. (2019) A multi-scale view of the emergent complexity of life: A free-energy proposal. In Georgiev, G., Smart, J., Flores Martinez, C.L., Price, M. (Eds.) Evolution, Development, and Complexity: Multiscale Models in Complex Adaptive Systems. Springer

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

Hohwy, J. (2016) The Self-evidencing Brain. Noûs 50(2), 259-285

Hutto, D. and Myin, E. (2013). Radicalizing Enactivism: Basic Minds Without Content. MIT Press

Kiebel, S. J., J. Daunizeau and K. J. Friston (2008). A hierarchy of time–Scales and the brain. PLoS Computational Biology 4(11): e1000209.

Kiebel, S. J., J. Daunizeau and K. J. Friston (2010). "Perception and hierarchical dynamics." Frontiers in Neuroinformatics 4(12). doi: 10.3389/neuro.11.020.2009.

Kiefer, A., Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. Synthese, 195(6), 2387–2415. https://doi.org/10.1007/s11229-017-1435-7

Kirchhoff, M. (2015) Extended cognition & the causal-constitutive fallacy: In search for a diachronic and dynamical conception of constitution Philosophy and Phenomenological Research, 90(2): 320-360

Kirchhoff M, Parr T, Palacios E, Friston K, Kiverstein J. (2018) The Markov blankets of life: autonomy, active inference and the free energy principle. J. R. Soc. Interface 15: 20170792. http://dx.doi.org/10.1098/rsif.2017.0792

Kirchhoff, M. (2018) Predictive processing, perceiving and imagining: Is to perceive to imagine, or something close to it? Philos Stud 175 (3), 751–767

Kirchhoff, M. and Robertson, I. (2018) Enactivism and Predictive Processing: a Non-representational view

Korbak, T. (2019) Computational enactivism under the free energy principle. Synthese, 1-21

Longo, G., Montévil, M., & Kauffman, S. (2012). No entailing laws, but enablement in the evolution of the biosphere. In Proceedings of the 14th international conference on genetic and evolutionary computation conference companion (pp. 1379–1392).

Myin, E. (2016). Perception as something we do. Journal of Consciousness Studies, 23(5–6), 80–104.

O'Regan, K., and A. Noë. (2001). A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences, 24: 939–1031

Ramstead, M., Badcock, P., and Friston, K. (2018) Answering Schrödinger's question: a free energy formulation. Physics of Life Reviews, 24, 1-16

Ramstead, M., Badcock, P., and Friston, K. (2019) Variational neuroethology: Answering further questions: Reply to comments on "Answering Schrödinger's question: A free-energy formulation" Physics of Life Reviews, 24, 59-66

Ramstead, M., Veissière, S., and Kirmayer, L. (2016) Cultural Affordances: Scaffolding Local Worlds Through Shared Intentionality and Regimes of Attention. Front. Psychol. 7:1090. Doi: 10.3389/fpsyg.2016.01090

Rescorla, M. (2016) Bayesian Sensorimotor Psychology. Mind & Language, 31(1), 3–36.

Suarez, M. (2003). Scientific representation: Against similarity and isomorphism. International Studies in the Philosophy of Science, 17(3), 225–244.

Tonneau, F. (2012) Metaphor and Truth: A review of Representation Reconsidered by W. M. Ramsey. Behavior and Philosophy, 39(40), 331-343

van Fraassen, B. C. (1980). The scientific image. Clarendon Press.

van Fraassen, B. C. (2008). Scientific representation: Paradoxes of perspective. Oxford University Press

Wiese, W. (2017). What Are the Contents of Representations in Predictive Processing? Phenomenology and the Cognitive Sciences, 16(4), 715–736.

6 Free-Energy Principle, Computationalism and Realism: a Tragedy

Authors

Thomas van Es 1

Inês Hipólito 2, 3

1 Centre for Philosophical Psychology, Department of Philosophy, University of Antwerp (Belgium)

2 Berlin School of Mind and Brain, Humboldt University (Germany)

3 Wellcome Centre for Human Neuroimaging, University College London (United Kingdom)

Abstract

The free energy principle provides an increasingly popular framework to biology and cognitive science. However, it remains disputed whether its statistical models are scientific tools to describe non-equilibrium steady-state systems (which we call the *instrumentalist* reading), or are literally implemented and utilized by those systems (the *realist* reading). We analyze the options critically, with particular attention to the question of representationalism. We argue that realism is unwarranted and conceptually incoherent. Conversely, instrumentalism is safer whilst remaining explanatorily powerful. Moreover, we show that the representationalism debate loses relevance in an instrumentalist reading. Finally, these findings could be generalized for our interpretation of models in cognitive science more generally.

6.1 Introduction

The free-energy principle (FEP) provides a theoretical framework that primarily aims to unify biological and cognitive science approaches to life and mind (Friston, 2013). Yet it also has ambitions to underwrite classical and quantum mechanics so as to become a theory of every 'thing' (Friston, 2019).[49] In essence, it serves as a mathematical description of life, and states that any living, non-equilibrium steady-state system can be associated with the minimization of free energy.[50] Moreover, a state space description of an organism can be associated with what is called a *generative model* to the extent that a generative model is a joint distribution over (hidden) states and observation (statistical observation). We will explain the specifics of the FEP in more detail below in Section 2.

[49] Every 'thing' as a system that can be modelled at non-equilibrium steady-state (NESS), such as typhoons, electrical circuits, stars, galaxies, and so on. NESS is a physical term that denotes any system that is far from equilibrium, and in a steady state with its environment. We elaborate on this in the next section.

[50] Our paper is neutral on the unifying ambitions of the FEP, and this discussion outside of the scope of this paper. Furthermore, there is an opposite proposal that focuses on entropy maximization instead (Vitas and Dobovišek 2019; Matyushnev and Seleznev 2006; Ziegler 1963).Yet these discussions are outside the scope of this paper.

For now, it is important to note that the generative model has played a key role in certain process theories associated with the FEP, such as predictive coding and processing (Hohwy, 2013; Clark, 2016) and active inference (Ramstead, Friston, Hipólito 2020; Friston et al. 2020; Tschantz et al. 2020; Parr, 2020), yet its status remains unclear.[51] According to predictive processing theories, the generative model is literally implemented by a human brain to calculate the potential states of the environment (termed a *realist* approach), whereas other approaches take it to be an insightful statistical description that a non-scientifically trained organism has no access to (termed an *instrumentalist* approach). There is another debate as to whether this model is representational in nature or not (Gładziejewski, 2016; Gładziejewski and Miłkowski, 2017; Kiefer and Hohwy, 2018; Bruineberg et al., 2016; Kirchhoff and Robertson, 2018). The representations debate is associated with the general debate in the cognitive sciences regarding the concept of representation as a useful posit (Ramsey, 2007; Hutto and Myin, 2013, 2017). The idea is that going non-representationalist may save the generative model's causal efficacy, albeit technically *via* the generative process (Ramstead et al., 2019). In this paper, we shall engage with both debates, and argue that the representationalism debate is not relevant to the FEP. Realism is doomed to fail regardless of whether it is representationalist or not, and, conversely, instrumentalism can thrive either way, or so we shall argue.

Neuroimaging techniques offer important insights into the nervous system, such that we can develop explanations from patterns of activity and/or neuronal structures. However, patterned activity will not answer the question of whether or not it is representational. Indeed, experimental, neuroimaging data per se cannot answer the question of whether in its activity and interactions, the brain represents anything. This would be analogous to conducting experimental research to know whether or not objects represent the law by which they fall. The answer for ontological questions is not in empirical experiments. Thinking that the nervous system *represents* by the same properties as those we use to explain thus means taking a philosophical standpoint. To do that, we need to offer a sound philosophical argument. Thinking that it does, or does not, is a philosophical standpoint.

Inheriting from debates in philosophy of science around instrumentalism *v*s. realism, analytic philosophy of mind debates whether or not mental activity should be conceived of as representational. Scientific realism would prescribe that the technical terms used in modelling a target system also exist in the target system. Realism about Bayesian inference would thus

[51] We discuss the relation between the FEP and its associated process theories in Section 2.

dictate that the activity in the nervous system entails or is an intellectual representation that results from calculus between *posteriors*, *likelihoods* and *priors* (Rescorla, 2016). Instrumentalist thinking would be sceptical to accept the metaphysical assumption that the nervous system employs any of the tools used by scientists to model its activity. For instrumentalists, our capability to model, say, the auditory system, with prediction formalisms such as Bayesian inference, does not imply that the auditory system itself operates by applying Bayesian inference.

The aim of this paper is to show that, philosophically, instrumentalist thinking is less controversial, yet remains explanatorily powerful and can yield important insights in organism-environment dynamics. An instrumentalist attitude about the FEP is a safer bet without losing the potentially high returns. After briefly describing the FEP in Section 2, we assess two proposals made in the realist logical space, that of Representationalist Realism (RR), and Non-Representationalist Realism (NRR) in Section 3. We reject both of them and in Section 4 we proceed to offer positive reasons to embrace instrumentalism about the FEP. Given the activity-dependence feature of neuronal activity, Dynamic Causal Modeling (DCM), under the FEP, seems to be the most suitable and promising set of instruments to preserve the character of neuronal activation as we empirically know it to be – activity in coupled systems. From this angle, realist arguments look like forcing the world to conform with the anthropomorphic instrumental lens we use to make sense of it.

6.2 Free Energy Principle: essentials

The FEP is a mathematical formulation that states that a self-organising system is a dynamical system that minimises its free-energy. The FEP is based on three aspects. First, the observation of *self-organisation*, which refers to our observation of patterns, in time and space from interacting components, plays a crucial role in life sciences (Wedlich-Söldner and Betz, 2018; Hipólito 2019; Levin 2020; Fields and Levin 2020). A self-organised system can be described in terms of the structured dynamics of its behaviour. These patterns can be thought of by the light of density dynamics. That is, the evolution of probability density distributions over ensembled states (known as *variational Bayes*). A self-organising system is a system that, far from equilibrium, is in a steady state with its environment, or in non-equilibrium steady-state (NESS). To be in a *steady state* is to be in one specific state, typically averaged out over time.

'NESS' thus implies environmental exchange to maintain steady states. As such, living systems are considered to be at NESS, because their exchanges with the environment allow them to maintain their physical and structural integrity (considered their 'steady' state). Of course, living systems are in constant flux and thus are only *by approximation* in NESS. This brings us to the important feature doing the explanatory labour: entropy. Entropy, as measure of how things are, where low entropy indicates maintenance of integrity (states concentrated in small regions of the state space), and high entropy, its dissipation (states dissipated in the state space). So, the FEP focuses on entropy reduction.

This brings us to the second aspect: living organisms can be described as (stochastic) dynamical systems possessing attractors. A phase space is a space in which all possible states of a system are represented, where each possible state corresponds to one unique point in the phase space. The gain of energy translates to the expansion of the phase state. Conversely, the loss of energy formally parallels the contraction of the phase space, meaning an increase of certainty and minimisation of entropy or the maximisation of dissipation of energy in the system.

Thirdly, the states in which the self-organising system is at a point in time can be identified by the interactive role they play within the (multilevel) self-organisation scheme. States within the state space can be statistically differentiated by the application of a Markov blanket (Friston, 2020; Hipólito, Baltieri et al. 2020; Hipólito, Ramstead et al. 2020). By this formalism, we can partition the system into internal, external, active, and sensory states. Although internal and external states do not statistically influence one another (as they are conditionally independent), active and sensory states do statistically influence one another to the extent blanket states (internal, sensory, and active) describe the patterned activity of an organism. By these lights, a system that is in NESS possesses a Markov blanket, though as a technical construct.

How can we formally account for the three aspects? The FEP prescribes the patterned activity of organisms in terms of minimisation of the free-energy as per Figure 1.

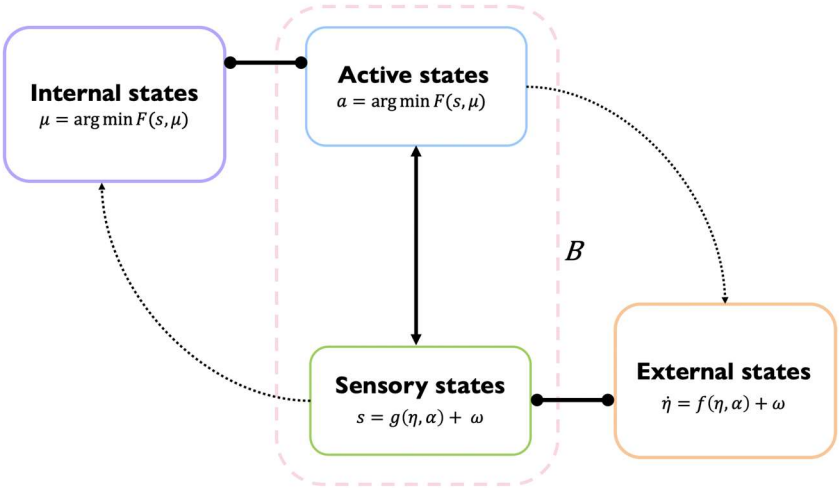

Figure 1. Minimisation of the free-energy by internal states and blanket states (B), comprehending sensory states and active state; and external states, which can be described by the equations of motion, as the function of (hidden states) of the world (η), active states (α), and noise or random fluctuations (ω).

Free-energy is a formal way of measuring the surprisal on sampling some data, given a generative model. Surprise refers to the "unlikeliness" of an outcome, as a measure of unlikeliness, within or in respect to a certain generative model. Mathematically, it qualifies how likely an outcome is, by measuring differences between posterior and prior beliefs of the observer. Technically, surprisal is the difference between accuracy (expected log likelihood) and complexity (i.e. Bayesian surprise or salience, as the informational divergence between the posterior probability and prior probability). Surprisal thus refers to the extent to which new data is 'surprising' to the prior model (surprisal should not be confused with psychological surprise in a day-to-day life setting or in information theory).

It is important to clarify that although the FEP is *related* to theories such as the Bayesian brain hypothesis (e.g. Knill and Richards 1996) and predictive coding (e.g. Hohwy 2013, Clark 2016), it does not entail them (at times a confusion in philosophy of mind). The FEP differs from predictive coding or the Bayesian brain hypothesis in a crucial aspect. The Bayesian brain hypothesis is the view that the brain performs inference according to Bayes's theorem, integrating new information in the light of existing models of the world. To do so, prior probability and likelihood are computed simultaneously to obtain the posterior probability. Predictive coding (Hohwy, 2013) differs from the Bayesian brain hypothesis since it implies that prediction comes first, and is then corrected or updated by data. In this setting, two representations, bottom-down prediction, and bottom-up error signal, either match or mismatch

(but see Orlandi and Lee 2018). The FEP differs from both the Bayesian brain hypothesis and predictive coding, by having at its target the reduction of entropy, rather than the maximisation of hypothesis likelihood given sensory data. The FEP does not join the discussion about the nature of computational processes (whether synchronous or sequential), because the FEP is a framework of states, not processes. The Bayesian brain hypothesis and predictive coding are process theories about how the principle is realised (Hohwy, 2020). The FEP, on the contrary, is a principle that things may or may not conform to. In this regard, the FEP, thus, stands in clear contrast with process theories such as the Bayesian brain hypothesis, predictive coding, or active inference.

The FEP is thus best seen as a research heuristic, a particular lense through which we can view and carve up the world. Associated process theories, then, are concerned with how the FEP is realized in real-world systems.[52] This crucial distinction sets us to realise that the FEP the FEP does not imply process aspects or features, such as representations, pertaining to the theoretical processes that aim to explain how the principle is realised. Yet the FEP does not in itself imply the representational tools employed by these process theories. Prior probabilities and likelihoods are tools used to explain the process by which variational free-energy is minimised. The FEP thus does not answer questions about the implementation of computational processes. Instead, the FEP targets the formal understanding of self-organising behaviour, not computational processes. It aims at explaining and understanding a system's behaviour from observing the self-organising system's patterns and making sense of them in terms of minimisation of variational free energy and entropy reduction.

6.3 Getting real about representations and models

The FEP provides powerful mathematical tools for the description and analysis of dynamic, self-organizing systems. However, the implications of these analyses are disputed. It is unclear what exactly they mean, what they say of the world or what we can do with them. Here we discuss the FEP along two axes, each with two possible values: 1) instrumentalism or realism, and 2) representationalism or non-representationalism, so that there are four possible lines of interpretation, i.e. combinations of philosophical takes on models in the FEP (see table 1).

[52] One could of course defend a predictive coding view of neural processing without subscribing to the FEP's grand ambitions, see Rao & Ballard (1999).

FEP options	Realism	Instrumentalism
Representationalism	REP-REA	REP-INS
Non-representationalism	NRP-REA	NRP-INS

Table 1: Philosophical combinations under the models of FEP. Representationalist realism (REP-REA), non-representationalist realism (NRP-REA), representationalist instrumentalism (REP-INS), and non-representationalist instrumentalism (NRP-INS).

Realism and instrumentalism, here, concern the models and statistical manipulations that make up the FEP, and whether they are thought to be used and manipulated by the systems under scrutiny, independent of scientific inquiry (REA), or, conversely, whether they are thought to be scientific tools, wrought by humans in specific sociocultural environments to study particular systems (INS). Representation is a famously contested term in (philosophy of) cognitive science. Here, we shall use it to refer to at least something with representational content. That is, anything that represents some target system, does so in a way that the target system may not be so (Travis, 2004). This implies that representational content minimally has two aspects: 1) directedness, and 2) truth, accuracy or correctness conditions. First, we shall discuss the realist types: REP-REA and NRP-REA, before turning to the instrumentalist approach in Section 4.

6.3.1 Representationalist realism doesn't work

REP-REA is the view that the models and statistical calculations we use in the FEP formalism are literally employed by either a brain or an organism in its navigation of the world.

Prime examples of the REP-REA view come in the form of process-theoretic offshoots of the FEP, such as predictive coding, predictive processing, or, more generally, PEM theories of cognition (see for accessible introductory texts Hohwy, 2013; Clark, 2016). By employing Bayesian epistemology, scientists refer to the model of the nervous system by using technical terms pertaining to the Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Such that, the *posterior*, as the probability of "A" being true, given that "B" is true is equal to the *likelihood*, as the probability of "B" being true given that "A" is true, times the *prior*, as the probability of "A" being true, divided by the probability of "B" being true.[53]

A realist description of the model in Bayesian technical terms involves the assumption that the activity of the nervous system itself entails the representational properties of the model. That is, the activity of the nervous system aims at an intellectual representation that results from combining *posteriors*, *likelihoods*, and *priors*. The use of this technical wording is so customary, that some scientists apply it interchangeably for the brain and the model of the brain. This is the case of Pouget and colleagues (2013) stating that "there is strong behavioural and physiological evidence that the brain both *represents probability distributions* and *performs probabilistic inference*" (p. 1170, emphasis added). This causes many philosophers to take for granted that the technical terminology used in models of the brain is applicable to the brain itself, helping to paint a picture of the brain as an "inference machine" or the "Bayesian brain hypothesis". (Helmholtz 1860/1962, Gregory 1980; Dayan et al. 1995; Friston 2012; Hohwy 2013; Clark 2016). According to these views, the agent (sometimes considered to be the brain, sometimes the organism, see Hohwy, 2016; Corcoran et al., 2020) is essentially a prediction machine. Subpersonally — that is, unbeknownst to the acting individual — the system predicts, in accordance with Bayes' theorem, what is most likely to occur next, given the current state of affairs and its knowledge of the world. According to PEM's best bet, this knowledge is thought to be of the causal-probabilistic structure of the world, and stored in a 'structural-representational' format (Gładziejewski, 2016; Gładziejewski and Miłkowski, 2017; Kiefer and Hohwy, 2018). As we will see, to get explanatory bite out of the brain as (literally) an inference machine, philosophers on the realist bench need to use the full force of the technical terms employed in the model. Terms such as *posteriors*, *likelihoods*, and *priors*, then directly refer to the nervous system, both in representational (Kiefer and Hohwy 2019; Clark, 2016; Hohwy 2018), and non-representational, or seemingly enactivism inspired proposals (Kirchhoff, 2018; Hohwy, 2018; Ramstead et al. 2019; Hohwy 2020).

At issue in the debate is whether the intellectual process that we apply by using scientific tools in the investigation of a target phenomenon needs necessarily to be supposed as

[53] See how Baltieri & Buckley (2017) address a similar issue.

an ontological feature of the target phenomenon. Indeed, Linson and colleagues (2018) attentively note that as convenient as it may be, expressions such as "the brain "is" Bayesian or "implements" Bayesian models can lend itself to misunderstanding cognition's ontological commitments" (p. 14). Although representational language is oft-used by scientists, it remains to be seen whether this is explanatorily additive or a mere gloss (Cheremo, 2009). The proposed structural-representational format is intended to address this worries, yet, as van Es and Myin (2020) argue, does not seem to solve the well-known problems of invoking representation in cognitive science.

Consider Ramsey's (2007) job description challenge: a representation must minimally fulfil its explanatory role qua its representational status (Ramsey, 2007). A representation as used in cognitive science, it is thought, must fit the job description of what representations *do*. That is, the representation is to be explanatorily powerful *in virtue of being* a representation (and not, say, a covariation relation with an inoperative representational gloss). Further, a representational relation is a three-part relation: 1) a target system, which is represented by 2) a representational vehicle, which in turn is used as such by 3) a representation-user with access to (1) and (2) and the representationally exploitable relations between the two (Nunn, 1909-1910, as cited in Tonneau, 2012). This is closely related to the mereological fallacy (Bennett and Hacker, 2003). In slogan-form, this says that 'brains don't use models or representations, agents do.'

With regards to the two latter points, most of the REP-REA work relies on the philosophical assumption of the organism or the brain as the representation-user with access to the target system and the representational vehicle. Indeed, this assumption permeates much of the technical work and philosophical thinking of the nervous system, made popular as an analogy between the brain and scientist (Helmholtz 1860/1962, Gregory 1980; Hohwy 2013; Clark 2016; Yon, Lange and Press 2019). The longstanding Helmholtzian view (1860/1962), that 'unconscious inferences' are much like the inferences scientists draw, is also supported by Clark (2016):

> the experimenter is here in roughly the position of the biological brain itself. Her task – made possible by the powerful mathematical and statistical tools – is to take patterns of neural activation and, on that basis alone, infer properties of the stimulus (p. 95).

One must be on guard in this respect. Organisms or brains, unlike scientists[54], do not possess a perspective of an external observer. Thinking that it does, requires a sound argument that is not yet offered in the literature. In fact, the hypothesis for the brain as an ideal observer has been often rejected in literature, most recently (Brette 2019; Hipólito et al. 2020) as a bad metaphor (Mirski and Bickhard 2019; Reeke 2019). In agreement with this, Mirski et al. (2020) call for encultured minds in replacement of error reduction minds. Finally, there is what Hutto and Myin (2013, 2017) have termed the hard problem of content. This, very essentially, is the problem with grounding representational content in a naturalistic manner.

It may be fruitful to briefly rehearse here the attempt to meet the job description challenge, and why it fails (Gładziejewski, 2016; van Es and Myin, 2020). Essentially, the positive account is that structural representations as used in, for example, predictive coding, can meet the job description challenge, and solve the hard problem of content (Gładziejewski, 2016; Gładziejewski and Miłkowski, 2017; Kiefer and Hohwy, 2018). Gładziejewski uses the *compare-to-prototype* strategy, which he borrows from Ramsey (2007). He, first, analyzes a prototypical representation — in this case, a cartographic map — and distinguishes 4 features that make a cartographic map the representational tool that it is, and, second, argues step by step how each feature is present in the predictive coding account of structural representation. Cartographic maps, he argues, are structural representations of the terrains they represent, because they (1) exhibit structural similarities to the target, "(2) guide the actions of their users, (3) do so in a detachable way, and (4) allow their users to detect representational errors" (Gładziejewski, 2016, p. 566).

A counterexample shows why this analysis fails at capturing what makes a cartographic map a representation in the first place. Specifically, van Es and Myin (2020) show that a cardboard box and a table top could meet the four conditions, without engaging in any sort of representation whatsoever. Say that you're walking, holding a cardboard box that you hope to place on the table top when home. The cardboard box is structurally similar to the table top at least in terms of its relatively flat surfaces. Indeed, structural similarity comes cheap, so condition (1) is met. This structural similarity is *exploitable*, as Gładziejewski stresses (see also Shea, 2007), and can be used to guide actions of the user. In this case, the structural similarities between the cardboard box and the tabletop can be exploited so that the box can be successfully placed on the tabletop. Moreover, the structural similarities are a *fuel to success* so that,

[54] See Bruineberg, Kiverstein and Rietveld (2018) for a similar approach.

counterfactually, had they not been in place, the actions would be unsuccessful (Gładziejewski and Miłkowski, 2017). Had the cardboard box's surfaces not been similarly flat, but instead convex, the box would not afford to be placed on the tabletop stably, and the box may have ended up falling off instead. As such, condition (2) is met. *Detachability* requires that the exploitable structural similarities are exploitable in some way in the absence of the target system. The map can be used at home to plan a trip, what turns to take where, in the absence of the terrain it represents, for example. Similarly, the exploitable structural similarities between the cardboard box and the tabletop can be used to plan where to place the box, whilst walking home and thus in absence of the tabletop. This means that condition (3) is met. Finally, *error detection* requires that the agent can detect when the supposed representation has erred in representing the target system. A cartographic map may be dated and not properly represent the current state of a city's roads. It is then the manner in which the structural similarities between the map and the terrain *do not* hold that *fuels* the failure of the navigational activity, which can be detected by way of feedback from the real world: you may not be able to take a turn that, following the map, you need to take to arrive at your destination. Returning to our cardboard box, a misalignment in the relations between the surface shape of the respective items, say, if the table top is convex or tilted and slippery, will result in a falling box. Surely, we can detect the fall by way of feedback from the real world. We may see it, it may fall on our feet, it may make a noise, etc. As such, the job description challenge was set up mistakenly, so van Es and Myin (2020) argue. It is not these four features that (conjointly) make a cartographic map a representation — lest a cardboard box represents a tabletop for the same reason. As such, REP-REA's best attempt at standing up to the job description challenge cannot get off the ground.

This also places pressure on the extent to which the other issues can be deemed resolved. The aforementioned brain as a representation-user problems remains unsolved. Further, the hard problem of content is about correctness or veridicality conditions, yet the 'error' detection minimally required here is met by a misaligned surface relation between a cardboard box and a tabletop, which falls short of *representational* error detection. After all, we may *fail* at doing something, without this being *representational* in nature.

6.3.2 Non-representationalist realism doesn't work

Acknowledging the deeply rooted issues with representations, there is a strand of FEP that advocates a non-representationalist approach. Though it is not always clear whether specific accounts are to be placed in instrumentalist or realist camps (see van Es, 2020 for discussion),

we shall discuss here a realist interpretation of the relevant literature, and explain exactly why it cannot work. We associate the NRP-REA literature with the slogan that the *brain* does not *have* a model, the *organism is* a model (Friston, 2013; Ramstead et al., 2019; Hesp et al., 2019). Here we will take 'to be a model' to mean that, essentially, the organism is, embodies or instantiates a model relative to its phenotype, the type of organism that it is, independently of our human, sociocultural modelling practices. This is to say that the model *really* exists, and is actively being used, manipulated or 'leveraged' by any and all self-organizing systems to minimize their free energy.

Kirchhoff and Robertson (2018) argue that the FEP falls short of ascribing representational models to acting organisms. Their target is Kiefer and Hohwy's (2018) notion that the agent has a measure of misrepresentation in terms of the KL-divergence between the prior probability distribution and the posterior probability distribution. The prior probability distribution here refers to the state *before* encountering new evidence, and the posterior probability distribution here refers to the state *after* encountering new evidence. Upon encountering new evidence, the model's probability distribution will be updated to reflect the newfound evidence and how it affects the different aspects of the model. In a sense, then, the difference between the prior and the posterior will be a measure of the extent to which the model has been changed in the updating process. This is then, for the system, a measure for the extent to which its initial model was misaligned. Further, *if* we take the generative model to be representational in nature, the KL-divergence becomes a measure of the extent to which the system *misrepresented*. Yet, Kirchhoff and Robertson point out that the model comparison in the KL-divergence only measures Shannon covariance, not representation (2018). Barring a representational assumption, this means, they suggest, that this falls short of providing a measure of misrepresentation, and only succeeds in providing a measure of covariational misalignment (Bruineberg and Rietveld, 2014; Bruineberg et al., 2019; Kirchhoff and Robertson, 2018). As such, what actually does explanatory work in FEP is the minimization of negative covariance, *not* the minimization of (representational) prediction error.

If we cannot invoke representations in our realist account of the FEP machinery, what does this leave us with? Key terms in the FEP conceptual toolkit are the *generative model*, a probability distribution over sensory states parameterized by the internal states, the *generative process* by which it is placed into contact with external states via active states, and Bayesian updating of the model (Corcoran et al., 2020; Ramstead et al., 2019; Friston, 2013). Without representations, one may wonder, can we still have a generative model? For this, we need to briefly explore what it takes for anything to be called a model. In the current discourse in

philosophy of science, there is a wide variety of accounts with regard to what makes a model, and how it is that they can tell us anything about their target systems.

There are many varieties of models in use in a scientific context, such as scale models, analogous models, idealized models and more. Standardly conceived, each of these is a model of its target in virtue of *representing* that target (Frigg and Hartman, 2020). Bayes's theorem itself is a placeholder, that when furnished with relevant information becomes a model that affords predicting the activity or behavior of the target system given certain conditions. Minimally, a model such as Bayes' theorem, is required to have three features: (1) access (furnishing information or data); (2) a target (neuronal activity); (3) structural similarity (similar causal relations). If we would create a Bayesian model of, say, neuronal activity, then access is accounted for by the furnishing information or data, the target is neuronal activity and structural similarity needs to hold by way of a dynamical covariational relation so that if *X* would wiggle in the registered neuronal activity, something needs to wiggle in the model as well.

If we take a model to be essentially representational, it should be clear, a mere covariation relation is insufficient to warrant model status. Moreover, Bayesian inference can be seen as the implementation of Bayes' theorem on a specific Bayesian model in light of new evidence. In light of this, it seems that without representation, the *generative model*'s status as a model needs to be revoked. This presents a serious problem to those defending NRP-REA (such as Kirchhoff et al., 2018; Ramstead et al., 2019; Bruineberg et al., 2016).

Let us consider this more closely. A generative model is a probabilistic mapping of potential external states relative to the internal states. If we have a multi-dimensional state space that describes a particular system's internal states, with an axis for each variable associated with the system, then the generative model is what tells us, given this state space description, the probability of the possible values for each variable of the external states. This can be extremely useful because each of those variables represents one or some behavioural features of the target system it is a description of. A description, of course, is a form of representation. If we are to take away the representational characteristics of the generative model, the variables over which it is a probability distribution do not actually represent anything at all[55]. It would be a probability distribution over variables that in no way stand in or

[55] We discuss one strand of definitions of 'model': the representational one. Nonetheless, this makes up the bulk of the philosophical literature on the efficacy and ontological status of models. Below we discuss the only outlier position.

are to be seen as surrogates for real-world characteristics or features.[56] It thus seems that, without representation, the generative model is no model at all, and thereby unfit to aid an agent in navigating its environment, as NRP-REA would have it.

NRP-REA seems like a no-starter. Unless, there would be a way of making sense of surrogacy, of something *standing in* for something else *without* reference to representation or representational content. Luck has it that Guilherme Sanchez de Oliveira challenges the representational capacity and motivation even for *scientific models*. *Prima facie*, this seems like the exact way out NRP-REA's generative models require: modeling without representation (2018, 2016). Yet de Oliveira's work may not be the hero they need. He argues that scientific modelling isolates a model from its context, and in doing so, "constrains our ability to see how the nature of the phenomenon is shaped by what brings it about (the individual scientists, the research context, disciplinary traditions, and technological possibilities in addition to properties of the target)" (de Oliveira 2016, p. 96). The challenge to representational properties is that "we get caught up on ethereal metaphysical concerns that have nothing to do with the phenomenon in the real world of scientific practice" (de Oliveira 2016, p. 96). Moreover, elsewhere de Oliveira argues that to judge models' epistemic virtue and their ontological status in terms of their representational relation to a target system is contradictory (2018). In a nutshell, if a model is inherently representational, and if modelling is thus a representational activity, this means that:

> scientists can use models to learn about target phenomena because models represent their targets, *and* that models represent their targets because scientists use them as representations of those targets—in short, this would mean that the reason scientists *can* use models to study real-world phenomena is that they *do* use them to study real-world phenomena. (de Oliveira, 2018, p. 14, emphasis in original)[57]

Vicious circularity ensues. Essentially, the *use* of models is justified in terms of their representational status, yet the representational status itself is grounded in our use thereof. Whether de Oliveira is correct in this analysis of modeling practices in science is irrelevant to our current debate, yet his proposed alternative *is* relevant.

---

[56] This point can be understood in Wittgenstein's understanding of 'nonsensical' propositions, where variables would be radically devoid of meaning, that is to say, transcend the bounds of sense. If we remove the representational characteristics of the generative model, the variables over which it is a probability distribution do not have any referent, i.e. are 'nonsensical' propositions.

[57] This argument is directed at 'mind-dependent' views of the representational relation of models, according to which, our *use* of models *as* representations is crucial to their representational status. See de Oliveira (2018, p. 9-12) for a discussion of mind-independent view that has long gone out of fashion.

A non-representational approach to models as de Oliveira (2018) suggests, keeps only what is essential to our modeling practices. There are at least two features we can distinguish: a model is 1) mediative or surrogative in that it mediates between the modeller and the target system or *stands in* (or surrogates) for the target system to the modeller, and 2) requires training in specific, socioculturally embedded modelling practices (de Oliveira, 2018). He further notes that mediation nor surrogacy are necessarily representational in nature: consider our use of toy guitars or miniature-sized footballs as surrogates for their professional counterparts. These surrogates, further, aid in the 'skill-development and learning transfer' practices. As it is for the toy guitar, so it is for the model, de Oliveira argues (2018). Indeed, scientific models result mostly from procedures and processes of negotiations materially extended across laboratories in the world, and, thereby, across cultures, and from experts to students. We use models to learn about complex systems, and use this knowledge in our manipulation of the target systems indirectly by, for example, informing policy makers. As such, in our scientific endeavors, models can be naturalistic and useful, whilst only counting as representational when embedded, manipulated, and viewed within the appropriate sociocultural practice (Hutto and Myin, 2013, 2017). Here too, it is important to note that these models are devised, employed and explored by agents, not by their brains. Conversely, a scientific model outside of social practices is simply a device with contingent relations to another system that in itself is senseless (in the Fregean sense) and, thereby, holds no explanatory capacity.

Now that we have sketched de Oliveira's motivation for an account of a non-representational approach to models, we need to see whether this helps NRP-REA's predicament. It essentially comes down to whether FEP's generative models display both (1) the mediative or surrogative, and (2) the skill development and learning transfer features of models, if they are ascribed to (or used, implemented, instantiated, or leveraged by) any free energy minimizing system. Feature (1) is easily shown, as the generative model (but more specifically the generative process by which the model is brought into contact with the external world) is considered crucial in determining action policies (Ramstead et al., 2019). Feature 2, however, is, as emphasized above, clearly sociocultural in nature. It is by becoming enculturated in a scientific ecosystem, being trained by experts in the practice, that we attain the relevant sensibilities with regards to construing and manipulating a model, as well as how to leverage it to further our understanding of the target system. The generative model, in NRP-REA, is to be used in some way or another by *any* free energy minimizing system, unconsciously. That is to say that the way the generative model is envisioned to be leveraged

by an organism does not take into consideration the practice that a trainee would need to undergo to become skilled at using complex, statistical models.

In sum, regardless of whether a model is to be seen as a representational device or not, the generative model, if it is to be given a realist reading as is done in NRP-REA, *cannot* reasonably be said to be a model taking into consideration the current state of the literature on the ontological and epistemic status of models. In this section we have first discussed the KL-divergence option if we take models to be essentially representational. We have argued that for a model to actually be about something, refer to something, i.e. it needs to be representational (under this notion), yet the KL-divergence approach resists this. Without this, the probability distribution we apply Bayes' rule to can't actually get off the ground. We have also discussed de Oliveira's option that models are not representational. Yet here too Bayesian inference doesn't hold without learning transfer, professional training, and so on. This means that Bayesian inference will not get off the ground. After all, Bayesian inference is a particular mode of manipulation of a Bayesian model of a target system. These manipulations are performed by agents embedded and trained in sociocultural modelling practices unavailable to the NRP-REA theorist's notion of a generative model as leveraged by an organism. Though we have provided a potential way out for NRP-REA by considering an account of modeling that does not rely on representation, it turned out to be a dead end. This means that NRP-REA, despite carefully avoiding the well-known representationalist's pitfalls, is incapable of balancing themselves on the Bayesian-enactivist tightrope. If neither representationalism nor non-representationalism can make a realist interpretation take off, we may need to consider whether instrumentalism fares any better, and if so, what good it actually does to go instrumentalist. Why not give up on the FEP project in its entirety if the models it describes are not literally employed by the organisms we study?

6.4 Why instrumentalism works

Instrumentalism is, broadly, the idea that the models we use to describe important and interesting statistical relations between, among others, organisms and their environments do just that: describe. It resists the temptation to conceive of organisms as having access to our human sociocultural heritage of making and exploiting models.[58] As such, instrumentalism in

[58] Humans, of course, do have access to our sociocultural heritage. *Prima facie*, this one might consider a 'humans-only' approach. However, the use of models does not, by way of sociocultural heritage, become *innate* (Satne and Hutto, 2015). We have been exposed to imagery all around us, exponentially so the younger you are, which influences our skillset. Though enculturated in a wide variety of representational practices, the particular

itself is characterized by ontological agnosticism with regards to what *actually* makes a system tick. Instead, it is concerned with accurately describing organism-environment dynamics and the interesting relations that may surface.

In this section, we want to first explicate why instrumentalism does not run into any of the issues that realism does. Of interest here is the way in which the question of representationalism transforms from being of vital interest to the FEP project to an interesting related question that helps conceptualize the framework. Second, we want to delve into how instrumentalism can work for us. This is important to emphasize, because otherwise it may seem like we are only losing explanatory ambitions, without gaining anything in return.

6.4.1 The representational collapse and the safer bet

In Section 3, we raised a few concerns with the realist perspective on the FEP. The realist perspective we considered as the model of the FEP, and thus Bayesian inference, is in one way or another literally used, employed, instantiated, embodied, or 'leveraged' by any free energy minimizing system, or at least organisms. We argued that the position is untenable, regardless of whether we take a representationalist or a non-representationalist stance. Bayesian inference using a statistical model of a target system is commonly seen as a representational activity, yet there is no naturalistically viable answer as to how this works outside of our own socioculturally developed representational practices as scientists or philosophers, as we discussed in Section 3.1. Subsequently, in Section 3.2 we find that the non-representational approach and its covariational escape has its explanatory concepts fall one by one. When it concerns essential organismic behavior, we show that without representational content, there is no model; without a model, there is no Bayesian inference; without Bayesian inference, there is no realism. As such, realism is untenable across the board. Yet instrumentalism is not *evidently* free of worries.

Here we shall briefly discuss the issues encountered by realism, why they don't concern the instrumentalist approach, and also why, in general, instrumentalism is a much *safer* bet. We use the same methods, the same conceptual tools, but how they are employed differs wildly. In an instrumentalist perspective, Bayesian inference, as well as any potentially associated representational activity, is not said to be performed, embodied or leveraged by any system other than those humans that have been trained to do so. The same applies to models, and

skill of employing Bayesian inference remains rather *niche*, making it a tough sell for universality. The distinction between activity being *conform* a computational principle and actually *computing* according to this principle is relevant, but is outside the scope of the current paper.

modeling activities, but of course also to any other sociocultural activity such as writing, whether that is formally, calligraphic or graffiti. In the instrumentalist take, organisms do not model anything in and of themselves, but they could potentially be trained by others to engage in certain modeling practices that aim to explain and predict scientific phenomena. FEP models, as well as the inferences we make with them about their target systems, are specific to our human scientific practices of studying the world by way of using idealized surrogates. Where these models originate in, and how they can serve as tools that help us understand the world, then, becomes a question for the history and philosophy of science, *not* for the cognitive sciences.

We see a similar transformation of the issue of representationalism. In the introduction of Section 3, we sketched the possibilities along the two axes of interest: realism *vs* instrumentalism, and representationalism *vs* non-representationalism, leading to four positions: REP-REA, NRP-REA, REP-INS and NRP-INS respectively. For the realist position, REP-REA and NRP-REA are extremely different accounts of how living and cognitive systems navigate their environment. Either the system forms a rich, representational model of (the causal probabilistic structure of) the external world (Hohwy, 2013; Gładziejewski, 2016), or the system covaries adaptively with its environment by 'leveraging' a stipulated generative model (Ramstead et al., 2019; Kirchhoff and Robertson, 2018). Notice, however, that *qua* the FEP, cognitive science, and biology, the representationalism question enters the domain of philosophy of science. Indeed, if we go instrumentalist, as far as our scientific endeavour is concerned, *it doesn't actually matter* whether the models we use are representational or not, it just matters *that they work* (de Oliveira, 2018, pp. 18-20). As such, instrumentalism does not solve the issues of realism, rather, the issues do not even apply to instrumentalism. In fact, they are *dissolved*.

This is particularly interesting when we consider the non-representationalist view presented in the literature (Kirchhoff and Robertson, 2018; Ramstead et al., 2019). In Section 3.1, we placed this view in the NRP-REA camp for argumentative purposes, conceding that the literature itself can technically be read in multiple ways (van Es, 2020). Yet we can see now that arguing for a non-representational take on the models as employed in the FEP, only makes sense under realist assumptions. Only if we *assume* the entire statistical machinery at work in the FEP is literally employed by free energy minimizing systems, does it really matter whether these models imply representationalism (and the problems this is accompanied by) or not. Consequently, this puts the enactivism-inspired 'no representation, just covariation' project in

the FEP literature in a bind. It is either doomed to fail (under realist assumptions) or irrelevant (under instrumentalist assumptions).

At this juncture, one may either deem instrumentalism the god-given gift without philosophical problems, *or* suspect that there is something deeply worrying about it. Or a bit of both, we don't judge. Yet it's exactly this *lack of judgment* that may seem suspect. The realist took a plunge, and, or so we argue, failed. They took a risk and came up empty. Yet it may seem the instrumentalist just waited by the sideline, and only remained safely untouched because they never moved in the first place. That is, it may seem the instrumentalist is only safe from issues because *it doesn't actually make any claims* about the world. It may seem empirically vacuous, without even the promise of helping us understand the world and its distinguishable systems any better. In the remainder, we shall argue that despite giving up the realist claims on the world, instrumentalism in the FEP has much explanatory capacity to offer with respect to new insights in making sense of systems' interactions in terms of patterned activity.

6.4.2 The stakes of instrumentalism or models in neuroscience

In neuroscience, we use different imaging techniques and formal languages to understand the activity of the nervous system. Formal, or mathematical languages are developed and applied to make sense of the overwhelming amount of data collected from imaging the brain, where different formalisms correspond to different models. If the model shows similar patterns of activation to those directly collected from functional neuroimaging, we can obtain not only insights into the neuronal activity itself, but also draw and test new hypotheses related to and within that model. The model, in scientific practice, is a representation of the nervous system to the extent it holds explanatory capacity. This is, as we know, the goal *par excellence* of computational neuroscience.

Indeed, computational neuroscience simulates the neuronal processes to infer models that explain and predict the phenomena. There are two major ways to model neuronal processes. One is Structural Causal Models (SCM), which typically applies machine learning or information theory to model the system in conformity with the presence or absence of 'information'. The goal in SCM is precisely to display topological maps of brain structures as per the presence or absence of 'information' amongst highly connected (neural) modules or

nodes.[59] This is the set of techniques *par excellence* of brain mapping.[60] The other way of modelling neuronal processes is by Dynamical causal models (DCM), employed to explain the activity-dependent patterns found in the nervous system. Applied to the FEP is the modelling of activity-dependence in coupled systems by means of dynamical formalisms. As simulation models that aim to hold predictive capacity, both models – SCM and DCM – apply the statistical tools of Bayesian epistemology[61], viz. Bayesian inference.

6.4.3 How instrumentalism can work for us

The main question for the FEP is, not about processes, but self-organising behaviour. As we have explained in section 2, the FEP aims at explaining and understanding a system's behaviour from observing the self-organising system's patterns and making sense of them in terms of minimisation of variational free energy and entropy reduction.[62] As a principle, the FEP is expected to apply to different levels of self-organisation.

The behaviour of (self-organising) systems can be described as acting to minimise expected free energy, and to reduce expected surprisal. Living systems, such as cells in a tissue, neurons in a network, brains in organisms, organisms in environments and so on, enacting their environments, could be thought of as actions for epistemic affordance. By epistemic affordance we mean actions that avoid dissipation (resolve uncertainty and, thereby, expected free energy).[63] In order to avoid dissipation, opportunities for resolving uncertainty become attractive. Appealing to dynamical systems theory, this can be described as a random dynamical attractor: a dynamical system in which the equations of motion have an element of randomness or fluctuations to them. An example of a random dynamical system is a stochastic differential equation, describing and accounting for the important aspect of noise. Brown (1827), examining the forms of particles immersed in water, "observed many of them very evidently in motion". Albert Einstein (1905) noted they arose directly from the incessant random pushes,

[59] Where the aim is to highlight the structure by determining (predicting the likelihood) of connections between modules in terms of information being exchanged between modules - thus by the presence or absence of information.
[60] See Pearl (2001); Spohn (2010); Bielczyk et al. (2019); Borsboom, Cramer Kalis (2019); Straathof et al. (2019).
[61] See (Talbot, 2016).
[62] We do not claim that FEP offers the ultimate answer to *all* behavior.Yet it may be key in making sense of certain biologically essential levels of cognition.
[63] This does not mean that propositional information is extracted from the environment.

or perturbations, to the particle made by molecules in the surrounding fluid. Langevin (1908) formulated the first stochastic equation to describe Brownian motion emphasising the dynamical behaviours observed in the interplay between deterministic processes and noise.[64] Randomness or fluctuations (such as Brownian motion, or even cell or neuronal activity) are 'noisy' to the extent that their origin implies 'degrees of freedom'. Notably, noise can drastically modify the even deterministic dynamics.[65] Importantly, this means that stochastic dynamical systems, accounting for noise, are equipped, at least in principle, to capture how existing states contribute to adaptation. State-space models are among the most suitable sets of techniques (Razi and Friston 2016) to model the unfolding activity or behavior of a system subject to fluctuations and noise, described by an ordinary differential equation (ODE):

$$\dot{\mathbf{x}}(t) = f\big(\mathbf{x}(t), \boldsymbol{\theta}, \mathbf{u}(t)\big) + \mathbf{w}(t) \qquad (1$$

Where f denotes the coupled dynamical system where θcorresponds to the parameters of the influences; $\mathbf{x}(t) = [x_1(t), x_2(t),...,x_n(t)]^T$ represents the rate of change over the time in state variables $\mathbf{x}(t)$. And, finally $\mathbf{w}(t)$ represents the random influences that can only be modelled probabilistically. Although random influences play an important role in 'stochastic' systems, they are typically de-emphasised in most formal applications by being replaced or absorbed into prior distributions over parameters. Considering however the relevance of noise, e.g. to stabilize unstable equilibria and shift bifurcations[66]; motivate transitions between coexisting deterministic stable states or attractors; or even induce new stable states that have no deterministic counterpart, taking random fluctuations as priors, we think, blurs the line between dynamical and deterministic systems. Models, instead of aiming to represent something, should be able to capture the essential aspect of random influences (the patterns we alluded to above) and thus offering a more comprehensive understanding of behaviour. In a real-world scenario, systems, like cells, neurons, or organs, can be described as subject to fluctuations or noisy

[64] This is especially relevant under the observation that at the very least noise acts as a driving force exciting internal modes of oscillations in both linear and nonlinear systems (where the latter corresponds to the enhanced response of a nonlinear system to external signals, see Jung, 1993; Gammaitoni et al., 1998; Lindner et al. 2004).
[65] Even if it is possible to use deterministic equations of motion to study a system subjected to the action of a large number of variables, the deterministic equations need to be coupled to a "noise" that simply mimics the perpetual action of many variables.
[66] The parameter value at which the dynamics change qualitatively (Arnold 2003).

environments. However these systems 'act', can be understood as avoiding dissipation (or resolving uncertainty and, thereby, expected free energy).

There is no reason to think of this form of action as intellectual thinking (i.e. representing). The intellectual part of the story is our scientific or philosophical attempt to make sense and understand observed behaviour. So, we describe behavior or actions that, from an external observer standpoint, look as if the subjects of enactment were asking 'what would happen if I did that' (Schmidhuber 2010). We can develop and use tools of process theories (e.g. predictive coding, predictive processing) to explain how systems resolve uncertainty (and thereby minimise the free energy). We can use formal terminology, such as *intrinsic value*, *epistemic value of information*, *Bayesian surprisal*, and so on (Friston 2017), to develop models that explain the neuronal processes enabling and underlying things becoming salient to a system to resolve uncertainty. For example, to develop models that explain neuronal excitatory and inhibitory projections in terms of predictions and prediction errors, respectively. In this scientific route, an open question for process theories in relation to the FEP is which theory, predictive coding, Bayesian filtering, belief propagation, variational message passing, particle filtering, and so on, if any, conforms to the FEP. More precisely, which model, if any, conforms with the FEP.

Yet from the fact that it is possible to model a process, it does not necessarily follow that the target phenomenon represents the intellectual tools we use to model it. Consider a moving object that can be explained by Newton's Law of Motion. That we can model the movement by that formalism, does not follow that the object represents the law by which it falls. Few people would claim that the object *represents (or embodies, instantiates, implements, employs, leverages)* the laws by which it moves. Because science does not back this up, those who wish to do so, are committed to a philosophical assumption that moving objects, like cells, or organs like the nervous system, represent laws, principles, or the intellectual tools we use to describe processes conforming to laws or principles (*posteriors*, *likelihoods*, and *priors*). Friston, Wiese and Hobson (2020) are on the guard on this matter, pointing that, from the fact that it is possible to map states, "does not mean that the resulting descriptions refer to entities *that actually exist*" (p. 17, emphasis added).

FEP is not in itself a commitment to the picture that an organism, and/or its nervous system literally is a hierarchical system that itself aims at representation (Baltieri and Buckley

2019; Gallagher 2020; Hipólito et al. 2020; Williams 2020). This is because the FEP targets understanding behavior, from the observation of its dynamical states, in terms of self-organisation towards the aim of avoiding dissipation. In the FEP, the notion of salience plays an essential role read according to the reduction of entropy. What becomes salient is what reduces entropy, put simply. It is these enactive and cultural aspects that are lacking from process theories aiming at developing representational models. Explaining why things become salient is an *explanandum* of the FEP. In this setting, active inference, as a process theory, is an important FEP associate tool to explore the processes enabling and underlying things becoming salient, because it accounts for salience as an attribute of the action itself in relation to the lived world. In pushing in this direction, active inference seems formally equipped to a more accurate description/model of real-world sociocultural scenarios. But active inference is a process theory, i.e. it aims at explaining *the processes by which* things become salient to an agent, not *why* they become salient - that is a goal of the FEP.

The instrumentalist account we propose here understands the use of models without the need to assume that the target system also engages in a representational activity. From the fact that we can generate a high probability value that allows us to draw claims about behaviour, from within our model of an enactive system, we are not licenced to assume that the enactive system itself represents the laws by which it adapts. Such a claim would imply a further claim: that nature, essentially, represents. This does not seem metaphysically reasonable. We think that instrumentalism associated with FEP offers sufficient explanatory power without falling into problematic realist assumptions. In what follows we explain how the FEP, as a tool, holds explanatory capacity for the investigation and understanding of organisms enacting the environment.

In Section 3 on realism, we discussed the KL-divergence argument against representational aspects of FEP (Kirchhoff and Robertson 2018). As we attempted to show in Section 4.1, this argument against representationalism only holds under a realist assumption. The KL-divergence 'solution' to the problems with representations becomes irrelevant. Indeed, only if the model is actually thought to be used, manipulated (or 'leveraged') by the organism, does it actually make sense to try and resolve representationalist worries. Yet in our instrumentalist account this is a non-issue.

In conclusion, we do not think that there are convincing reasons to believe that organisms or systems engage in representation, nor to think that our scientific models are themselves necessarily representational. Situated in a sociocultural practice, models allow us to make culturally informed inferences about the likelihood of something being the case with the target system (i.e ontological claims). So, the instrumentalism we propose does not assume that generative models used in the modelling are models that are used by organisms or systems themselves, nor that models are representational outside of the culture they are developed in. Neutral with regards to realist ascriptions, our instrumentalist account for the offers sufficient explanatory power to explain the behaviour of systems or organisms without falling into unnecessary philosophical problems.

6.5 Conclusion

In this paper we have defended the instrumentalist take on the FEP, arguing that the realist approach is a non-starter, regardless of whether it is representationalist or not. Crucially, the question as to whether systems do or do not model their environment will not be decided by neuro-imaging studies or the models we employ in interpreting the data. This is a philosophical matter that should be dealt with by way of philosophical argumentation. We have argued that the representationalist realist (REP-REA) position does not hold up because of the as of yet missing naturalistic grounding of representations independent of sociocultural practices, including structural representations (van Es and Myin, 2020; Hutto and Myin, 2013, 2017). The non-representationalist realist position (NRP-REA) purports to solve the issues of REP-REA by removing representational content from the story. Yet it does not hold up because without content, there is no model and no Bayesian inference. The instrumentalist does not face the same problems, as they do not ascribe the modeling activity to the organism under scrutiny. The question of representationalism then turns into a general philosophy of science debate on the ontology of models in science, on which the validity or usefulness of the FEP does not hinge (de Oliveira, 2018). The instrumentalist position, then, means that we take the statistical machinery to be a helpful description of real life systems, potentially offering deep insights into the relevant statistical relations between organism and environment. The instrumentalist does *not* take the models we make of the organisms to be employed by the organisms themselves *in virtue* of our capacity to model them.

The difference between realism and instrumentalism is thus primarily ontological in nature: in realism, there is an ontological claim with regards to the status of models in living systems, whereas in instrumentalism there is no such ontological claim. This may be seen as a weakness, as it looks as though the instrumentalist position only gives up explanatory ambitions relative to the FEP realist. This is true. However, the ambitions given up on, we argue, are never going to be met. If this is on the right track, the realist's ambition is a fata morgana, if you will. As such, instead of chasing ghosts, the instrumentalist position is more realistic in their ambitions. There is, within this more modest framework, still plenty of insight to be gained into the workings of life and cognition by way of dynamic causal modeling (DCM). In sum, we argue that modesty and ambition go hand in hand when it comes to models and the FEP.

References

Baltieri, M., & Buckley, C. L. (2017, September). An active inference implementation of phototaxis. In *Artificial Life Conference Proceedings 14* (pp. 36-43). One Rogers Street, Cambridge, MA 02142-1209 USA journals-info@ mit. edu: MIT Press.

Baltieri, M., & Buckley, C. L. (2019). Generative models as parsimonious descriptions of sensorimotor loops. *arXiv preprint arXiv:1904.12937*.

Brown, R. (1828). XXVII. A brief account of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *The Philosophical Magazine*, 4(21), 161-173.

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

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

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

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Corcoran, A. W., Pezzula, G., and Hohwy, J. (2020) From Allostatic Agents to Counterfactual Cognisers: Active Inference, Biological Regulation, and The Origins of Cognition. *Biology and Philosophy, 35*(3). https://doi.org/10.1007/s10539-020-09746-2

Crauel, H., Flandoli, F. (1994) Attractors for random dynamical systems. *Probab. Th. Rel. Fields 100***,** 365–393. https://doi.org/10.1007/BF01193705

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The helmholtz machine. Neural computation, 7(5), 889-904.

Einstein, A. (1905). On the movement of small particles suspended in stationary liquids required by the molecularkinetic theory of heat. *Ann. d. Phys, 17*(549-560), 1.

Fields, C., & Levin, M. (2020). Scale-Free Biology: Integrating Evolutionary and Developmental Thinking. *BioEssays*, 42(8), 1900228.

Frigg, R. and Hartmann, S. (2020) Models in Science in *The Stanford Encyclopedia of Philosophy (Spring 2020 Edition)*, Zalta, E. N. (ed.), URL = <https://plato.stanford.edu/archives/spr2020/entries/models-science/>.

Friston, K., Parr, T., Yufik, Y., Sajid, N., Price, C. J., & Holmes, E. (2020). Generative models, language and active inference. *PsyArXiv*. DOI: 10.31234/osf.io/4j2k6

Friston, K. (2012). The history of the future of the Bayesian brain. *NeuroImage*, *62*(2), 1230-1233.

Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface, 10*, Article 20130475.

Friston, K. (2019). *A free energy principle for a particular physics.* Unpublished manuscript.

Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, *159*(3), 417-458.

Friston, K. J., Fagerholm, E. D., Zarghami, T. S., Parr, T., Hipólito, I., Magrou, L., & Razi, A. (2020). Parcels and particles: Markov blankets in the brain. *arXiv preprint arXiv:2007.09704*.

Friston, K. J., Fortier, M., & Friedman, D. A. (2018). *Of woodlice and men: A Bayesian account of cognition, life and consciousness. An interview with Karl Friston*. ALIUS Bulletin, 2, 17-43.

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

Gallagher, S. (2020). Action and interaction. Oxford University Press.

Gammaitoni, L., Hänggi, P., Jung, P., & Marchesoni, F. (1998). Stochastic resonance. *Reviews of modern physics*, *70*(1), 223.

Gregory, R. L. (1980). Perceptions as hypotheses. Philosophical Transactions of the Royal Society of London. B, Biological Sciences, 290(1038), 181-197.

Gulli, R. A. (2019). Beyond metaphors and semantics: A framework for causal inference in neuroscience. Behavioral and Brain Sciences, 42.

Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: causally relevant and different from detectors. *Biology & philosophy, 32*(3), 337–355. https://doi.org/10.1007/s10539-017-9562-6

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese, 193*(2), 559–582.

Hesp, C., Ramstead, M., Constant, A., Badcock, P., Kirchhoff, M., & Friston, K. (2019). A multi-scale view of the emergent complexity of life: A free-energy proposal. In G. Georgiev, J. Smart, C. L. Flores Martinez, & M. Price (Eds.), *Evolution, development, and complexity: Multiscale models in complex adaptive systems* (pp. 195–227). Springer.

Hipólito, I., Baltieri, M., Friston J, K., & Ramstead, M. J. (2020). Embodied Skillful Performance: Where the Action Is. *Synthese*.

Hipólito, I., Ramstead, M., Constant, A., & Friston, K. J. (2020a). Cognition coming about: Self-organisation and free-energy: Commentary on "The growth of cognition: Free energy minimization and the embryogenesis of cortical computation" by Wright and Bourke (2020). *Physics of Life Reviews*.

Hipólito, I., Ramstead, M., Convertino, L., Bhat, A., Friston, K., & Parr, T. (2020b). Markov blankets in the brain. *arXiv preprint arXiv:2006.02741*.

Hipólito, I. (2019). A simple theory of every 'thing'. *Physics of life reviews, 31*, 79-85.

Hohwy, J. (2013). The predictive mind. Oxford University Press.

Hohwy, J. (2018). The predictive processing hypothesis. *The Oxford handbook of 4E cognition*, 129-146.

Hohwy, J. (2020). Self-supervision, normativity and the free energy principle. *Synthese*, 1-25.

Hutto, D. D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Mit Press.

Hutto, D. D., & Myin, E. (2017). *Evolving enactivism: Basic minds meet content*. MIT press.

Jung, P. (1993). Periodically driven stochastic systems. *Physics Reports*, *234*(4-5), 175-295.

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. Routledge handbook to the philosophy of psychology, 2nd ed. Oxford, UK: Routledge.

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

Kirchhoff, M. (2018). Predictive brains and embodied, enactive cognition: an introduction to the special issue. *Synthese*.

Lemons, D. S., Gythiel, A., & Langevin's, P. (1908). "Sur la théorie du mouvement brownien [On the theory of Brownian motion]". *CR Acad. Sci.(Paris)*, 146, 530-533.

Levin, M. (2020, July). Robot Cancer: what the bioelectrics of embryogenesis and regeneration can teach us about unconventional computing, cognition, and the software of life. *In Artificial Life Conference Proceedings* (pp. 5-5). One Rogers Street, Cambridge, MA 02142-1209 USA

Lindner, B., Garcıa-Ojalvo, J., Neiman, A., & Schimansky-Geier, L. (2004). Effects of noise in excitable systems. *Physics reports*, *392*(6), 321-424.

Linson A, Clark A, Ramamoorthy S and Friston K (2018) The Active Inference Approach to Ecological Perception: General Information Dynamics for Natural and Artificial Embodied Cognition. *Front. Robot. AI 5*:21. doi: 10.3389/frobt.2018.00021

Martyushev, L. M., & Seleznev, V. D. (2006). Maximum entropy production principle in physics, chemistry and biology. *Physics reports*, *426*(1), 1-45.

Mirski, R., & Bickhard, M. H. (2019). Encodingism is not just a bad metaphor. *Behavioral and Brain Sciences*, 42.

Mirski, R., Bickhard, M. H., Eck, D., & Gut, A. (2020). Encultured minds, not error reduction minds. *Behavioral and Brain Sciences*, 43.

Nunn, T. P. (1909-1910). Are secondary qualities independent of perception? *Proceedings of the Aristotelian Society, 10,* 191-218.

Orlandi, N. & Lee, G. (2018). How Radical is Predictive Processing? in Eds., Colombo, Irvine, & Stapleton, *Andy Clark & Critics.* Oxford University Press

Parr, T. (2020). Inferring What to Do (And What Not to). *Entropy, 22*(5), 536.

Pearl, J. (2001). Bayesian networks, causal inference and knowledge discovery. *UCLA Cognitive Systems Laboratory, Technical Report*.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature neuroscience*, *16*(9), 1170–1178. https://doi.org/10.1038/nn.3495

Ramsey, W. M. (2007). *Representation reconsidered*. Cambridge University Press.

Ramstead, M. J., Friston, K. J., & Hipólito, I. (2020). Is the free-energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy*, *22*(8), 889.

Ramstead, M. J., Kirchhoff, M. D., Constant, A., & Friston, K. J. (2019). Multiscale integration: beyond internalism and externalism. *Synthese*, 1-30.

Ramstead, M. J. D., Kirchhoff, M. D., & Friston, K. J. (2019). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior, 28*(4), 225-239.

Rao and Ballard, (1999) Predictive Coding in the Visual Cortex: a Functional Interpretation of Some Extra-classical Receptive-field Effects. *Nature Neuroscience, 2*(1):79-87

Razi, A., & Friston, K. J. (2016). The connected brain: causality, models, and intrinsic dynamics. IEEE Signal Processing Magazine, 33(3), 14-35.

Reeke, G. N. (2019). Not just a bad metaphor, but a little piece of a big bad metaphor. *Behavioral and Brain Sciences*, *42*.

Rescorla, M. (2016). Bayesian sensorimotor psychology. *Mind & Language, 31*(1), 3–36.

Satne, G., & Hutto, D. (2015). The Natural Origins of Content. *Philosophia*, *43*(3), 521–536. https://doi.org/10.1007/s11406-015-9644-0

Shea, N. (2007). Consumers need information: Supplementing teleosemantics with an input condition. *Philosophy and Phenomenological Research, 75*, 404–435.

Tonneau, F. (2012) Metaphor and truth: A review of *Representation Reconsidered* by W. M. Ramsey, *Behavior and Philosophy, 39/40*, 331-343.

Travis, C. 2004. The silence of the senses. *Mind 113* (449): 57–94.

Tschantz, A., Baltieri, M., Seth, A. K., & Buckley, C. L. (2020, July). Scaling active inference. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

van Es, T. (2020). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior*. https://doi.org/10.1177/1059712320918678

van Es, T. and Myin, E. (2020) Predictive processing and representation: How less can be more. In Mendonça, D., Curado, M., and Gouveia, S. S. (eds) *The philosophy and science of predictive processing*. Bloomsbury.

Vitas, M., & Dobovišek, A. (2019). Towards a general definition of life. *Origins of Life and Evolution of Biospheres*, *49*(1-2), 77-88.

von Helmholtz, H. (1962). Handbuch der physiologischen optik. 1860/1962. & Trans by JPC Southall Dover English Edition.

Wedlich-Söldner, R., & Betz, T. (2018). Self-organization: the fundament of cell biology.

Williams, D. (2020). Predictive coding and thought. Synthese, 197(4), 1749-1775.

Yon, D., de Lange, F. P., & Press, C. (2019). The predictive brain as a stubborn scientist. Trends in cognitive sciences, 23(1), 6-8.

Zarghami, T. S., & Friston, K. J. (2020). Dynamic effective connectivity. *Neuroimage, 207*, 116453.

Ziegler, H. (1963) Some extremum principles in irreversible thermodynamics with application to continuum mechanics. In: Sneddon, I.N., Hill, R. (eds.) *Progress in Solid Mechanics,* North-Holland, Amsterdam, pp. 91–193

Part III

7 Autism as gradual sensorimotor difference: From enactivism to ethical inclusion

Authors

Thomas van Es 1

Jo Bervoets 1,2

1 Department of Philosophy, University of Antwerp, Belgium

2 Leuven Autism Research (LAuRes), KU Leuven, Belgium

Abstract

Autism research is increasingly moving to a view centred around sensorimotor atypicalities instead of traditional, ethically problematical, views predicated on social-cognitive deficits. We explore how an enactivist approach to autism illuminates how social differences, stereotypically associated with autism, arise from such sensorimotor atypicalities. Indeed, in a state space description, this can be taken as a skewing of sensorimotor variables that influences social interaction and so also enculturation and habituation. We argue that this construal leads to autism being treated on a par with other sensorimotor atypicalities such as blindness or atypical height. This leads to our conclusion that, insofar there is an ethical call to inclusion in our public sphere regardless of contingent bodily difference, an enactivist take on autism naturally leads to extending such inclusion to autism. Moreover, our analysis suggests a concrete way forward to achieve inclusion of autistics: by being more attentive to the autistic sensorimotor specifics.

7.1 Introduction

Autism is a field of intensive cognitive science theorizing and research. There it is primarily conceived as a social-cognitive deficit. The most prominent suggestion is the Theory of Mind hypothesis (Baron-Cohen 2000), but other cognitive theories such as Weak Central Coherence (Happé and Frith 2006) and (deficits of) Executive Functioning (Frith 1996) are part of this perspective. Such theories have been criticized for unwarrantedly positing cognitive modules to perform some kind of representational calculus in the brain (Hipolito et al. 2020), as well as for their ethical implications (McGeer 2007; Bervoets and Hens 2020). Severe criticism of autism as a deficit to social-cognitive functioning also comes from neurodiversity scholars (Milton 2017, Chapman 2020). Therefore, recent autism research increasingly views autism as, fundamentally, a *sensorimotor atypicality*, such as the 'High and Inflexible Precision of Prediction Error' (HIPPEA) account (Van de Cruys et al. 2014), but see also (Mottron et al. 2006; Markram and Markram 2010) for different proposals. All these theories are in the wide

sense theories of sensorimotor atypicality and therefore, as theories of a different autistic embodiment, amenable to enactive treatment. They are not wedded to a homogeneous autistic phenotype consisting of social-cognitive deficits and instead naturally align with an addition of hypersensitivity in the DSM5 (APA 2013)[67]; an addition that is in large part a consequence of a type of self-advocacy of which the neurodiversity movement was born (Pellicano 2013).

Sensorimotor theories of autistic embodiment (Van de Cruys et al. 2014; Mottron et al. 2006; Markram and Markram 2010) instead embrace heterogeneity of the autistic phenotype by trying to account in a statistical way for how social-cognitive issues *may* emerge *given* sensorimotor atypicalities *and* societal and personal contexts. Constant et al. (2018) develop this notion of emergence based on HIPPEA. A crucial concept that these emergentist accounts share with enactivism is that of *scaffolding* (Krueger and Maiese 2018), meaning any specific trajectory of an autistic is never wholly reducible to autism as such, but includes sensorimotor and social interactions totally specific to any autistic individual. Enactivism, broadly understood, is a strand in cognitive science that stresses the importance of embodiment, autonomy and sensorimotor interactivity over representations and brain states to understand cognition and behavior (Varela et al. 1991; Thompson 2007; Di Paolo et al. 2017; Hutto and Myin 2017). This sensorimotor, or perception/action-driven, approach is more open to taking into account autistic lived experience and - by that token - able to overcome the ethical issues mentioned in an interdisciplinary and productive way (Bervoets and Hens 2020).

There is a lot of recent work (Hipólito et al 2020; van Grunsven 2020) addressing how enactivism can lead to seeing autism in line with the neurodiversity tenets. That is, it does not reduce associated issues to an inherently problematic individual neurology. Instead these issues are seen as, at least in large part, the results of contingent social practices at odds with natural human diversity. However, there is, as of yet, no detailed bottom-up account of how specifically the enactive construal of autism as the type of sensorimotor atypicality here considered entails an ethical and inclusionary approach. Even if autism and enactivism have been linked in much depth via the key concept of 'participatory sense-making' (Di Paolo & De Jaegher 2012; De Jaegher 2013; Fuchs & De Jaegher 2009), a constructive account linking autistic embodiment via enactivist principles to a better inclusion of neurodiversity has not been concretely given. In fact, an enactivist treatment of autism and inclusion of autistics and

[67] The DSM-5 also mentions hypo-responsivity but according to sensorimotor theories this can be seen as connected to, and derivative of, hypersensitivity on sensorimotor accounts, see for instance Van de Cruys et al. (2019).

their lived experiences are everything but self-evident, see van Grunsven (2020) referring inter alia to Hutto (2003).

In this paper we shall explore how an autistic sensorimotor difference *can* lead to social atypicalities that autism is stereotypically associated with. We will do this on the one hand by employing a sensorimotor atypicality view on autism to show how skewing of certain variables associated with sensorimotor interaction influences autistic (developmental) trajectories.[68] We will on the other hand rely on critical reading of *Linguistic Bodies* (Di Paolo et al. 2018) to see how in an enactivist framework social and sensorimotor habits are mutually constrained, leading to diverse tensions between autistic and neurotypical people. Taking both elements together, we argue autism to be one of many gradually distributed sensorimotor differences (such as height) which we overcome in each and every social encounter. Insofar a normative ethical claim consists in furthering inclusion in our public sphere of contingent bodily differences, an enactivist take on autism then seamlessly leads to extending such inclusion to autism. Given our analysis is based on a precise understanding of autistic embodiment, it also suggests a concrete way forward to achieve autistic inclusion through recognizing the specific autistic sensorimotor atypicalities.

The paper is straightforwardly organized in three sections corresponding to the key elements in the above paragraph. The next section deals with how autistic sensorimotor atypicalities, as seen through the lens of HIPPEA, translate into a specific skewing of social trajectories. Section 3 links this specific skewing to sensorimotor and social habit-forming, and the effect this can have on intersubjective social autonomous interactions. Section 4 builds on this to develop an ethical claim. The argumentative structure is then as follows.

If, as we argue,

1) autism is fundamentally a gradual sensorimotor atypicality that, as sensorimotor habits do, potentially branches out to social atypicalities, and,

2) broadly following Di Paolo et al. (2018), the overcoming of sensorimotor differences is inherent to any social interaction,

then, we infer,

3) we *should* strive to include autistic ways of being despite the atypicalities by which autism is perceived, furthering inclusivity in the public sphere in the process.

[68] Our treatment below relies on the HIPPEA account, yet is not limited to this specific account. Its key premises are limited to sensorimotor atypicality on the one hand and the enactivist framework on the other. This also means (see section 4) that our argument is as such not limited to autism but extends to any specific difference in embodiment.

Furthermore, in our conclusion we suggest that this principled enactive treatment of overcoming gradual differences - or differences in degree - can be leveraged concretely to further inclusion also of other cognitive or neurodevelopmental 'disorders'.

7.2 Autism as atypical sensorimotor embodiment

The recent trend in autism research is to try to capture autistic atypicality as flowing from an underlying sensorimotor difference. This aligns with common experiences reported in self-advocacy and qualitative research focusing on sensory sensitivity (Hens & Langenberg 2018; Pellicano 2013) and is already enshrined in the diagnostic criteria of DSM-5 (APA 2013). Accepting atypical sensorimotor embodiment as a core commonality underlying autistic heterogeneity allows a). to avoid ethical issues equating autism with higher cognitive (dys)function (Bervoets and Hens 2020) and b). to link autism research to enactivist views that eschew relying on higher cognitive brain modules and rather rely on emergentist notions like 'scaffolding' (Krueger and Maiese 2018). Whilst, in principle, the benefits of our account are therefore compatible with any sensorimotor difference postulated as underlying autism, we rely for a detailed exposition of how sensorimotor atypicality influences developmental trajectories on HIPPEA (Van de Cruys et al. 2014). We do this because it allows us to make this connection as precise as possible for the reader. Also, autistic scaffolding accounts based on HIPPEA already have been proposed by, for instance, Constant et al. (2018), the underlying Predictive Coding view has been theoretically linked to dynamic misattunement (Bolis et al. 2017) and they are being intensively researched empirically, for example: Palmer et al. (2015) or Zaidel et al. (2015).

In essence, what we want to show is that viewing autism as atypical sensorimotor embodiment leads, in a state space description of the agent, to skewing certain variables associated with the sensorimotor interaction that is crucial for social normativity (see section 3). Given such skewing we can then make precise how autistic developmental trajectories are such that they give rise to the type of social misattunement that has been too often, stereotypically, mistaken for the social-cognitive essence of autism. Before we go into the specifics let us first illustrate this with a familiar paradigm case of sensorimotor difference: blindness. Nobody would say that it is impossible for blind people to get to certain states of social interaction. What is different between blind and non-blind people is *how* they get to the physical place in which the social interaction takes place. They clearly will not arrive there with the same environmental (accessibility) conditions. The blind person will rely on, for

example, a touching stick or echolocation or acoustic guidance. The non-blind person will, typically, rely on visual cues and would - failing those - feel quite lost. The sensorimotor difference then is generally accepted to impact only the *how* of navigating physical space and not *that* it is possible to navigate to a certain physical place. What we will argue below is that, in an enactive view of autism as a sensorimotor atypicality, what is basically in question is the *how* of achieving certain states in a physical (or more abstractly: state) space.

HIPPEA, or in full "High and Inflexible Precision of Prediction Error in Autism", is a conception of autism within the predictive coding theory. Predictive coding in cognitive science, roughly, is a theory on perception according to which the brain is essentially a prediction machine that aims to minimize the divergence between its internally generated prediction of the external world and the actual input (Hohwy 2013). In HIPPEA it is hypothesized that autism consists in a specific way of dealing with perceptual uncertainty. Following HIPPEA autism is an atypical way to process a difference (the "Prediction Error") between perceptual input and prior expectations. This atypicality then specifically consists in atypically giving a "High and Inflexible" weight to the difference between both (in other words. it consists in giving an atypical importance to "Precision"). The theory is phrased as a perceptual theory but this - certainly on an enactive reading, see further below - implicates action as well. Indeed perception has been already for a long time recognized as an active process (Gibson 1966) of mutual engagement with one's environment. In this case of atypical importance of precision, we would expect to see an atypical sensorimotor interaction with one's environment; an interaction more attuned to the actual specifics of the environment than to abstracting away from those specifics in order to flexibly switch from one abstract form of interaction to another. This then leads to a learning style that is at odds with the requirements of a typical environment. Given the aforementioned scaffolding, autistic developmental trajectories will - but *only* statistically speaking (Bervoets 2020) - lead to the emergence of social-cognitive elements of the DSM-5 (APA 2013). A developed example of this process related to atypical development of a self is worked out by Constant et al. (2018).

Let's now take a closer, enactive look at how atypical precision may skew variables which are associated with sensorimotor interaction. Specifically, in enactivism we should be wary of reifying an atypicality related to precision as part of a computational module in the brain. Instead, this atypicality is to be seen as capturing the specific way autistic bodies tend to covary with their environment[69]. The difference between typical and atypical precision *is* a

[69] For a complete argument to this effect see van Es (2020a; 2020b).

difference in such covariation. Autistic bodies tend to covary in a more precise way with their environments, picking up regularities in their environment in a more precise way. By that token they will tend to be less sensitive to social habits which rely on making abstraction from perturbations in the environment (which, said in passing, can be seen as an enactive reinterpretation of the terms "High" and "Inflexible" underlying the HIPPEA acronym). Coming back to the analogy we made above with respect to blindness, this atypicality is not a static feature of a 'modeler' or of 'a model' restricted to be in certain atypical states. Rather, it concerns the more nuanced, atypical manner of attuning to one's environment, i.e. an atypical way of navigating states afforded by one's environment.

Autistic atypicality is, then, maybe less 'visible' to external observers (noting that even blindness may be compensated for by for instance echolocation, therefore remain rather invisible, see Servick (2019)) but in itself very akin to classical sensorimotor differences such as visual impairment. Autistic people will tend to rely a lot more on current tangibles in their environment and less on carrying out previously internalized habits aligned with social conventions (for the enactive development of such a scaffolding, see: Krueger and Maiese 2018).[70] In a physical space metaphor: they rather tend to "feel" their way around in the here and now rather than "reenacting" behavior based on some abstract cues. This does not mean they can get to less (or indeed more) places. It just means they tend to get there in a different way (see also Hipólito and Hutto 2020).

This means that when we compare autistic and typical bodies, they navigate similar territory or state spaces in different ways. An enactive reading of HIPPEA, and by extension of the other putative differences in 'sensorimotor apparatus', leads to higher sensitivity to environmental perturbation. Seen from a neurotypical eye, autistic embodiment is correlated to 'overfitting' sensory data but from an enactive point of view it just is another, atypical, way of navigating one's immediate environment. Given an autistic body is more sensitive to (or deals more precisely with) its immediate environment, its actions depend more on the 'tangible' here and now than on an averaging out of such tangibles based on 'intangible' cues that have become

[70] There is a bit of complexity here. DSM-5 (APA 2013) also specifies Restricted and Repetitive Behaviors. This implies that autistic often display an insistence on sameness holding on to specific habits or repetitive patterns of behavior. This might at first glance seem at odds with our reading of HIPPEA which is focused on the immediate tangibility of the environment. However, in our view this can be seen as a developed coping strategy of autistics in making their environment more predictable so as to be less overburdened by spurious changes. We cannot develop this point further in the scope of this paper but see it as a crucial aspect of an enactive reading of HIPPEA in which such behavior can be explained without reference to inflexibility of internal representations. We thank Sander Van de Cruys for pointing this out to us. Whilst an important outstanding issue in autism research, we do not believe its resolution is critical for the argument made in this paper. We also refer to footnote 1 on hyper- and hypo-sensitivity for a similar connection between two apparently contradictory phenomena.

associated with certain expectations of default interaction patterns. Phrased like this one can see the link to other sensorimotor differences such as blindness: the difference with 'typical' agents lies in *how* certain states are reached and *not* in a principled restriction of *what* states can, ultimately, be reached. This fits with the atypicality view of autism, opposed to the deficiency view.

The skewing of how certain states are reached by an autistic body clearly has an impact on sensorimotor interactions with its physical environment. It will be more sensitive to triggers in such an environment, tending to covary with them in a more 'precise' way. This way it picks up on regularities and patterns a 'typical' body will simply not pick up on. At this junction there is neither value nor disvalue in this difference; it just is. However, if assessed from the habits formed by typical bodies or from the interaction patterns naturally sustained between typical bodies, autistic differences will tend to be perceived as flouting the conventions, or the norms, of 'typical' interaction or as some intrinsic 'deficiency'.[71] The relations between the individual differences, the emergence of shared norms and valuation are complex as can be gauged from the above: an increased sensitivity to one's physical environment may quickly be perceived as an insensitivity to one's social environment. In order to work out these relations and their link to interactional differences between autistic and neurotypical people, we will in the next section critically build on the enactivist framework of *Linguistic Bodies* (Di Paolo et al. 2018) to see how social and sensorimotor habits are mutually constrained. This provides us with the right foundations to develop our account of ethically overcoming these tensions by exploiting some intrinsic freedom created by the social space with respect to the physical space.

7.3 From sensorimotor atypicality to social normativity

In the enactive literature, Di Paolo et al. (2018) distinguish three distinct, yet intimately intertwined levels of autonomous organization relevant to an organism: the biological and sensorimotor levels of organization are tied to each individual, whereas the social level is intersubjective, and unique to each social encounter. These levels are hierarchically layered so that the biological level of organization lies at the foundation, and enables the formation of any other form of autonomy. *Autonomy* is defined as precarious operational closure (Di Paolo and

[71] That such a (dis)valuation of autistic difference as deficiency is compatible with enactive theorizing has been shown by Janna van Grunsven (2020). The view from 'typical bodies' can be compatible with the enactive views as developed by Gallagher (2004) and Hutto (2003) on autistic differences. The view from autonomous sustainability of typical interaction patterns is left open by the 'participatory sense-making' approach by Di Paolo, Cuffari and De Jaegher (2018). In the next section we critically build on the latter approach in order to capitalize on its strengths, at the same time avoiding this weakness.

Thompson 2014; Maturana and Varela 1980). This means that an autonomous system consists of a self-enabling process network, which captures *operational closure*, that is naturally inclined to disintegration, which implies *precariousness*. A process network is *self-enabling*, if each process in it enables or is enabled by at least one other process in the network. A process *enables* another process when its continuation aids the continuation of the other process; and is *enabled by* another process when the other process's continuation aids its own continuation. In this sense, the self-enabling nature of the network is what allows it to, albeit temporarily, resist the natural incline towards disintegration.

Sensorimotor autonomy does not concern the physical integration of the system, but instead its sensorimotor interactivity. Here too, it is thought, do we find clusters of activities or processes that are self-enabling. These autonomous sensorimotor activity clusters are also called *habits* (Di Paolo et al. 2017; Barandiaran 2017). Crucially, sensorimotor autonomy is irreducible to biological autonomy, and thus does not need to be in service of it either. Think of the many routines that populate our daily activities, many of them irrelevant to biological sustenance. Of course, the *restriction* of sensorimotor autonomy *can* have biological effects by way of stress, for example.

Social autonomy is intersubjective, and unique to each social encounter. This, following Di Paolo et al. (2018), does not explicitly concern each individual's autonomy, but instead the autonomy of the social interaction itself. In this sense, each social interaction is characterized as a self-enabling network (the 'self' here referring to the interaction, not to the 'selves' that are interacting) that, unless actively maintained, naturally inclines towards disintegration. What helps sustain an activity and what doesn't depends on the participants and the environment.

Arguably, not all three levels are actually associated with each organism. Some, the argument goes, seem to only have biological autonomy so that their sensorimotor interactivity *can* be reduced to, and is merely in service of, their biological autonomy, whereas others engage at least in some sensorimotor activities for the sake of it: think of a pigeon rolling in the snow, for example.

7.3.1 Levelled sensitivity

An autonomous system cannot be defined without at least implicitly defining the environment it is embedded in. This is because in continuously enabling its own structural integrity, maintaining its autonomy, a system necessarily interacts with its environment. This continuous interaction can *bias* an autonomous system to the interactional regularities that are encountered. Put differently, the interactional history of a system allows it to become attuned to the unfolding

interactions: to the particular ways that certain movements influence sensory signals (also called *sensorimotor contingencies*, see O'Regan and Noë 2001; Di Paolo et al. 2017). A *sensitivity*, then, is nothing more than a system's structural bias towards certain interactional patterns or regularities borne out of an interactional history, which includes biological, sensorimotor and social interactivity. This structural bias could be unpacked in terms of reliable covariance between the system and its environment (Bruineberg and Rietveld 2014; Hutto and Myin 2017; van Es 2020).[72] For two systems to *covary*, a change in one brings about a change in the other. *Reliable* covariance is then a covariation relation that does not have a one-to-one mapping, but implies covariance to a certain degree. There is no definite threshold for this, but in our case, minimally one may expect that the system's structural integrity is to be maintained for anything to count as reliable.

The sorts of interactional regularities that a system becomes sensitive to, depend of course on the interactions it is and has been able to engage in: its interactional history. This means that when considering sensitivities, we need to consider the system's *embodiment*, which determines the possible modes of interaction (Varela et al. 1991). After all, a system cannot become sensitive to environmental changes that its senses cannot pick up on one way or another. People, for example, cannot become sensitive to the ways particular winds affect one's flying capabilities, whereas birds are unlikely to become sensitive to a standardized pen grip. Furthermore, interaction only happens in an environment. The sorts of interactions available are thus also dependent on the environment one is in. Much like birds in a world that does contain pens, humans only a few centuries ago could not have become attuned to pen-writing activities either. Indeed, any activity is necessarily performed *by* a system, *in* an environment, thus *the possibilities for interactions* are determined by the system and its environment, and so *the interactional history* out of which interactional sensitivities are borne is, also, determined by both the system and its historical trajectory through an environment.

A sensitivity is a system's attunement to the particular way that certain movements influence the system in one way or another. The different autonomous levels of organization

[72] Di Paolo et al. (2018) and Di Paolo et al. (2017) take this a step further, and take the system's precarious autonomy to imply a normativity relative to each level of organization. More than sensitivity, this normativity is intended to account for a supposed 'judging,' or 'evaluating' capacity. Each system's biological autonomy thus implies an 'inherent normativity' which offers a standard of evaluation by which to judge the possibilities of interaction in the environment as either harmful, positive, or neutral. There is a debate in the enactivist literature with regards to whether this individualistic account of normativity betrays a hidden representationalist commitment (Hutto and Myin 2013, 2017). Our present argument does not hinge upon this debate. We can speak of interactional sensitivities obtained on an individual level, and introduce normativity on a social, intersubjective level, not perfectly in line with, but amenable to both takes in enactivism (Di Paolo et al. 2018; Di Paolo et al. 2017; Hutto and Myin 2013, 2017).

are influenced rather differently by one and the same motoric movement. Consider how, biologically, eating foods provides energy for continued sustenance, whereas sensorimotorically the foods' way of resisting the chewing pressure entails a certain texture, the influences on the tongue as it swirls and moves the food about entail a certain flavour (Noë 2004). Socially, thus intersubjectively, eating the food may prolong the interaction — not eating may be rude — while simultaneously imply a potential break — perhaps the interaction is over once the food is finished. The different ways particular interactions influence the respective autonomous levels implies that the developed sensitivities are also different. Whereas biologically the system may become attuned to, say, energy intake regularities such as the frequency or the amounts, sensorimotorically the system may become attuned to textures and flavours, the spaces food is eaten in, the tableware it is eaten out of or the tools it is eaten with. The individual system may also become attuned to social regularities particular to food-eating practices, such as the presence or absence of certain people, the appropriate manner of using the cutlery, and whether or not to engage in, say, conversation.[73]

Though conceptually we can divide the three levels, in practice they are intimately intertwined. We don't, e.g., *experience* multiple levels of organization, nor do different levels of embodiment refer to multiple physical bodies. There are however ways we can distinguish them in practice. Consider that someone may sensorimotorically be a bespectacled person, so that their spectacles are an inherent part of their 'sensorimotor self', whereas, biologically, the spectacles do not figure in the self-enabling network. The same could arguably be said of shoewear or even clothes. Moreover, we sometimes find ourselves in conflict, where the different interactional biases conflict, such as becoming peckish during an academic presentation so that biologically one may have become attuned to energy intake in the current circumstances, whereas socially eating behavior may be detrimental to the continued interaction. Conversely, each and every activity is always associated with all three levels of autonomy, to the extent the respective levels are associated with the system (Di Paolo et al. 2018). Every human activity minimally takes energy, is some sort of sensorimotor interaction in a world that is inherently social in nature. Though distinct, no level of autonomy ever 'acts

[73] There is a slight disparity here between the discussed levels of organization and their respective sensitivities. According to Di Paolo et al. (2018), biological and sensorimotor autonomy are both *individual*, yet social autonomy is inherently intersubjective. Following the in-text logic, this implies that the autonomy of the interaction, instead of the participating individuals, becomes biased towards certain interactional patterns. There may be something to this, as Di Paolo et al. (2018) show. Consider recurring fights with a friend that neither of you actually wants to have. It may be, indeed, helpful to think of this in terms of a bias of the social encounter itself, instead of the individual participants'. However, in this paper, we shall also be concerned with the *individual* interactional sensitivities pertaining to social encounters.

on its own'. It is then in this particular sense, that the interactional regularities that any system becomes attuned to are dependent on its levels of autonomous embodiment as well as the environment it figures in.

7.3.2 How autism levels up

Now that the basic framework is laid out, we can start looking at how the sensorimotor atypicalities common to autistics fit in. Briefly rehearsing, the interactional regularities one becomes sensitive to (or interactional sensitivities) are dependent on a system's interactional history. Organismic interaction depends on organismic embodiment as well as its environmental constitution. This means that the interaction changes whenever the organism or the environment changes; the organism comes to covary reliably with its environment.

The particular ways in which the organism has come to covary with its environment, has become sensitive to its environment, has historically interacted with its environment, can be described as a state space trajectory. Imagine a multi-dimensional state space, with one axis for each variable associated with the system.[74] The state of the system is then a multi-dimensional coordinate, with a particular value for each variable. Whenever a variable changes, the state of the system changes, pushing it to a different multi-dimensional coordinate. As an organismic system is constantly in flux, over time we will see the changing coordinates of the system as traversing a particular *trajectory* through the multi-dimensional state space.

Autism, we argue, is best taken to consist fundamentally in particular sensorimotor atypicalities. Within the state space formalism, this implies that certain sensorimotor-related variables in the state space may incline, or be skewed, to display particular values or dynamics.[75] This skewing is consistent throughout their developmental trajectory, and thus consistently influences their modes of environmental interaction. Consequently, it impacts interactions, over time it impacts interactional history, and thus interactional sensitivities. Sensorimotorically, an autistic will thus display an atypical state space trajectory, meaning an autistic becomes sensitive to different interactions and forms different habits. This may explain

[74] What exactly it means for a variable to be 'associated' with a system is an interest-dependent matter. If we are to model a person's kinetic dynamics when going down a slide on a playground, variables pertaining to their cycling habits are not considered 'associated' with the system for current purposes. However, when we are interested in modeling this person's risks for cardiovascular diseases, we may want to consider those. In this sense, a state space description of a real life system need not be complete to be scientifically valuable.

[75] Whether the inclination is visible in particular variables or dynamics depends of course on which variables were chosen to figure in the state space description. A system that accelerates strongly either skews to high values in the 'acceleration' variable or quickly rising dynamics in the 'movement speed' variable. It depends on how we set up the formalism.

certain common basic sensorimotor tendencies among autistics (Constant et al 2018.). Yet it underdetermines the *exact* influence it will have on any specific autistic individual. Each trajectory is of course not only dependent on a particular skewing of certain variables, but it is a holistic pathway: each and every aspect of the individual, as well as each and every aspect of their environment affects the trajectory in their own ways. As such, not only is each trajectory unique, but each autistic trajectory is also *uniquely atypical*. In this way, our model supports the large heterogeneity in the atypicalities usually associated with autism. Now we can see how this can 'scale up' to social interactivity (but see Gallagher 2017; Zahnoun 2018 for criticism on the need to 'scale up').

Per our explanation in Section 3.1, anything happening on the sensorimotor level of embodiment impacts the system's social interactional sensitivities. Sensorimotor atypicalities entail a mode of interaction with the world different from that conventionalized by interaction between typical people. As such, the social interactivity will be influenced by the autistic's sensorimotor atypicalities just as their picking up of objects is impacted by them. This starts at the very fundament of social development. Consider that joint attention schemes, for example, play out differently when one's relevant sensorimotor habits are mutually attuned than when they are not. Over time, this may result in atypical social habits, such as particular manners of speaking, body language or a preference for certain environments to engage in social interactions in. As such, consistent sensorimotor atypicalities will influence the social interactional regularities that the agent becomes sensitive to. Put differently, sensorimotor atypicalities entail atypical pattern-forming on a social level, which in itself implies a skewing in the interactional regularities that the person becomes attuned to on a social level. At this juncture, it is important to understand that autism is just one example of a possible skewing of a sensorimotor trajectory. Height, weight, or visual or acoustic acuity more broadly may just as well be atypical, and are accompanied by their own atypical habits both sensorimotorically and socially.

As such, placing the atypicality of autistics at a sensorimotor level allows for the wide variety of noted atypicalities in autistic social interaction (Jaswal and Akhtar 2019). One person's trajectory may have resulted in habits that eschew fantastical imaginings (see for example DSM-5 (APA 2013) or Curry and Ravencroft 2002), whereas another's interactional history may have resulted in habitual interaction in fantastical worlds (see for example Lyons & Fitzgerald 2012). However, they may, coincidentally, share a habit of avoiding eye contact (Valiyamattam et al. 2020). They also may not (Jones et al. 2017). This all depends on the particular state space trajectory of each individual, as it does for every person, autistic or not.

Moreover, this trajectory is a sum of the organism's state space *and* its environment's, which means that the same person will develop different habits if it were to have grown up in a different environment. This is to say that from that point on, their state space trajectories will come to vary wildly and they will grow up to be different persons altogether.

7.3.3 From individual sensitivity to social normativity

Individually, each organism has its own embodiment, accompanied by its own set of sensitivities on a biological, a sensorimotor and a social level. However, each social interaction also has its own autonomy, according to Di Paolo et al. (2018). Following the definition of autonomy above, this implies that each social interaction is composed of a self-enabling network that naturally inclines towards disintegration. This means that, if the processes in the network were to stop, the entire interaction would break. Only by continued processual activity can the whole network be kept active: a conversation cannot be sustained without continued interactivity.

Moreover, each social encounter is unique and entails its own state space description. The potential modes of interaction that could either sustain or break the intersubjective autonomous organization are thus also unique to each social encounter. Like the organismic case, what sustains or breaks a social interaction is dependent on the processes that make up the network itself, as well as the environment it is embedded in. Naturally there is some variability in participants' sensitivities: some people may be used to speaking at a particular volume that may seem too loud or too soft to others. Yet some environments also may afford speaking loudly better than others. The range of speaking volume which best sustains or breaks a particular social interaction is thus dependent on both the participants' sensitivities as well as the environment the interaction figures in. Indeed, contrary to the organismic case, the interactional processes that make up this network are composed of all participants', and thus depend, but cannot be reduced to, all the individual participants' own interactional sensitivities (Di Paolo et al. 2018).

This makes the interactional state space description an intersection of at least participants' individual interactional regularities and sensitivities, their environment, and their individual relation to the shared environment. An atypically tall person, for example, may be adjusted to looking slightly downward in social interactions, which means they may not be well attuned to social interactions with someone even taller. As it is for the atypically tall person, so it is for the autistic. This is to say that an autistic's own interactional history will in part determine what will sustain and what will break a social encounter they are engaged in, just as

much as their interactional partner's interactional history does. One may have developed a habit of not looking someone straight into the eyes during conversation. The co-participant may have developed a strong fondness of eye contact. This may create a tension in the social encounter that is to be overcome if the social encounter is to be sustained, but not unlike any other tension specific to each and every social encounter between two uniquely different individuals embedded in a unique environment.

It is this theme of overcoming differences that we take up in the next section where we link it to the ethical theme of inclusivity of diversity. What we have argued for in section 2 and 3 respectively is that 1) autism is fundamentally a gradual sensorimotor atypicality which branches out to social atypicalities, and that 2) overcoming of sensorimotor differences is, in general, inherent to any social interaction. Section 4 then infers that we *should* also overcome the social tensions flowing from sensorimotor differences associated with neurodiversity on the same grounds as furthering inclusivity of other diversities.

7.4 The ethical normativity of overcoming differences in degree

It is hardly controversial that differences in visual ability or height should not lead to exclusion of those on the extremes of our human spectrum from our community. However, we can also arrive at a principle of inclusion via classical ethical literature. Indeed, grounding of our moral stance in our social communicative practices is, for instance, given by Strawson (2008/1962) via his notion of 'reactive attitudes'.[76] Crucial in this leading ethical framework is negotiation of ethical interaction across individuals characterized by individual difference; see Wallace's (2019) development of a call to cosmopolitanism from reactive attitudes. In another Strawsonian interpretation, McGeer (2019) works out an account of inclusion of individuals based on the ability to *scaffold* towards social communicative practices where initial differences can be overcome. This is also in line with disability literature in general where inclusion is central, and this independent of meta-ethical debates. This is clear from, for

[76] The 'reactive attitudes' are attitudes of blame, resentment, forgiveness, gratitude etc. which we discern in the other and the other with us. They allow us to build practices in which we can hold each other responsible. There are two remarks to be made about our using this framework as an illustration of our argument. First, there is a lot of controversy on whether such an account does not risk excluding others based on the ability to recognize and show the reactive attitudes. It has been used as a way to rationalize either (partial) exclusion of autistics (Shoemaker 2015) or to positively argue for a 'normalization' of autistics to recognize conventional social cues (Richman and Bidshahri 2018). However, we use this theory for illustration precisely because of its ability to abstract away from sensorimotor (a)typicalities and its consequent bringing into focus the communicative interaction as such. Second, we use this theory, just like we used HIPPEA above, as a concrete illustration of our argument without it being a necessary premise in what we contend. Any ethical frameworks that can develop a need for inclusivity over and above individual differences will do to support our conclusion.

instance, Barnes' take on *The Minority Body* (2016), and neurodiversity literature specifically, as a particularly acute example we can take Milton's (2017) expression of the *double empathy problem*. In the latter, Milton argues that understanding autism as a deficit in social-cognitive ability actually is tantamount to a lack of empathy from the point of view of a neurotypical agent. In the former, Barnes argues that attributing issues to individual disabled bodies is, equivalently, an inability to see from the perspective of the typical bodies how those issues actually stem, at least for the largest part, from lack of societal concern for accessibility of minority bodies. Our argument can be seen as integrating these into a *Minority Brain* view (Bervoets and Hens 2020).

Now, as we have seen in section 3, the sustenance of social interactions is in a crucial way independent on specific underlying sensorimotor preferences. We do not have to react in a specific way to be able to signal, for instance, blame or gratitude. Just how blind and seeing people navigate - see our example in section 2 - the same physical space differently, tall people and people of average height will navigate differently in an interactional state in which they can recognize a reaction of blame or gratitude. The ethical call of inclusion is to separate the *how* we achieve an ethical interaction from the *that*, the possibility of reaching it. Here we need to clearly separate the unique trajectories each and every individual traverses in order to reach some point of social and ethical interaction and the fact *that* such points can be reached. In enactive terms: whilst everybody lays down a different path while walking (Thompson 2007), there is no reason why these paths may not intersect in participatory sense-making (Di Paolo et al 2018, specifically Chapter 8).

Let us make this more precise with respect to autism and autistic embodiment. Where autistics are sensorimotorically skewed to attune, literally more precisely, to the here and now of environmental perturbations this is not the case for the majority of (neuro)typical people. It is in this way understandable that the latter mistake their specific way of coming to reactive attitudes for minimal, or essential, preconditions of achieving them. Whether it is by shaking hands, recognizing specific facial expressions or in other ways abstracting from other environmental cues deemed inessential, typical people will tend to mistake a sensorimotor preference for certain habits in social interaction for the essence of that social interaction. Autistic people will nevertheless from childhood on have difficulty attuning to these habits *precisely* because of a statistical atypicality in the way in which they tend to covary with environmental cues. We have shown that this skewing of the variables associated with navigating the multi-dimensional state space should not be reified into stereotypes of social interaction, because it is but one of a great many historical factors that contribute to how unique

autistic individuals wind up trying to engage in social interactions optimized for a sensorimotor embodiment that is not theirs. In a society more forgiving of the preconditions of expressing reactive attitudes, they may more easily develop shared habits than in one that is less forgiving; and obviously the longer it takes to be recognized as one who is able to communicate, the more this part of one's history will further skew one's potential to bridge the gap between one's own sensorimotor preferences and social interaction which is deemed foundational for inclusion in our moral community. A specifically acute example of the overcoming of such gaps is given in the context of early language development by De Jaegher (Di Paolo et al. 2018, Chapter 8).

None of this is to say that these social, ethical, interactions are disembodied. Far from it, in line with Di Paolo et al. (2018), each human interaction is necessarily embodied. But, because of the relative autonomy of sensorimotor and social interactions we are not limited to a specific way, a specific habit, of sensorimotor interaction to achieve a specific, ethical interaction. Rather, by prescribing specific sensorimotor interactions as basic for ethical inclusion we violate the independence of the social and ethical field and - see our argument in the previous paragraph - risk to actively contribute to exclusion through 'ableist' bias. Indeed, there is a sense in which all of disability, at least the social model aspects of disability, can be seen as too prescriptive a notion of sensorimotor interaction that needs to underlie ethical interaction.

As is clear from the above, the overcoming of differences in degree (whether in visual acuity, height, culture, ...) is something we do daily among people. We recognize just societies by their capacity to overcome the superficial tensions that may rise from such differences, their tendency to embrace cosmopolitanism, per Wallace (2019). What we argued for is that the enactive construal of autism as an atypical embodiment leading to a specific skewing of sensorimotor preferences leads to recognizing it as one such difference in degree. Looking at how such difference is overcome in actual lived practice helps us explain the precise nature of this underlying difference and - consequently - may help us to create more inclusive conditions by relaxing sensorimotor requirements which underlie social interaction of (neuro)typical people. Clearly, there's a cost to cosmopolitanism on the side of typical people who need to adapt to new habits (or unlearn old habits) but we believe that not doing so (and consequently leave all the cost of adaptation to the atypical) is a case of bigotry. This does not mean that the final result will not be a compromise between a cost incurred on both sides - see the double empathy problem of Milton (2017) - but it does mean there is an ethical (as well as scientific) call to at least figure out what is the best compromise, even if this is a path we can only lay down in walking.

7.5 Conclusion

We have argued that an enactivist construal of autism entails an inclusionary approach in line with the neurodiversity tenets specifically and cosmopolitan ethics generally. That this is not self-evident is clear from other enactivist approaches to autistic embodiment, see Van Grunsven (2020), which start either from the neurotypical interactional habits (Hutto 2006; Gallagher 2003) or from the mere autonomy of specific sensorimotor interactions (Di Paolo et al. 2018). Our approach instead starts from a relative - and ethical - independence of social interaction and so leaves room for an inclusionary approach to autistic embodiment seen in its sensorimotor specificity. Basically we show that there are many ways to peel an orange and in our daily practice we overcome sensorimotor differences that unavoidably arise from different life histories. Autistic embodiment is just one of such bodily differences and understanding it by appeal to a good empirical theory of autism, such as atypical precision, allows us to direct our search for an increased mutual understanding. After all, we are all continuously changing bodies and are in need of being cut slack in order to maintain social, ethical, interaction.

In the process, we have also engaged with the enactivist literature. We have expanded enactivism to provide a space for speaking of individual social sensitivities, over and above the biological and sensorimotor individual sensitivities. This has allowed us to expand the merely descriptive enactivist analysis into the normative debate. This provides new tools for the ethical project of inclusion, by specifying particular social tensions that we should work to overcome, as we do already elsewhere. With this, we avoid slipping into enactivist developments of autistic embodiment that, ultimately, point back to basic capabilities and therefore deficiencies related to autism (Van Grunsven 2020). Furthermore, we connected the enactivist notion of *scaffolding* with its use in McGeer's work, allowing for an ethical shift of focus from the overt capacities *typically* used in social interactivity to the minimally - shared - ability allowing social interactivity at all (McGeer 2019; Krueger and Maiese 2018).

Crucially, the toolkit provided here can be put to use in other disability studies as well, such as Tourette Syndrome or ADHD, and may serve as a general framework from which to analyse and help resolve social tensions. In general, this can bring ethics and empirical science more closely together *via* enactivism. Yet we don't think a philosophical framework is the solution to inclusion; if this were the case the struggle for better inclusion of 'physical' disabilities would long be over. The solution lies in individuals 'actively' trying to overcome difference, hereby providing the evidence that this overcoming is possible and the difference not a difference in human quality but merely one of physical degree. Still, this framework can

be, in our opinion, a good support for such activism in showing that inclusion should not be a mere targeting of individual difference but requires an effort on both sides of the interaction.

References

American Psychiatric Association (2013). Diagnostic and Statistical Manual of Mental Disorders (DSM-5§). Washington, DC: American Psychiatric Association Publishing. doi: 10.1176/appi.books.97808904 25596

Barandiaran, X. E. (2017). Autonomy and enactivism: towards a theory of sensorimotor autonomous agency. *Topoi 36*, 409–430. doi:10.1007/ s11245- 016- 9365- 4.

Barnes, Elizabeth. (2016). *The Minority Body: A Theory of Disability.* Oxford University Press.

Baron-Cohen, S. (2000). Theory of mind and autism: a review. *Int. Rev. Res. Ment. Retard. 23*, 169–184. doi: 10.1016/s0074-7750(00)80010-5

Bervoets, Jo, and Kristien Hens. (2020). Going Beyond the Catch-22 of Autism Diagnosis and Research. The Moral Implications of (Not) Asking 'What Is Autism? *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2020.529193.

Bolis, D., Balsters, J., Wenderoth, N., Becchio, C., and Schilbach, L. (2017). Beyond autism: introducing the dialectical misattunement hypothesis and a Bayesian account of intersubjectivity. *Psychopathology 50*, 355–372. doi: 10. 1159/000484353

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

Chapman, R. (2020). The reality of autism: on the metaphysics of disorder and diversity. *Philos. Psychol. 33*, 799–819. doi: 10.1080/09515089.2020.17 51103

Constant, A., Bervoets, J. Hens, K., and Van de Cruys, S. (2018). Precise Worlds for Certain Minds: An Ecological Perspective on the Relational Self in Autism. *Topoi*. https://doi.org/10.1007/s11245-018-9546-4.

Curry, G. and Ravenscroft, I. (2002). *Recreative Minds*. Oxford University Press, Oxford. doi:10.1093/acprof:oso/9780198238089.001.0001

De Jaegher, H. (2013). Embodiment and sense-making in autism. *Front. Integr. Neurosci. 7*:15. doi: 10.3389/fnint.2013.00015

Di Paolo, E., & Thompson, E. (2014). The enactive approach. In L. Shapiro (Ed.), *Routledge handbooks in philosophy. The Routledge handbook of embodied cognition* (p. 68–78). Routledge/Taylor & Francis Group

Di Paolo, E., and De Jaegher, H. (2012). The interactive brain hypothesis. *Front. Hum. Neurosci. 6*:163. doi: 10.3389/fnhum.2012.00163

Di Paolo, E., Buhrmann, T., Barandiaran, X., (2017) *Sensorimotor Life: An Enactive Proposal*. Oxford University Press

Di Paolo, E., Cuffari, E. C., & De Jaegher, H. (2018). *Linguistic Bodies: The Continuity Between Life and Language.* Cambridge, MA, USA: MIT Press.

Frith, U. (1996). Cognitive explanations of autism. *Acta Paediatr. 416*, 63–68. doi: 10.1111/j.1651-2227.1996.tb14280.x

Fuchs, T., and De Jaegher, H. (2009). Enactive intersubjectivity: participatory sense-making and mutual incorporation. *Phenomenol. Cogn. Sci. 8*, 465–486. doi: 10.1007/s11097-009-9136-4

Gallagher, S. (2004). Understanding Interpersonal Problems in Autism. *Philosophy, Psychiatry, and Psychology 11* (3):199-217.

Gibson, J. J. (1986). *The Ecological Approach to Visual Perception*. Psychology Press , London.

Happé, F., and Frith, U. (2006). The weak coherence account: detail-focused cognitive style in autism spectrum disorders. *J. Autism Dev. Disord. 36*, 5–25. doi: 10.1007/s10803-005-0039-0

Hens, K., and Langenberg, R. (2018). *Experiences of Adults Following an Autism Diagnosis*. Cham: Springer. doi: 10.1007/978-3-319-97973-1

Hipólito, I, Hutto, D, & Chown, N. (2020). Understanding autistic individuals. Cognitive diversity not theoretical deficit. In Rosqvist, H. B., Chown, N., and Stenning, A (eds), *Neurodiversity Studies. A New Critical Paradigm.* Routledge, London. https://doi.org/10.4324/9780429322297

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Hutto, D. D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Mit Press.

Hutto, D. D., & Myin, E. (2017). *Evolving enactivism: Basic minds meet content*. MIT press.

Hutto, Daniel D. (2003). Folk psychological narratives and the case of autism. _Philosophical Papers_ 32 (3):345-361.

Jaswal, V., & Akhtar, N. (2019). Being versus appearing socially uninterested: Challenging assumptions about social motivation in autism. *Behavioral and Brain Sciences, 42,* E82. doi:10.1017/S0140525X18001826

Jones, R. M., Southerland, A., Hamo, A., Carberry, C., Bridges, C., Nay, S., Stubbs, E., Komarow, E., Washington, C., Rehg, J. M., Lord, C., & Rozga, A. (2017). Increased Eye Contact During Conversation Compared to Play in Children With Autism. *Journal of Autism and Developmental Disorders*, *47*(3), 607. https://doi.org/10.1007/s10803-016-2981-4

Krueger, Joel & Maiese, Michelle (2018). Mental institutions, habits of mind, and an extended approach to autism. *Thaumàzein 6*:10-41.

Lyons V., and Fitzgerald M., (2012). Critical Evaluation of the Concept of Autistic Creativity, In Fitzgerald, M. (ed) *Recent Advances in Autism Spectrum Disorders - Volume I*, IntechOpen. doi: 10.5772/54465

Markram, K., and Markram, H. (2010). The intense world theory – a unifying theory of the neurobiology of autism. *Front. Hum. Neurosci. 4*:224. doi: 10.3389/ fnhum.2010.00224

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and Cognition—The Realization of the Living*, ser. Boston Studies on the Philosophy of Science. Dordrecht, the Netherlands.

McGeer, V. (2007). Why neuroscience matters to cognitive neuropsychology. *Synthese 159*, 347–371. doi: 10.1007/s11229-007-9234-1

McGeer, V. Scaffolding agency: A proleptic account of the reactive attitudes. *Eur J Philos.; 27*: 301– 323. https://doi.org/10.1111/ejop.12408

Milton, D. (2017). *A Mismatch of Salience: Explorations of the Nature of Autism from Theory to Practice*. West Sussex: Pavillion.

Mottron, L., Dawson, M., Soulières, I., Hubert, B., and Burack, J. (2006). Enhanced perceptual functioning in autism: an update, and eight principles of autistic perception. *J. Autism Dev. Disord. 36*, 27–43. doi: 10.1007/s10803-005- 0040-7

Noë, A. (2004). *Action in perception*. MIT press.

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

Palmer, C. J., Paton, B., Kirkovski, M., Enticott, P. G., & Hohwy, J. (2015). Context sensitivity in action decreases along the autism spectrum: a predictive processing perspective. *Proceedings of the Royal Society B: Biological Sciences, 282*(1802), 20141557

Pellicano, E. (2013). Sensory symptoms in autism: a blooming, buzzing confusion? *Child Dev. Perspect. 7,* 143–148. doi: 10.1111/cdep.12031

Richman, KA, Bidshahri, R. Autism, theory of mind, and the reactive attitudes. *Bioethics. 32*: 43– 49. https://doi.org/10.1111/bioe.12370

Van de Cruys, S., Kelsey, P. & Hohwy, J., (2019): Explaining hyper-sensitivity and hypo-responsivity in autism with a common predictive coding-based mechanism, *Cognitive Neuroscience*, DOI: 10.1080/17588928.2019.1594746

Servick, K., (2019, October 1). *Echolocation in blind people reveals the brain's adaptive powers*. Science Mag (retrieved 13th of June 2020), https://www.sciencemag.org/news/2019/10/echolocation-blind-people-reveals-brain-s-adaptive-powers

Shoemaker, D. (2015). *Responsibility from the Margins*, Oxford University Press

Strawson, P. F. (2008). *Freedom and Resentment and Other Essays.* Routledge. https://doi.org/10.4324/9780203882566.

Valiyamattam, G. J., Katti, H., Chaganti, V. K., O'Haire, M. E., & Sachdeva, V. (2020). Do Animals Engage Greater Social Attention in Autism? An Eye Tracking Analysis. *Frontiers in Psychology*, *11*, 1.

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., et al. (2014). Precise minds in uncertain worlds: predictive coding in autism. *Psychol. Rev. 121*, 649–675. doi: 10.1037/a0037665

van Es, T. (2020a). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior*. https://doi.org/10.1177/1059712320918678

van Es, T. (2020b) Minimizing prediction errors in predictive processing: from inconsistency to non-representationalism. *Phenom Cogn Sci 19***,** 997–1017. https://doi.org/10.1007/s11097-019-09649-y

van Grunsven, J. (2020). Perceiving 'Other' Minds: Autism, 4E Cognition, and the Idea of Neurodiversity. *Journal of Consciousness Studies 27* (7-8):115-143

Varela, F. J., Thompson, E., and Rosch, E. (1991). *The embodied mind: cognitive science and human experience*. Cambridge, MA: MIT Press.

Wallace, R. (2019). The Moral Nexus. PRINCETON; OXFORD: Princeton University Press. doi:10.2307/j.ctv3znwhn

Zahnoun, F. (2019) On representation hungry cognition (and why we should stop feeding it). *Synthese*. https://doi.org/10.1007/s11229-019-02277-8

Zaidel, A., Goin-Kochel, R. P., & Angelaki, D. E. (2015). Self-motion perception in autism is compromised by visual noise but integrated optimally across multiple senses. *Proceedings of the National Academy of Sciences, 112*(20), 6461-6466

8 Between pebbles and organisms: Weaving autonomy into the Markov blanket[77]

Authors

Thomas van Es 1

Michael D. Kirchhoff 2

1 Centre for Philosophical Psychology, Department of Philosophy, Universiteit Antwerpen, Belgium

2 School of Liberal Arts, Faculty of Arts, Social Sciences and the Humanities, University of Wollongong, Wollongong, Australia.

Abstract

The free energy principle (FEP) is sometimes put forward as accounting for biological self-organization and cognition. It states that for a system to maintain non-equilibrium steady-state with its environment it can be described as minimising its free energy. It is said to be entirely scale-free, applying to anything from particles to organisms, and interactive machines, spanning from the abiotic to the biotic. Because the FEP is so general in its application, one might wonder whether this framework can capture *anything specific* to biology. We take steps to correct for this here. We first explicate the worry, taking pebbles as examples of an abiotic system, and then discuss to what extent the FEP can distinguish its dynamics from an organism's. We articulate the notion of 'autonomy as precarious operational closure' from the enactive literature, and investigate how it can be unpacked within the FEP. This enables the FEP to delineate between the abiotic and the biotic; avoiding the pebble worry that keeps it out of touch with the living systems we encounter in the world.

8.1 Introduction

The free energy principle (FEP) is a *principle first* approach to what it takes for a system to exist. Rather than empirical investigation, the FEP starts from a mathematical *principle* that a system is thought to conform to if it exists. Indeed, FEP researchers seek to provide a general theory unifying biology and cognitive science formulated almost entirely from mathematical principles in physics and information theory (see e.g., Friston 2010 2013; Hohwy 2020; Kirchhoff et al. 2018; Linson et al. 2018; Ramstead, Kirchhoff, Friston 2019). The ambition is to secure a definition of existence by appealing to constructs in physics and information theory, and then employing those constructs to derive a principle of self-organization and cognition (Friston 2019; Hesp et al. 2019). In a nutshell, the FEP states that a system that maintains non-

[77] The text of this chapter has been accepted for publication as van Es, T., Kirchhoff, M. (2021) Between pebbles and organisms: Weaving autonomy into the Markov blanket. *Synthese*

equilibrium steady-state (NESS) with its environment can necessarily be cast as minimising free energy.[78] This particular observation can consequently be exploited to show a wide variety of interesting relations to hold between a NESS system and its environment.

Yet the FEP's mathematical toolkit is not only applicable to living systems. It is said to be entirely *scale-free* in its applicability. That is, it is intended to apply to any system able to maintain its organisation despite tendencies towards disorder: from chemotaxis in cells (Friston 2013; Auletta 2013), neuronal signalling in brains (Friston et al. 2017; Parr & Friston 2019), tropism in plants (Calvo & Friston 2017), synchronised singing in birds (Frith & Friston 2015) to decision-making and planning in mammals (Daunizeau et al. 2010; Friston 2013; Williams 2018). It has also been applied to model adaptive fitness over evolutionary timescales by casting evolution in terms of Bayesian model optimisation and selection (Campbell 2016; Hesp et al. 2019). However, this widespread applicability of the FEP can be taken as a fault, rather than an advantage.

Indeed, there is a general concern about the FEP's ability to speak to the essential organizational dynamics of biology, because it can seem utterly disconnected from biology. More specifically, the FEP is sometimes considered incapable of uniquely addressing the organisational dynamics of living systems (van Es 2020; Colombo and Wright 2018). Because the FEP implies an entirely scale-free dynamics in which *any* self-organising NESS system can be cast in terms of self-evidencing, some worry that this particular view cannot capture the specific details of biological organisation that is of interest to the biological sciences. If true, this undercuts the grand unifying ambitions of many FEP researchers.

We address this worry here. We start by rehearsing the basic tenets of the FEP, with particular focus on the Markov blanket formalism and how it relates to Bayesian inference (sect. 2). We proceed to explicate the aforementioned worry by considering the application of the FEP formalism to a pebble and discuss how the FEP seems to fall short in delivering the tools to distinguish pebbles from organisms (sect. 3). *Prima facie,* its scale-free applicability makes it seem like it is unable to carve any interesting joints between the abiotic and the biotic, which would hinder the prospect of a FEP biology. Kirchhoff et al. (2018) make an initial attempt to address this problem, suggesting that *autonomy* is what distinguishes living from

[78] The term 'non-equilibrium steady-state' refers to *self-sustaining* processes in a system requiring input and output to avoid relaxing into *thermodynamic equilibrium* (= systemic decay/death). It is important to mention here that the notion 'steady-state' in non-equilibrium systems is an approximation to some specified duration of time - e.g., circadian rhythms over a 24 hour clock cycle or the homeostatic processes involved in maintaining on average and over time a specific body temperature. So strictly speaking, biological systems are not in steady states; rather, to say that a system is in a steady-state, X, at a particular time, is effectively to say that the probability density over the system's states during some period of time was X.

non-living systems. The overarching claim there is that autonomy is the capacity of a system to modulate its exchange with its environment. Here we supplement this initial treatment. We first look at 'autonomy' from an enactive viewpoint (sect. 4). We then sketch the contours of how the notion of 'autonomy' from the enactive literature could be emulated with the tools available to the FEP formalisms. This allows us to understand what constitutes an autonomous system rather than merely using the notion of autonomy as a mark by which to delineate life from no-life (sect. 5).

8.2 Markov blankets, free energy and Bayesian inference

The FEP speaks to what characteristics a system must exhibit for it to exist (Friston 2013). Its basic premise is that any random dynamical system "that possesses a Markov blanket will appear to actively maintain its structural and dynamical integrity" (Friston 2013, p. 2).

A Markov blanket is a statistical separation of states that is applicable to any thing that exists (Hipólito 2020). It is a set of blanket states that separates a system's internal states from external states (Pearl, 1988; Beal 2003). The blanket states shield (in a statistical sense) internal from external states, and vice versa. They can be partitioned into sensory states and active states. Sensory states capture the influence of external states on internal states. Active states capture the influence of internal states on external states. Intuitively, any thing can be separated statistically from that which it is not (Palacios et al. 2020).

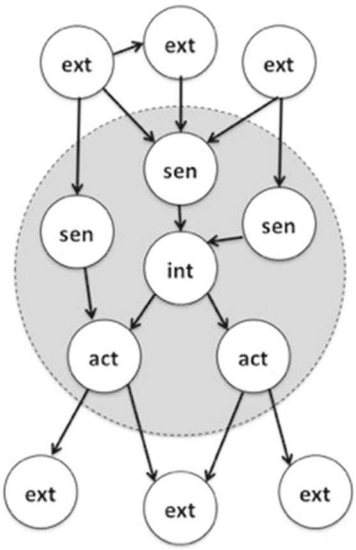

Figure 1 is a schematic representation of a Markov blanketed system. The gray circle delineates the Markov blanketed system that separates internal states (int) from external states (ext). The blanket states, sensory states (sen) and active states (act) are displayed as surrounding the internal states. The arrows depict modes of influence. External states influence only one another or sensory states,

while influenced only by active states or one another. Internal states are influenced only by sensory states, influencing only active states. In terms of modes of influence, internal states are separated from external states. (figure taken from Bruineberg, et al. 2018).

In this statistical formulation, the separation between internal and external states implies that these states are *conditionally independent*, given the states that comprise the Markov blanket. If we want to figure out the *external* states and we *know* the values of the *blanket* states, *knowing* the values of the *internal* states *will not offer additional predictive value*, and vice-versa. This is so by definition, because the blanket states already capture any possible influence the internal states could have on the external states. A brief example may clarify this. Say you observe that it is cold. This could be either due to an open window or to an air conditioning system that is set too strong. If you would observe that, say, the air conditioning is set excessively high, the observation that it is cold now does not offer further information with regards to whether or not the window is open. That is, in this case, the observed cold and the open window are conditionally independent, given that the air conditioning is on blast (Kirchhoff and Kiverstein 2019; Beal 2003). In terms of the Markov blanket formalism, the observed cold could be cast as the internal states, the state of the window could be formalized as the external states with the states of the air conditioning serving as the blanket states. This example is important because it indicates the widespread applicability of the formalism. Indeed, it is not necessarily obvious to associate the boundaries induced by the Markov blanket with physical boundaries, though it does seem to lend itself well to this particular application. We should nonetheless remain wary about overstating the implications of this *statistical* partitioning of states when considering its application onto other systems (van Es 2019, 2020).

Now that it is clear what a Markov blanket is, we can delve into its relation with free energy minimisation and Bayesian inference. This is a technical story. According to the Second Law of thermodynamics, the *entropy* of any closed system increases indefinitely over time. Any system that exists, or any Markov blanketed system that retains its structural integrity over time, seems to temporarily slow down the increase of entropy for as long as it remains intact (Friston 2012, 2013, 2019; Schrödinger, 1944). Of course, any such 'resistance' is only temporary, as entropy increases upon disintegration, which, in the case of biotic systems, means death.

For any such system, you can establish a multi-dimensional state space with as many dimensions as there are variables represented in the state space. Each point in the state space

corresponds to a unique intersection of values for each variable. In this state space, you can mark a bound of states within which a system can remain intact, outside of which it cannot (Friston, 2012, 2013). For as long as the system remains intact, the system will continuously 'revisit' the states within this bound. This is so by definition, as we define the bound by the range of values within the system remains intact. With regards to organisms, the viable bound differs per species: humans remain intact under quite different circumstances than fish do, for example. Insofar as this bound counts as a description of the states in which the system can be found when alive, it is also considered to be a mathematical description of a phenotype (Friston 2013; Kirchhoff et al 2018).[79] On average and over time, any living system is thus likely to be found within the bound of viable states and unlikely to be found outside of it. That is, we may *expect* a system to be within a bound of states that it typically remains within *on average* (Friston, Wiese, Hobson 2020). This implies a probability distribution that can be laid over the state space so that each state is assigned a probability value (Ramstead et al. 2019; Corcoran et al. 2020; Friston 2013). At any given time the system is encountered, it is highly likely to occupy a state within the viable bound, and highly unlikely to occupy a state outside of this. This means that states within the bound are considered high-probability states, whereas states outside of it are considered low-probability states.

Furthemore, if a system's *internal* states remain within a particular range, this must mean that the *influences* on those states are similarly bounded. An example should clarify this. Consider an egg and spoon race. An egg-and-spoon runner will need to ensure that the influence on the egg of their running the race remains within certain bounds, lest the egg move out of the spoon and break. Let us apply the Markov blanket partitioning method. We shall take the internal states here to be the egg's, and the influences it receives via the spoon shall be the sensory states, the runner is here the environment impacting on the spoon and comprises the external states. This means that an egg-and-spoon runner can be cast as keeping a *tight bound on the sensory states* of the Markov blanketed egg for as long as it remains in the spoon.

As it is for the egg-and-spoon runner, so it is for *any* system that remains intact over time. Relative to the viable bound of the internal states of the system, then, we can also establish *a state space for the sensory states* within which the system can remain intact, outside of which it cannot. Here too, we can determine a probability distribution where states within the bound are ascribed high probability, those outside of it are ascribed low probability. This is a probability distribution *over external states*, as it relates to the influences on the internal states

[79] See Colombo and Wright (2018) for criticism on the viability of this application onto an organismic system.

by the external states. In other words, it defines the possible external states that there could be relative to the internal states, given that the internal states remain within the viable bound. Of interest here is that the internal states themselves provide all we need (the 'sufficient statistics') to compute the probability distribution over the external states. As such, by knowing the viable bound of the internal states, we can compute the viable bound of the system's sensory states.

Further, in Bayesian probabilistic theory, *surprise* is a quantity defined as the *improbability* of a particular state (Shannon, 1948). If the surprise of sensory states (or 'sensory surprise', not to be confused with agent-level surprise with regards to an unexpected sensation) is high, the sensory states currently occupy a low probability area in the state space. As low-probability states are those that endanger the system's structural integrity, surprise is kept low, or minimized, as long as the system remains intact. However, sensory surprise is a probabilistic measure of sensory states. The entire state space of sensory states includes all possible modes of influence the external states could possibly exert on the internal states. This is, in principle, an infinite set. Computing sensory surprise directly is thus intractable (Friston 2009).

This is where (variational) *free energy* comes in. Free energy, in the statistical usage of the term, is a functional of the internal and sensory states a system is in.[80] In this case, free energy is thus, more specifically, the function of a function of the sensory states that is parameterized by the internal states. Because of this, the value of free energy limits the possible values of the internal and sensory states. To see why, consider a solution to a simple summation problem in arithmetics, say it's 15, and the terms of the equation are non-negative. This means that none of the terms of the problem can exceed the value of 15. Minimizing the value of free energy, then, minimizes an upper bound on the probability of sensory states. This ensures that sensory states remain in high-probability areas in the state space, which in turn implies that sensory surprise is minimized. Minimizing free energy can thus be seen as approximately minimizing the otherwise intractable value of sensory surprise (Friston and Stephan 2007). Moreover, as free energy is a function of only the internal and sensory states, it is in principle computable (Kiebel, Daunizeau and Friston 2008; Friston and Ao 2012).

In Bayesian probability theory, *negative surprise* is equivalent to Bayesian model evidence. *Minimizing* surprise thus *maximizes* Bayesian model evidence. The process by which Bayesian model evidence can be maximized is called *Bayesian inference*, sometimes referred to as self-evidencing (Friston, Killner, Harrison 2006; Hohwy 2016). Bayesian inference then refers to the particular way a probability distribution needs to be updated in light of new

[80] A functional is a function of a function.

evidence (Beal 2003). Bayesian inference describes the permissible 'moves' one can make in the formal system of Bayesian probability theory. We can now see that for any system to remain intact over time, its entropy needs to be minimized on average over time, which means expected free energy needs to be minimized, which in turn implies the minimization of sensory surprise, which is done by way of a formal operation called Bayesian inference.

The above story is employed in the FEP as a mathematical description of the homeostatic processes of biotic systems (Friston 2013). This works, very roughly, as follows. In the Markov blanket formalism, the Markov blanket is thought to carve out ontological joints: the internal states map onto the organism itself, and the external states map onto the environment (Kirchhoff and Kiverstein 2019). The partitioning blanket states map onto the organism's modes of interaction so that sensory states are associated with sensory receptor activity, and active states are associated with the system's influence on its environment, such as action. It remains a current debate to what extent this application of the Markov blanket should be taken literally or instrumentally (van Es 2020; Bruineberg et al. 2020; Hohwy 2016). In this paper, we will remain neutral in this debate, and instead explore only what can be done within the formalism, regardless of how it may or may not be implemented in any real system.

In a realist interpretation, to 'engage' in Bayesian inference is considered a fundamental aspect of life, as without it, the organism would go outside of its viable bounds. This is called *active inference*, and is thought to account for both action and perception by the same guiding principle (Friston 2013). The probability distributions are embodied and/or encoded by the organism (and/or the brain). They are to be manipulated, updated and leveraged by the organism. Through active inference, the organism updates the probability distributions in the face of newfound evidence, and uses this to infer action policies for its interaction with the world. Long term activities are thought to require counterfactual inference, which is associated with the minimization of *expected free energy* or free energy on average over time (Corcoran et al. 2020). Rather than updating the probability distribution to remain within its viable bounds, this should be seen as the inference of a possible trajectory through the state space conditioned on bodily movement. This allows the organism to adapt to environmental fluctuations. After all, the distribution of states within which an organism can remain alive cannot be simply 'updated' when confronted with an environment likely to push the system

outside of viable bounds. Active inference thus plays a central role in the realist FEP story of biological systems.[81]

8.3 Pebble meets Markov blanket

One person's meat is another person's poison: the scale-free applicability of the FEP's Markov blanket formalism may be taken as a vice, rather than a virtue. In this section we take up a specific challenge to the FEP that flows from what seems like an overly generous application of the FEP formalism to a wide variety of phenomena: the pebble challenge. It challenges the FEP's ambitions to describe the organizational dynamics of life precisely because its mathematical formalisms apply equally well to pebbles, and other abiotic systems as they do to biotic ones. One might therefore worry that the FEP fails to say anything specific about biology, unless characteristics we take to be specific to biology are not so specific at all. We describe this challenge in more detail now.

Friston & Stephan (2007) anticipates this kind of challenge to the FEP. They ask, "What is the difference between a plant [a biotic system] and a stone [an abiotic system]?" (2007, p. 422) They say that the plant "is an open non-equilibrium system, exchanging matter and energy with the environment, whereas the stone is an open system that is largely at equilibrium" (2007, p. 422). There is something to this initial observation. Plants are open systems, i.e., energy and mass can flow between the system and its surroundings. The same, of course, can be said of stones as environmental forces impinge on their surface area, and their own existence influences their environment by, say, releasing heat during the day, or altering pathways for organisms (Olivotos & Economou-Eliopoulos 2016). At first glance, it thus seems that the FEP applies in the same way to stones, plants and humans.

The FEP (as we saw above) starts from the simple observation "that *for something to exist* it must possess (internal or intrinsic) states that can be separated *statistically* from (external or extrinsic) states that do not constitute the thing" (Friston 2019, p. 4, emphases added). This Markov blanket formulation would apply to a pebble as follows. The Markov blanket defines the conditional independencies between two sets of states: the system and the environment. Pebbles are composed of minerals with different properties, lattice structure,

[81] The extent to which this story should be taken in a realist sense so that each biotic system literally performs advanced statistical operations, or in an instrumentalist sense so that each biotic system's interactional dynamics merely correspond to (or 'instantiate') the dynamics described in Bayesian inference is still debated (van Es 2020; see Ramstead, Kirchhoff, Friston 2019; Corcoran et al. 2020). A discussion of this debate is outside the scope of this paper, and it is unnecessary for our current purposes.

hardness and cleavage. We can associate these variables as the internal states comprising the system. On shingle beaches, the second set of states (the environment) would be other pebbles, and so on. In rivers, the water could be cast as the external states. As seen in Section 2, it is possible to cast a spatial boundary for anything that exists in terms of a Markov blanket (Friston 2013). The pebble has a clear boundary separating internal states and external states. The sensory states of a pebble can be associated with the effects of external causes of its boundary - stressors such as pressure, temperature and so on. Its active states would correspond to how the pebble effects external states - e.g., via release of heat back into the environment. The Markov partitioning rule governing the relation between states dictates that external states act on sensory states, which influence, but are not themselves influenced by internal states. Internal states couple back to external states, via active states, which are not influenced by external states (Palacios et al. 2020). Given that the Markov blanket formulation for a pebble is possible, it follows that internal pebble states are conditionally independent of external states in virtue of the Markov blanket states.[82]

What does this mean for our FEP analysis of the pebble, given what we have seen in Section 2? Under the FEP, the mere presence of a Markov blanket implies that internal states can be understood as if they minimise the free energy over the states that make up their Markov blanket. Technically, since minimising free energy is the same as performing approximate Bayesian inference, it follows that one can associate the internal pebble states (and its blanket states) with Bayesian inference. As such, it seems that if (1) anything that exists over time can be described in terms of a Markov blanket which implies that expected free energy is minimized by way of Bayesian inference, and (2) pebbles exist, then (3) pebbles can be described as having a Markov blanket, whose dynamics will appear as though they minimize free energy by way of Bayesian inference. The formalisms of the FEP that we employed here therefore seem too general to distinguish between pebbles and organisms. Below we will discuss what is needed for a formalism to properly address autonomy in Section 4. In Section 5 we will see how FEP's toolkit can be leveraged to make a headway in providing a principled distinction between pebbles and organisms.

[82] We could, for example, determine the surface molecules of the pebble to be sensory states, adjacent molecules to be active states, and the remainder of the pebble's molecules to be internal states, with the environment cast as external states. The molecules we cast as active states are then shielded from influence of the external states, while still able to influence the external states, though vicariously through sensory states. Of course, a pebble is merely an example and this could apply to many abiotic systems. Thanks to an anonymous reviewer for pointing this out.

8.4 Autonomy meets pebble

The pebble challenge need not be a knockdown argument against the ambitions of the FEP to address biology and cognitive science. Here we consider a possible reply to it. Our agenda will be to introduce the notion of *autonomy* from enactive philosophy of cognitive science.[83] Kirchhoff et al. (2018) appeal to this notion in order to distinguish between *mere active inference* and *adaptive active inference*. The former can be shown to apply to abiotic systems such as pebbles (from above) and the generalised synchrony induced in coupled pendulum dynamics. Adaptive active inference is introduced to make sense of the idea that living organisms are able to actively change or modulate their sensorimotor coupling to their environment - which is needed to actively monitor and predict changes to perturbations that challenge homeostatic variables, which may, sometimes, go out of bounds. However, the modulation of sensorimotor coupling is merely a (contingent) feature of an autonomous system. Operational closure and precariousness jointly define autonomy. We build on Kirchhoff et al's (2018; see also Kirchhoff & Froese 2017) argument by showing how autonomy is underwritten by the concepts of operational closure and precariousness (cf. Di Paolo & Thompson 2014).

8.4.1 Operational closure and precariousness

Operational closure is central to the conceptualisation of autonomy (Di Paolo & Thompson 2014). It is characterized as a form of *organization* in the sense that it specifies the particular way any system's component parts are organized in relation to one another. By specifying the organized 'unity' (the system) via this formalism, we also implicitly define its environment. Furthermore, by defining the system and its environment, we also specify the boundary through which the system interacts with its environment (Beer 2004, 2014; Maturana and Varela, 1980).

A system is operationally closed if the processes that make up the system constitute what is known as a self-enabling network. This means that each of the network's processes enables and is enabled by at least one other process in the network. It is empirically possible to determine whether any particular system is operationally closed by mapping out the causal

[83] Autonomy is a central theoretical construct of the enactive approach to life and mind (Varela, 1979; Varela et al., 1991; Thompson 2007; Di Paolo & Thompson 2014; Di Paolo et al. 2017). Enactivism is a theoretical framework with roots in theoretical biology, dynamic systems theory, and phenomenology. In enactivism, the notion of autonomy as operational closure has received special attention in attempting to unearth the self-organisational dynamics essential to life. Yet the literature so far has fallen short of construing operational closure in terms of the FEP's conceptual toolkit. Here we will make a first attempt at conceiving of an operationally closed system as being composed of a network of Markov blanketed systems that stand in a mutually enabling relation to one another.

processes relevant for the system and how they relate to one another. In particular, one must look for *enabling* relations. Any one process is said to enable another process if its continuation is partly or wholly constitutive of the enabled process. To explain how this works, it may help to look at the diagram of an operationally closed system (Figure 2 below). In this toy system, we distinguish five component processes: A, B, C, D, and E represented as nodes in the figure. The arrows between them represent enabling relations, so that A can be seen to enable process B. Following the arrows, we can identify a closed loop in the enabling relations pertaining to processes A, B, and C. This means that the continuation of A enables the continuation of B, which enables the continuation of C, which comes full circle and enables the continuation of process A: the ABC network is thus *self-enabling*. But what about processes D and E? E can be seen to enable process A, yet remains outside of the network as it is not enabled by a process in the network. D, on the other hand, *is* enabled by a process in the network, but doesn't loop back and enable a process in the network itself. This is why ABC can be identified as a self-enabling network, while D and E fall outside the boat.

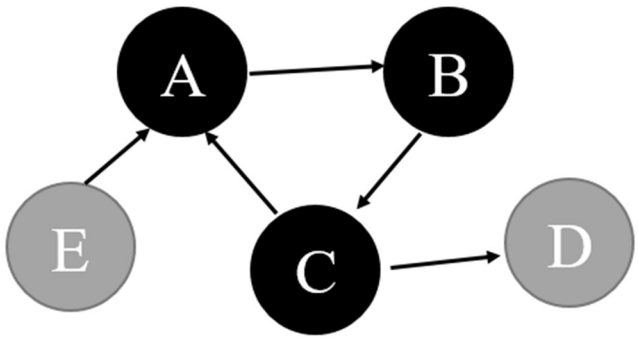

Figure 2, a diagram displaying an operationally closed network of enabling relations. Each node in the figure represents a process in the network, and each arrow represents an enabling relation. The operationally closed network is marked by the black nodes; processes outside the operationally closed system are marked in grey. Each node that is part of the operationally closed network is marked by having at least one outgoing and one incoming arrow from another node in the operationally closed network as is described in-text (inspired by Di Paolo and Thompson 2014).

Precariousness signifies a natural inclination to decline. In Figure 2 above, for example, process A is precarious if it would cease were it not enabled by E and C. It may be that not each enabling process is *per se* necessary or sufficient in enabling. If A is precarious, this does mean however, that jointly, its enablers are both necessary and sufficient for the continuation

of A. As each node in the network is precarious, the network itself is too. This is crucial for the notion of autonomy in the enactive approach (Di Paolo 2005).

A paradigmatic case that displays operational closure and precariousness is a single cell. A cell is constituted by a complex network of interrelated causal processes, but, for didactic purposes, we distinguish three. The first process comprises the metabolic network. The second process is the membrane-generation of the cell that separates the network from the environment. The third process consists of the active regulation of matter and energy exchanges of the cell, via the membrane-induced barrier, with its external environment. By way of this third process, the system can absorb nutrients from and expel wastes into its environment to continue its metabolism, looping back into process one.

The metabolic network, process 1, can be divided into subprocesses. A central aspect of metabolism is the production of enzymes, which exhibits a form of closure in itself. Enzymes are precarious. As such, when particular enzymes need to be produced, this occurs "in metabolic pathways helped by other enzymes, which in turn are produced with the participation of other ones … in a *recursive* way" (Mossio and Moreno 2010, p. 278, emphasis added). That is, the metabolic network in itself can be said to be "enzymatically closed" (Mossio and Moreno 2010, p. 278). This production network enables process 2: the generation of a membrane that separates the network from its environment. This semipermeable barrier is necessary for the system to actively regulate its exchanges with the environment. It both allows the system to take in matter and energy from the environment, and protect its internal network from external perturbation of the metabolism (Ruiz-Mirazo and Mavelli 2008; Thompson 2007). The exchange with the environment enabled by the barrier's separation is process 3. The limited openness is exploited to allow for the absorption of nutrients from the environment which can stimulate the maintenance of the membrane itself, but also "contribute to the production of an 'energy currency'" (Ruiz-Mirazo and Mavelli 2008, 376; Skulachev, 1992). Via trans-membrane mechanisms, this 'currency' is cashed out in internal metabolic reactions, transformed to serve as energy resources to maintain and actively regulate its boundary conditions (Ruiz-Mirazo, Mavelli 2008). This is to say that process 3 loops back into enabling process 1 and 2. These enabling relations are visualized in Figure 3 below. Here we can see

that operational closure and precariousness jointly correctly marks a cell as an autonomous system.

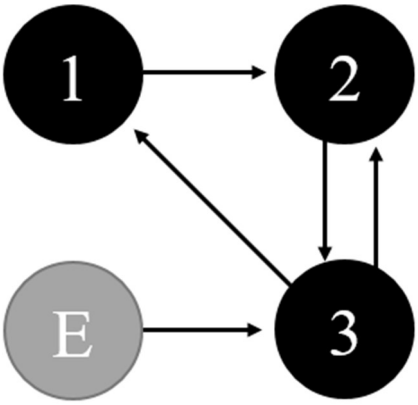

Figure 3 illustrates the simplified process network relevant to a single cell. Process 1, which captures the metabolic network, is represented by the 1 in the top-left. Process 2, membrane-generation, is represented by 2 in the top-right. Process 3, the active regulation of matter and energy exchanges with the environment is represented by 3 in the bottom-right. The environment is represented by the E in the bottom-left. The arrows between the represented processes stand for enabling relations as described above. We see that 1, 2 and 3 form a self-enabling network as per the definition above. Each process in the network enables and is enabled by at least one other process in the network. 1 enables 2 and is enabled by 3. 2 enables 3 and is enabled by 1 as well as 3. 3 enables both 1 and 2, and is enabled by 2 and E. The network here described thus represents an operationally closed system.

8.4.2 Autonomy and the pebble

A pebble is not autonomous. Given that autonomy is intended to solve the pebble challenge, it is important to subject the pebble to the same analysis: is a pebble operationally closed and precarious? If not, this indicates that autonomy as used here is an adequate concept to distinguish between abiotic and biotic systems. We distinguish four causal processes that are relevant to the formation and maintenance of the pebble's structural integrity on a shingle beach, two of which are directly considered to be determinants of a pebble's shape and size: particle abrasion and particle transport. These two processes may be more or less relevant depending on the particular geological location (Domokos and Gibbons 2012; see also Landon, 1930; Kuenen, 1964; Carr, 1969; Bluck, 1967). Particle transport refers to the transport of the pebble by the river. Particle abrasion refers to the collusion with other pebbles (and other materials) that occurs primarily during particle transport. The remaining two processes are the fluid flows of the river and the environment that consists of abraders of a hard enough consistency to allow for particle abrasion.

The four processes in the network are thus: fluid flows (A), environmental abraders (B), particle abrasion (C) and particle transport (D). Fluid flows enable particle transport, and can reasonably be considered to enable particle abrasion too. Assuming there are no other moving objects in the river, the pebble will be unlikely to move from its location and is thus unlikely to be abraded by other materials, if it is not swept anywhere by the fluid flows. Environmental abraders only enable particle abrasion. Particle abrasion in itself does not enable any other process in the network. Particle transport only enables particle abrasion. This means that fluid flows only enable other processes, but are themselves not enabled by any other process in the network. The enabling relations are specified in Figure 4 below. This means that A cannot be part of a self-enabling network. Environmental abraders only enable particle abrasion, and are not themselves enabled by other processes in the network and thus B suffers the same fate as A. C, particle abrasion, is enabled by all other processes in the network, but does not actually enable any other process, and can also not figure in a self-enabling network. Process D, particle transport, is the only process that is both enabled by and enables another process in the network, being enabled by fluid flows, enabling particle abrasion. This enabling chain, however, never loops back into enabling the continuation of particle transport. As such, Process D too cannot be part of a self-enabling network. Summing it up, there is no self-enabling network to be found in the processual network surrounding pebbles. This means that, under the operational closure formalism, pebbles are not marked as autonomous.[84]

[84] Our treatment of the pebble case may seem disanalogous with our treatment of the cell case. The discussion of the cell case treated a few important *internal* processes such as metabolism and membrane-generation next to the *external* processes concerned with exchanges with the environment. Our take on the pebble case seems to lack in internal counterparts to the external processes. This speaks to what the operational closure formalism indicates, which is that the pebble simply is not an operationally closed system. This means that, in terms of this formalism, there is no 'internal' to speak of that *could* operate (semi-)independently of the external processes.

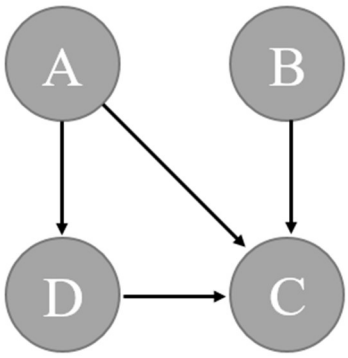

Figure 4 represents the process network relevant to a pebble on a shingle beach. The nodes with letters A, B, C and D in the figure represent the processes A, B, C, and D mentioned in-text respectively. The arrows represent enabling relations so that the arrow going down from A to C means that A enables C. Each node is coloured gray to indicate that the network is not operationally closed, as no process except for D enables and is enabled by at least one other process in the network. The network can thus not be said to be self-enabling.

8.5 Autonomy meets Markov blanket

Operational closure and precariousness provide the principled distinction between autonomous and non-autonomous systems. It is this distinction that seems difficult to capture within the Markov blanket formalism of the FEP: indeed, following Section 3, it seems as though both organisms and pebbles can be said to minimise free energy and can thus be cast as engaging in Bayesian inference. In Section 4 we have seen two notions from the enactive literature that are apt at capturing the difference between biotic and abiotic systems. As such, there is good reason to attempt to incorporate the enactive notion of autonomy into the FEP (Kirchhoff et al. 2018; Palacios et al. 2020).

A few FEP conceptions of autonomy exist in the literature, so it is important to discuss these and why they fall short of capturing operational closure and precariousness. According to one usage of autonomy, the internal and active states of any Markov blanketed system are considered *autonomous states*, because their values are not directly influenced by the environment (Friston, Wiese, Hobson, 2020). Yet this does not aid in a distinction between biotic and abiotic systems, as any Markov blanketed system by definition has internal and active states. One may also think the presence of active states in a system is crucial, as active states are what, in the FEP formalism, allow an organism to modulate their exchange with its environment. Yet, recall from Section 3 that a pebble also has active states. It would be strange to think that a pebble's existence has no influence on its environment merely because it does not *act* on its environment. A pebble's mass will influence the state of the water that may

surround it, or the movement of the adjacent pebbles on a shingle beach, and influences the behaviour trajectories of organisms in its vicinity. These sorts of influences will be formalised as active states in the Markov blanket formalism. As such, we will be able to identify external states dependent on the pebble's active states in the same way we can do so for organisms.

Yet, one may object, a pebble's exchange with its environment is much, shall we say, simpler, than an organism's. This is roughly what is captured in the distinction between between *active* particles and *inert* particles, discussed in Friston's 2019.[85] The distinction here rests on what is called the 'information length' of a system, the technical specifics of which are outside of the scope of this paper.[86] Broadly, one could say the information length of a system corresponds to the size of the 'viable bound' of the system under scrutiny as we have discussed it in Section 2. This means that a high information length is associated with systems whose internal states display a large degree of variability, whereas low information length is associated with systems whose internal states remain largely static, or consistently revisit a very small set of states. This seems to make headway into distinguishing biotic from abiotic systems, yet fails to draw a divide *in kind,* offering only a gradual distinction *in degrees*, leaving room for a grey area between biotic and abiotic. Take the pebble, for example. For the sake of argument, let us concede that the pebble's information length is sufficiently low to be termed an *inert* particle. Yet consider now a shingle beach, consisting of a large amount of pebbles, that lies at water. We can consider the beach as a whole to have its own Markov blanket, forming an *ensemble* of the individually blanketed pebbles at the beach. The complexity of the internal states of the shingle beach as a whole as it maintains its integrity (continues being a shingle beach) in spite of the environmental fluctuations (the water flowing on and off-shore, weather circumstances, etc.) increases exponentially as we imagine it to be larger, comprising more distinct and varied pebbles, each of which 'respond' differently to the varying temperatures and kinetic forces. This increases the associated information length of the shingle beach. We could do this for increasingly complex abiotic systems until, one could imagine, the information length starts to look a lot like that of a single bacterium. The crucial point here is that relying on a system's information length may not necessarily pick out biotic systems exclusively, and remains a difference *in degree*, as opposed to a difference *in kind*.[87]

[85] Thanks to an anonymous reviewer for pointing this out.

[86] See Friston's (2019) unpublished manuscript for a description in technical detail.

[87] What this means is that the grey area is not *inherently* a fault, yet conceding this does not help us in distinguishing the pebble clearly from the organism.

A final suggestion that could be thought to pick out organisms over pebbles is that of non-equilibrium steady-state (NESS). According to the FEP's more recent formulations[88], any Markov blanketed system that is at NESS with its environment can be cast as minimizing free energy (Ramstead, Badcock et al. 2019). For a system to be in a non-equilibrium steady-state so defined means that the system is far from equilibrium and, *in virtue of* systematic environmental exchange, remains in the same state over time. Yet being in a steady state implies that the system remains in *the same* state over time (Gagniuc 2017). This means that, for a dynamically changing organism in constant flux, this only holds by approximation or within certain specific timeframes. An extreme example are butterflies that just got out of their cocoon, which corresponds to a massive change in the organism's states, but humans can just as well hardly be said to occupy the same state over time.[89] A pebble is in no need of environmental exchange to remain a pebble and is thus not at NESS. Yet it is also known that NESS does not uniquely pick out biotic systems (see for example Bernard and Doyon, 2015; Pourhasan, 2016). As such, it remains of import to look at the enactive approach to autonomy and how this could be approached from within the FEP.

8.5.1 On self-individuation

A system is considered operationally closed only if it exhibits a network of self-enabling processes. That is, each process in the network enables and is enabled by at least one other process in the network. This means that any operationally closed system is inherently composed of multiple individually distinguishable component processes. Taken together, these individually distinguishable component processes form a larger network that *self-individuates*, and generates its own boundary between itself and its environment. The Markov blanket formalism is well-equipped to capture this hierarchical boundary generation of processes (Palacios et al. 2020). If we take each component process to have a Markov blanket, and the larger, operationally closed network to have a Markov blanket too, the generation of a self-enabling and self-individuating process network can be cast as the hierarchical self-organization of a Markov blanketed *ensemble* of Markov blankets. Palacios et al. (2020) show

[88] Contrary to, say, Friston (2013), the condition of a system being at NESS with its environment seems to have replaced the initial clause of being locally ergodic (see Friston 2019; but also Hipólito 2020; Ramstead, Badcock et al. 2019) . Discussion of this change and its philosophical implications are outside of the scope of this paper, but see Bruineberg et al. (2020) for preliminary discussions.

[89] The FEP may be able to accommodate 'wandering sets' (see Birkhoff 1927) which could account for changes to a system's viable bound over time, though it remains to be seen whether this could accommodate drastic and sudden changes such as the butterfly's (Friston 2019).

how, with a few crucial assumptions, single cells can be shown to aggregate quite naturally into a larger ensemble. In this particular way, we can consider each node of the network to be Markov blanketed, and the ensemble-network to be Markov blanketed in itself, as shown in Figure 5 below. The nodes of the operationally closed network need not be operationally closed themselves, which means that the nodes themselves need not invite being divided further into another layered network. We can thus ground operational closure in terms of Markov blanket ensembles without inviting an infinite regress. This maps onto a single cell organism too. Consider that each organelle of a single cell can be distinguished statistically from the rest of the cell, thus establishing a Markov blanket (Palacios et al. 2020), without in itself being operationally closed and thus not in itself requiring to be composed of a self-enabling process network under the current definition of operational closure.

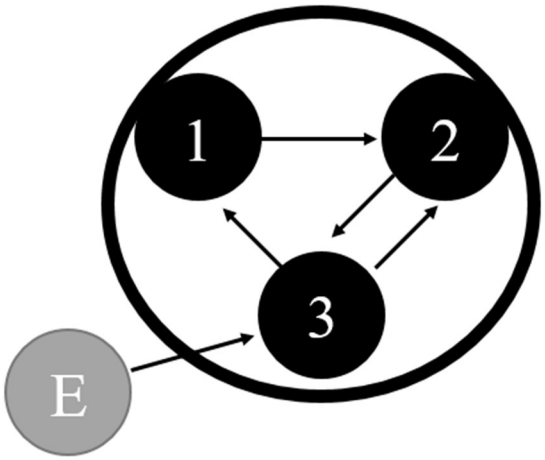

Figure 5 describes the operationally closed single cell system with a Markov blanket around the ensemble of process networks that make up the system. The process relations described are just as they were in Figure (x: cell). The circle around the self-enabling network of 1, 2 and 3 represents the Markov blanket around the ensemble.

Although this captures a key feature of operational closure, self-individuation (or membrane-generation in terms of a single biotic cell), it falls short of accounting for the conditional enabling network that differentiates autonomous from non-autonomous systems. Hierarchical self-organization is only part of the enactive story of autonomy. Indeed, the pebble challenge could be reformulated as a *shingle beach challenge* so that the beach can be cast as an ensembled Markov blanketed system that engages in Bayesian inference, composed of individually Markov blanketed pebbles. The distinction between abiotic and biotic thus remains blurred, even in a hierarchical perspective.

8.5.2 On operational closure and enabling relations

There are a few differences between just any Markov blanketed system and an operationally closed Markov blanketed system that we need to capture. Operational closure is a particularly *structured* manner of self-organization (Maturana and Varela, 1980). Increased structure over time implies that the long term entropy (informally a measure of disorder; Friston 2013) is low, which means that sensory surprise must be low too. The ensemble's states are constituted by the component states, which means that the component states inherit this low surprise. This is a key aspect of understanding operational closure in the FEP.

We can exploit the lower surprise internal to the network further by, for the sake of exposition, ignoring the system's environment. For each particular node, its sensory states are entirely determined by the active states of the other nodes in the network (Palacios et al. 2020). More specifically, if A enables B, that means that the active states of A must have an important influence on B, which in turn means that the active states of A are significantly determinant of the sensory states of B. Conversely, if B is enabled only by A, its sensory states are entirely determined by the active states of A. This means that, within the network, each node's sensory states are determined by the active states of its enablers. This implies that the sensory surprise of any node is at a nearly absolute minimum, *given* the active states of the enabling nodes.

In light of this, an enabling relation is closely related to the notion of *coupling*. Any two nodes can be said to be coupled when they are in a relation of *mutual* influence (Friston 2013). In active inference, the generative models associated with two coupled systems will approach one another over time, giving rise to what is known as *generalized synchrony*. As the coupled two systems continuously interact, they become attuned to one another; they adapt to one another (Friston 2013). This attunement means that the influence they have on one another becomes increasingly well accommodated. In mutual attunement, this entails changes in the *extrinsic* probability distribution in the state space, so that the sensory states associated with the active states of the coupled system are increasingly likely. On the scale of the network, this means that the nodes as part of the network, *i.e.* on a network-level scale, are in a tight coupling relation. This is to say that each node's influence will enable, and thus largely determine, another node's states that will, by virtue of being part of the network, couple back the initial node to enable and largely determine its own states either directly or indirectly. An operationally closed network, then, can be taken as a tightly coupled network of Markov blanketed nodes.

Note however, that, *prima facie*, the notion of coupling is not necessarily applicable to any two nodes in an enabling relation within the network. We thus cannot simply transcribe the enabling relations between nodes as coupling relations. A coupling relation is *symmetrical* insofar it prescribes mutual influence. This does not mean that the interaction needs to be *identical in both* directions of influence, but it does imply that the interaction is minimally bidirectional: the active states of one node determine the sensory states of another node *and vice versa*. Taken in this sense of direct influence, an enabling relation is not. An enabling relation can be asymmetrical, as we see in nodes 1 and 2 in Figure 5 above. This means that we would miss out on asymmetrical enabling relations if we were to transcribe them as coupling relations in a model. Moreover, an enabling relation concerns a specific type of influence that one node has on another. Consider that any random two systems may, for a certain duration over time, be coupled in a mutually *disruptive* fashion. This means that rather than *enabling* one another, they instead *inhibit* one another. This distinction too may be lost if we were to transcribe enabling relations as coupling relations. Crucially, however, we *can* say that the individual nodes are at least indirectly coupled to one another from a network perspective.

The network perspective can also capture the sense in which operational closure depends on a network's *constituents*. Consider, for example, how the free energy of a Markov blanketed ensemble of Markov blankets depends on the free energy of its constituents (Friston, 2013). Nonetheless, the shingle beach considerations in Section 5.1 remain valid.

8.5.3 On precariousness and limits

The Markov blanketed ensemble of Markov blankets has been important in our characterization of self-individuation and operational closure, so one could think it to cover precariousness as well. The idea is that an ensemble's free energy is determined by the free energy of its constituents, which means that if the constituents' free energy is minimised, the free energy of the ensemble is minimised too. It is important to note that this hierarchical dependence is a crucial feature of precariousness in organisms. That is, organisms and their component processes are inherently precarious. Yet what is typical of precariousness is not the hierarchical dependence relation: it is the natural inclination to decline.

More important here is the FEP requirement of a system to be at NESS with its environment. As we have seen, this implies that the system requires continued environmental exchange to maintain its state. *Barring the limiting remarks* of the applicability of NESS to real, living organisms noted above at the start of Section 5, the continued environmental

exchange requirement for maintaining its state *is exactly what precariousness demands*. Some employ this feature of the FEP to construe cancer, for example (Manicka and Levin 2019; Kuchlin et al. 2019).[90]

Furthermore, the low sensory surprise of an enabled node, given the active states of enabling nodes may also be able to capture an organism's *precariousness*. Recall that precariousness appears on two levels in an autonomous system. Each process in the network is precarious, and the network as a unity is too. Network-level precariousness is built into the FEP at its very core. Any system needs to put in work to be able to maintain its boundaries with the environment and continue existing. This means that without this work, the system will disintegrate, which is to say the system is naturally inclined to cessation, yet remains intact due to the 'efforts' of the system. In this sense, the organism can be taken as precarious.

However, this line of thinking invites an unintended implication on the node level. Consider that high-probability sensory states are those for which they are largely determined by their enablers' active states. The cessation of a process, further, is associated with leaving expected bounds. When a process ceases, its active states will thus by definition leave expected bounds. This implies that the sensory states of an enabled node would be highly surprising (given its ceasing enabler's active states) so that it's likely to enter an unviable state and cease as well. This seems to entail that if any random enabling node would cease, the sudden increase in sensory surprise for the enabled nodes would sooner or later cause each other process in the network to fall like dominoes. After all, their own cessation will cause a spike in sensory surprise in the nodes they enable, and so on. As the network is composed only of processes that both enable and are enabled by at least one other node in the network, no single process will be spared. In certain cases, this is to be expected. Consider our toy description of a single cell in Section 4.1 above. If we were to cease any of the processes in that network, the entire network would collapse. Each process is essential for the continuation of the network. However, this is only a contingent fact of our toy description. As stated above, it is not necessary for each enabling process to be individually necessary or sufficient for the continuation of the enabled process. This flexibility is key in our understanding of operational closure, yet is orthogonal to the domino effect we find on a node-level of description. This shows that, though this approach is able to capture certain characteristics, it is not capable of incorporating precariousness on both a network- and a node-level of description.

[90] Thanks to an anonymous reviewer for pointing this out, which inspired further considerations regarding operational closure as well.

Further, if we intend to capture the essential organizational dynamics for biotic systems, abstracting away the environment misses the point. By defining what something is (the system, or the *unity*), we indirectly define that which it is not (the environment) (Beer 2004; Friston 2019). This is exacerbated by the fact that for each probability distribution over internal states, there is an associated probability distribution over external states that specifies the expected influences of external states (Friston, Wiese, Hobson 2020). Even in the presence of an external environment, an operationally closed system intrinsically defines its environment as well as its boundary through which it can interact with the environment (Beer 2004 2014; Friston 2012). In an ecological situation, any one node's surprise is thus not at nearly absolute minimum, but can still be said to be *particularly* low, given the active states of its enablers.

In sum, we have presented some ways to consider conceptualizing operational closure and precariousness in terms of a tightly coupled network of Markov blankets. There is a sense in which tightly bound network-scale coupling, and particularly low sensory surprise of enabled nodes given the active states of enabling nodes, can capture operational closure and precariousness. This can be taken as a proof of concept. Further simulational research may aid further in the incorporation of autonomy into the FEP by putting the approach here to work.

8 Conclusion

Many FEP researchers hold the FEP to support a grand unifying ambition to account for a wide variety of phenomena, among others the organizational dynamics of living and cognitive systems. Yet a common criticism is that it is overly general and cannot distinguish between biotic and abiotic systems, making it seem uninteresting from a biological perspective. We addressed this worry by elaborating on earlier suggestions to incorporate the enactive notion of autonomy into the FEP framework. In Section 4, we described how operational closure and precariousness are concepts fit to handle the pebble challenge. In the subsequent section, we made a first attempt at incorporating the enactive language of autonomy into free energy language. We discuss different aspects of autonomy in the enactive approach and how they could potentially be transcribed into the FEP formalism. The FEP quite naturally accounts for self-individuation, a corollary of operational closure. The same applies to the bi-directional dependence relation of an operationally closed system and its component processes and a Markov ensemble and its nodes. Yet the enabling relation central to operational closure proves more challenging. There are implications with regards to the statistical relations between nodes for any operationally closed system such as an enabled node's low sensory surprise in light of

its enablers' active states that we show the FEP can account for. Precariousness is also shown to be difficult to incorporate on a node-level (although the ensemble level is able to capture some basal features of precariousness), and the complexity of an ecological environment places limits on surprise-minimization descriptions as leveraged before. The FEP can thus emulate a limited version of autonomy as it appears in the enactive approach. Simulation modeling can further help incorporate this notion of autonomy into the FEP formalism.

References

Auletta, G., (2013). Information and metabolism in bacterial chemotaxis. *Entropy*. Doi: 10.3390/e15010311

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 (Vol. 3)*. MIND Group: Frankfurt am Main. https://doi.org/10.15502/9783958573055.

Beal, M. (2003) Variational Algorithms for Approximate Bayesian Inference. PhD thesis University of Cambridge

Beer, R. D. (2004). Autopoiesis and cognition in the game of Life. Artificial Life, 10(3), 309–326.

Beer, R. D., (2014) The Cognitive Domain of a Glider in the Game of Life. *Artificial Life, 20*, 183-206. https://doi.org/10.1162/ARTL_a_00125

Bernard, D., Doyon, B. (2015) Non-Equilibrium Steady States in Conformal Field Theory. *Ann. Henri Poincaré 16***,** 113–161. https://doi.org/10.1007/s00023-014-0314-8

Birkhoff, G.D., 1927. *Dynamical systems*. American Mathematical Society, New York.

Bluck, B. J. 1967 Sedimentation of beach gravels; examples of South Wales. J. *Sedimentary Res. 37*, 128–156. doi:10.1306/74D71672-2B21-11D7-8648000102C1865D

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. https://doi.org/10.1007/s11229-016-1239-1

Bruineberg, J. and Dolega, K. and Dewhurst, J. and Baltieri, M. (2020) *The Emperor's New Markov Blankets* [Preprint]. *PhilSci Archiv* http://philsci-archive.pitt.edu/id/eprint/18467

Carr, A. P. (1969) Size grading along a pebble beach: Chesil beach, England. J. *Sedimentary Petrol. 39*, 297–311. doi:10.1306/74D71C3A-2B21-11D7-8648000102C1865D

Campbell, J.O., (2016) Universal Darwinism as a process of Bayesian inference. *Front. Syst. Neurosci. 10*(49). DOI: 10.3389/fnsys.2016.00049

Calvo, P., and Friston, K. (2017). Predicting green: really radical (plant) predictive processing. Journal of the Royal Society Interface, 14(131): 20170096. Doi: 10.1098/rsif.2017.0096

Colombo, M., Wright, C. First principles in the life sciences: the free-energy principle, organicism, and mechanism. *Synthese* (2018). https://doi.org/10.1007/s11229-018-01932-w

Conant, R. C., and Ashby, W. R. (1970) Every good regulator of a system must be a model of that system, *Int. J. Systems Sci., 1*(2), pp. 89–97

Corcoran, A. W., Pezzula, G., and Hohwy, J. (2020) From Allostatic Agents to Counterfactual Cognisers: Active Inference, Biological Regulation, and The Origins of Cognition. *Biology and Philosophy*, *35*(3). https://doi.org/10.1007/s10539-020-09746-2

Domokos, Gibbons (2012) The evolution of pebble size and shape in space and time. *Proc. R. Soc. A 468*, 3059–3079. doi:10.1098/rspa.2011.0562

Daunizeau, J., den Ouden, H. E. M., Pessiglione, M., Kiebel, S. J., Stephan, K. E., Friston, K. J. (2010) Observing the Observer (I): Meta-Bayesian Models of Learning and Decision-Making. *PLoS ONE 5*(12): e15554. https://doi.org/10.1371/journal.pone.0015554

Di Paolo, E. A. (2005). Autopoiesis, Adaptivity, Teleology, Agency. Phenomenology and the Cognitive Sciences, 4(4), 429. https://doi.org/10.1007/s11097-005-9002-y

Di Paolo, E., & Thompson, E. (2014). *The enactive approach.* In L. Shapiro (Ed.), *Routledge handbooks in philosophy. The Routledge handbook of embodied cognition* (p. 68–78). Routledge/Taylor & Francis Group

Friston, K. J., & Frith, C. D. (2015). Active inference, communication and hermeneutics. *Cortex; a journal devoted to the study of the nervous system and behavior*, *68*, 129–143. https://doi.org/10.1016/j.cortex.2015.03.025

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. (2012). A free energy principle for biological systems. *Entropy*, 1392 14(11):2100–2121.

Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, 10(86). https://doi.org/10.1098/rsif.2013.0475

Friston, K. (2019). A free energy principle for a particular physics. Unpublished manuscript.

Friston, K., Ao, P. (2012) Free-energy, value and attractors.*Computational and mathematical methods in medicine*.

Friston, K., Kilner, J. & Harrison, L. A (2006) free energy principle for the brain. Journal of Physiology-Paris 100, 70–87, https://doi.org/10.1016/j.jphysparis.2006.10.001.

Friston, K. & Kiebel, S. (2009) Cortical circuits for perceptual inference. *Neural Networks 22*:1093–104.

Friston, K. J., Parr, T., and de Vries, B. (2017a). The graphical brain: belief propagation and active inference. *Netw. Neurosci.* 1, 381–414. doi: 10.1162/NETN_a_00018

Friston, K., and Stephan, K. E. (2007). Free energy and the brain. Synthese, 159, 417–458.

Friston, K. J., Wiese, W., Hobson, J. A., (2020) Sentience and the Origins of Consciousness: From Cartesian Duality to Markovian Monism. *Entropy 22*(5), 516; https://doi.org/10.3390/e22050516

Gagniuc, Paul A. (2017). *Markov Chains: From Theory to Implementation and Experimentation*. USA, NJ: John Wiley & Sons.

Hesp, C., Ramstead,M., Constant, A., Badcock, P., Kirchhoff, M.D., and Friston, K. (2019). A Multi-scale view of the emergent complexity of life: A free energy proposal. In M. Price et al. (eds), *Evolution, Development, and Complexity: Multiscale Models in Complex Adaptive Systems*. Springer

Hipólito, I. (2019). A simple theory of every 'thing.' *Physics of Life Reviews*, *31*, 79–85. https://doi.org/10.1016/j.plrev.2019.10.006

Hohwy, J. (2016) The self-evidencing brain. *Noûs 50*(2), 259–285. doi: 10.1111/nous.12062

Hohwy, J. (2017). How to Entrain Your Evil Demon. In T. Metzinger & W. Wiese (Eds.). Philosophy and Predictive Processing: 2. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573048

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

Kiebel S. J., Daunizeau J., Friston K. J. (2008) A Hierarchy of Time-Scales and the Brain. *PLoS Comput Biol 4*(11): e1000209. https://doi.org/10.1371/journal.pcbi.1000209

Kirchhoff, M. D. (2018). Autopoiesis, free energy, and the life-mind continuity thesis. *Synthese*, 195(6), 2519-2540.

Kirchhoff, M., Froese, T. (2017). Where There Is Life There Is Mind: In Support of a Strong Life-Mind Continuity Thesis. *Entropy, 19*(4). https://doi.org/10.3390/e19040169

Kirchhoff, M.D., Kiverstein, J. How to determine the boundaries of the mind: a Markov blanket proposal. *Synthese* (2019). https://doi.org/10.1007/s11229-019-02370-y

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. JOURNAL OF THE ROYAL SOCIETY INTERFACE, 15(138). https://doi.org/10.1098/rsif.2017.0792

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

Kuchling F, Friston K, Georgiev G, Levin M. (2019) Morphogenesis as Bayesian inference: A variational approach to pattern formation and control in complex biological systems. *Phys Life Rev.* doi: 10.1016/j.plrev.2019.06.001.

Kuenen, Ph. H. 1964 Experimental abrasion of pebbles. VI. Surf action. *Sedimentology 3*, 29–43. doi:10.1111/j.1365-3091.1964.tb00273.x

Laland, K., Matthews, B., and Feldman, W. (2016). An introduction to niche construction theory. *Evolutionary Ecology*, 30, 191-202.

Landon, R. E. 1930 An analysis of beach pebble abrasion and transportation. *J. Geol. 38*, 437–446. doi:10.1086/623739

Linson, A., Clark, A., Ramamoorthy, S., and Friston, K. (2018) The Active Inference Approach to Ecological Perception: General Information Dynamics for Natural and Artificial Embodied Cognition. *Frontiers in Robotics and AI*, *5* (21). doi: 10.3389/frobt.2018.00021

Manicka S, Levin M. (2019) Modeling somatic computation with non-neural bioelectric networks. *Scientific reports. 9*(1):18612. doi: 10.1038/s41598-019-54859-8

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition*. Boston: Reidel

Mossio, M. and Moreno, A., (2010) Organisational Closure in Biological Organisms. *Hist. Phil. Life Sci.*, 32 (2010), 269-288

Olivotos, S., & Economou-Eliopoulos, M. (2016). Gibbs Free Energy of Formation for Selected Platinum Group Minerals (PGM). *Geosciences*, *6*(1), 2. doi:10.3390/geosciences6010002

Parr, T., and Friston, K. (2017). Working memory, attention, and salience in active inference. *Scientific Reports*, 7: 14678 | DOI:10.1038/s41598-017-15249-0

Palacios, E. R., Razi, A., Parr, T., Kirchhoff, M., & Friston, K. (2020). On Markov blankets and hierarchical self-organisation. Journal of Theoretical Biology, 486, 110089. https://doi.org/10.1016/j.jtbi.2019.110089

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann.

Pourhasan, R. (2016) Non-equilibrium steady state in the hydro regime. *J. High Energ. Phys. 2016,* 5. https://doi.org/10.1007/JHEP02(2016)005

Ramstead, M. J. D., Constant, A., Badcock, P. B., & Friston, K. J. (2019). Variational ecology and the physics of sentient systems. *Physics of Life Reviews*, *31*, 188–205. https://doi.org/10.1016/j.plrev.2018.12.002

Ramstead, M., Kirchhoff, M. D., Constant, A., and Friston, K. (2019). Multiscale Integration: Beyond Internalism and Externalism. *Synthese*, 10.1007/s11229-019-02115-x.

Ruiz-Mirao, Mavelli (2007) On the way towards 'basic autonomous agents': Stochastic simulations of minimal lipid–peptide cells. *BioSystems, 91*, 374-387

Schrödinger, E., (1944) *What is life?* Cambridge: Cambridge University Press.

Skulachev, V .P., (1992) The laws of cell energetics. *Eur. J. Biochem. 208*, 203–209.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal, 27*(3):379–423.

Thompson, E. (2007). Mind in life: Biology, phenomenology, and the sciences of mind. Cambridge, MA: Harvard University Press.

van Es, T. (2019). Minimizing prediction errors in predictive processing: from inconsistency to non-representationalism. *Phenomenology and the Cognitive Sciences*. https://doi.org/10.1007/s11097-019-09649-y

van Es, T. (2020) Living models or life modeled? On the use of models in the free energy principle. *Adaptive behavior*. https://doi.org/10.1177/1059712320918678

Varela, F. (1979) Principles of Biological Autonomy. The North-Holland Series in General Systems Research, Vol. 2. Elsevier North-Holland

Varela, F., Thompson, E., Rosch, E., (1991) *The Embodied Mind: Cognitive Science and Human Experience*, Cambridge, MA: MIT Press

Williams, D. (2018). Predictive coding and thought. *Synthese*, 197, 1749-1775.

Concluding Remarks

This thesis is a collection of pieces I have (co-)written with a specific focus on the free energy principle and the associated theory of predictive processing and enactivism. The free energy principle is important in the first part as a background theory supporting predictive processing, but becomes more important as we move through the different discussions. In general, the discussions have been aimed at conceptual clarification. That is, I have focused on the internal theoretical structures of different proposals, rather than on how to test their empirical merit. The first half is primarily a critical discussion of existing work and trends, whereas towards the end there is emphasis on the direction future lines of research can and should take. The middle part provides a turning point where we leave the more realist interpretations of the free energy principle under the guise of predictive processing behind, and move towards what I argue to be a more fruitful, instrumentalist perspective. In the final two chapters we explore the fertility of an enactive free energy principle. I want to take the time here and reflect on the takeaway message of the different sections in this thesis.

Part I focused primarily on realist interpretations of the free energy principle and the purported need for representation in explaining cognition. Each chapter discussed a specific defence of the need for representations and argued them to be unsatisfactory. Broadly, in Chapter 2 the proposed definition was deemed insufficient to distinguish representational from non-representational relations; in Chapter 3 the need for representation and inference was shown to have contradictory implications; in Chapter 4 I argued that non-representational processes could explain the phenomena that were thought to require representation.With the rejection of the representational approach, Part I cleared the playing field for Parts II and III.

More specifically, in Chapter 2, Erik Myin and I discussed the notion of *representation* at work in the predictive processing framework, specifically focusing on the discussion from Gladziejewski (2016). Gladziejewski attempts to justify the use of representations in predictive processing by way of a *compare to prototype* strategy as championed by Ramsey (2007). He takes cartographic maps as his prototype, and determines four distinct features, which he argues conjoinedly describe what makes a cartographic map representational. We discussed these features one by one, and showed that none of them actually help in getting closer to distinguishing representation from non-representation. Moreover, we provide a counterexample that contains all four features, but would not reasonably be called representational. This brought into question whether the initial prototype analysis was perhaps flawed. The flaw, we argued, is that it misses the social factor on the basis of which we,

socioculturally, co-determine which structures represent others, by collectively *using* the structures, and correcting others for incorrect usage. This implies that representations exist only as the products of interactional practices between people. The quest to unearth representations in, say, a brain, is thus conceptually confused. This means that, if we are to reap the benefits of the insights of predictive processing, we will need to rethink how these considerations relate to the actual organisms we describe, as was done in Part II.

Chapter 3 continued the investigation of the representational paradigm as it appears in predictive processing. Here I argued that a specific contradiction appears when we take the predictive processing framework seriously, with particular focus on its theoretical commitment to inferentialism and hierarchical layering based on Markov-blanketed boundaries as discussed in Hohwy (2016). The very basic idea is that a Markov blanket is a statistical partitioning method that distinguishes different set of states in a model. Hohwy interprets these blankets to constitute an *epistemological* boundary between the states within the blanket and those outside of it. This implies that anything that happens within the boundary does not require representation, and anything beyond the boundary does. This becomes an issue when the boundary appears at multiple hierarchical scales where there are multiple smaller boundaries that appear within a system that has its own boundaries. This implies that from the perspective of one of the smaller, partial systems, the other partial systems require representation. However, when this partial system is seen as a component of the larger bounded system, they do not require mutual representation as they are both *within* the larger blanket. As such, the partial systems both need to and need not represent the other partial systems within the larger blanket, depending on how we characterize the situation. As such, it cannot be that the Markov blanket carves out *actual* epistemological boundaries. Instead, as we have seen in later chapters in Part II and Part III, the Markov blanket formalism is best taken as a statistical identification tool, instead of a real feature of living systems. Taken in conjunction with the findings of Chapter 2, the representational approach is increasingly unattractive. Its appeal to representations is unfounded, yet even if we grant them the representations, the position is contradictory.

Whilst the previous two chapters were primarily concerned with how representationalism *doesn't* work, the final grapple with representationalist thought in Chapter 4 provides an alternative. Specifically, I discussed the embedded view proposed by Orlandi (2014, 2012, 2013). She suggests a *middle passage* between representationalism and non-representationalism, and has been criticized by both sides for being inconsistent. Essentially, under the embedded view we can explain how agents display sensitivity to environmental regularities without invoking representations, whilst we retain the need for representations to

explain our engagement with the *absent*. I argued that the tools she suggests to explain agential sensitivity to environmental regularities can also be used to explain the examples of absence she discusses, such as partial occlusion. It is exactly because of our previous interactional history with partially occluded objects that we are sensitive to the ways to interact with partially occluded objects right now. This shows not only that the representational paradigm is undesirable as argued in the previous two chapters, but also that we may well account for the wide variety of organismic activity that we find in nature without appealing to internal representations of any kind.

In sum, Part I discussed some of the issues with representational approaches in their application to explaining how we become attuned to environmental regularities. What we have argued is that representational commitments invite contradictions and fall short in delivering what they purport to do. Moreover, assuming representations is also unnecessary. The latter we see specifically in Chapter 4, where the non-representational explanation put forward by Orlandi was shown to also cover those situations she considered to require representation.

The predictive processing approach is often interpreted as a specific interpretation or implementation of the free energy principle. Part I argued that this interpretation is untenable, yet this does not imply that the free energy principle itself is without merit. Part II investigated this potential merit, and explored the possible interpretations of the free energy principle that would net us the most. Specifically, it gets explicit on the question of what to do with models in the free energy principle. Should we consider the model to be literally used by any organism in their navigation of the world, or should we instead treat these models as 'merely' useful tools we can use to analyze complex patterns of activity?

The predictive processing literature, as we have seen in Part I, takes a firm stance on the question, and suggests to do the former: organisms literally encode a complex Bayesian model in their brains to predict the best course of action. Yet the free energy principle literature may at first blush seem at odds with itself. At one point it seems like the entire approach is a mere metaphor for action and perception loops, at another it seems we can find Markov blankets everywhere in nature, as though they are properties of things in themselves, lying in wait for our discovery. In Chapter 5, I analyzed this tendency, and argued that the formalism is mistakenly used to justify philosophical claims that it does not actually have any relevance to. Instead, it seems that the philosophical conclusions only follow if we assume them to be built into the premises, making the entire reasoning viciously circular. In Chapter 6, co-written with Inês Hipólito, we continue this project and provide a systematic overview of the different interpretations of the free energy principle that we see in the literature. We argue that only an

instrumentalist position actually holds weight. An interesting implication is that the question of representationalism with regards to the free energy principle that stood central in the discussions of predictive processing transforms in its entirety. Rather than concerning whether organisms form representations of their environment, the question concerns scientific practice. Although in no way an unimportant discussion to be had, it does not matter for philosophy of cognitive science in the same way.

Part I can be seen as primarily a rejection of the representationalist direction. Part II, then, can be seen as exploring the various directions, and deciding on one of them. Part III, finally, can be seen as taking the first steps moving forward on this path. More specifically, Part III first employed the enactive framework to better understand certain phenomena, and then attempted to incorporate the enactive framework into the free energy formalism.

Following this plan, Chapter 7, co-written with Jo Bervoets, took a different direction from the rest of the thesis in its exclusive focus on the enactive framework. In this, we argued that the enactive conceptualization of an agent's sensitivity to sensorimotor interactions can be used to understand autism as an atypical pattern of sensorimotor interactivity. Of particular interest is how this explains the broad heterogeneity of atypicalities displayed by autistics on a social level, whilst also providing space for the broadly shared similarities on non-social sensorimotor dynamics. Chapter 7 gave a taste of what we can do with the enactive approach. The question then arised whether the free energy framework could potentially underwrite the framework, incorporating its conceptual developments. In Chapter 8, co-written with Michael Kirchhoff, we explored the options. We discussed the notion of autonomy as operationally closed precariousness in the enactive approach and discerned whether the free energy principle formalism could describe the particular dynamics picked out by enactivism. The answer was mixed. There are certain features that lend themselves well for being modeled in the free energy formalism, whereas others are not as easily captured. This implies that, though there is a lot of promise in the free energy principle, it also has explanatory limits. As such, we will need to look outside of the free energy principle's boundaries for inspiration.

In conclusion, despite exploring various issues in the philosophy of cognitive science, this thesis is unified by a number of underlying themes. First, there is the notion of anticipation and the sense in which our current behavior is informed by our interactional history. We have seen this in Part I most prominently and the discussion of an enactive conceptualization of autism puts this idea to work. Second, throughout there is a recurrent grappling with the free energy principle and its process theories: what any of it means and how it can help us understand our target systems better. In Part I predictive processing is rejected and an

alternative approach to understanding our sensitivity to interactional regularities is investigated, in Part II the breadth of options is explored, advocating for an instrumentalist take, and in Part III this perspective is put to work. While only briefly raising its head in Part III, there is an enactivism-inspired undercurrent running through all the work presented here that you can see primarily in the style of argumentation and the issues that are focused on. There is a consistent rejection of neurocentrism, of over-intellectualisation of organismic activity, proposing instead to study the organism-environment system as a whole. With this, we can distinguish an enactive perspective on the study of organismic activity, given further depth by free energy inspired statistical analysis. I submit that this suggests a fruitful pathway for future research to walk on.

Complete Bibliography

Allen, M., and Friston, K. (2016) From cognitivism to autopoiesis: towards a computational framework for the embodied mind. Synthese, 1–24.

American Psychiatric Association (2013). Diagnostic and Statistical Manual of Mental Disorders (DSM-5§). Washington, DC: American Psychiatric Association Publishing. doi: 10.1176/appi.books.97808904 25596

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 (Vol. 3)*. MIND Group: Frankfurt am Main. https://doi.org/10.15502/9783958573055.

Auletta, G., (2013). Information and metabolism in bacterial chemotaxis. *Entropy*. Doi: 10.3390/e15010311

Badcock, Friston, Ramstead (2019) The hierarchically mechanistic mind: A free-energy formulation of the human psyche

Baltieri, M., Buckley, C. L. (2017, September). An active inference implementation of phototaxis. In *Artificial Life Conference Proceedings 14* (pp. 36-43). One Rogers Street, Cambridge, MA 02142-1209 USA journals-info@ mit. edu: MIT Press.

Baltieri, M., Buckley, C. L. (2019). Generative models as parsimonious descriptions of sensorimotor loops. *arXiv preprint arXiv:1904.12937*.

Baltieri, M., Buckley, C. L., & Bruineberg, J. (2020). Predictions in the eye of the beholder: an active inference account of Watt governors. In *Artificial Life Conference Proceedings* (pp. 121-129). One Rogers Street, Cambridge, MA 02142-1209 USA journals-info@ mit. edu: MIT Press.

Barandiaran, X. E. (2017). Autonomy and enactivism: towards a theory of sensorimotor autonomous agency. *Topoi 36*, 409–430. doi:10.1007/ s11245- 016- 9365- 4.

Barnes, Elizabeth. (2016). *The Minority Body: A Theory of Disability*. Oxford University Press.

Baron-Cohen, S. (2000). Theory of mind and autism: a review. *Int. Rev. Res. Ment. Retard. 23*, 169–184. doi: 10.1016/s0074-7750(00)80010-5

Beal, M. (2003) Variational Algorithms for Approximate Bayesian Inference. PhD thesis University of Cambridge

Beer, R. D. (2004). Autopoiesis and cognition in the game of Life. Artificial Life, 10(3), 309–326.

Beer, R. D. (2014) The Cognitive Domain of a Glider in the Game of Life. *Artificial Life, 20*, 183-206. https://doi.org/10.1162/ARTL_a_00125

Bernard, D., Doyon, B. (2015) Non-Equilibrium Steady States in Conformal Field Theory. *Ann. Henri Poincaré 16***,** 113–161. https://doi.org/10.1007/s00023-014-0314-8

Bervoets, J, and Hens, K. (2020). Going Beyond the Catch-22 of Autism Diagnosis and Research. The Moral Implications of (Not) Asking 'What Is Autism? *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2020.529193.

Birkhoff, G.D., 1927. *Dynamical systems*. American Mathematical Society, New York.

Bluck, B. J. 1967 Sedimentation of beach gravels; examples of South Wales. J. *Sedimentary Res. 37*, 128–156. doi:10.1306/74D71672-2B21-11D7-8648000102C1865D

Bolis, D., Balsters, J., Wenderoth, N., Becchio, C., and Schilbach, L. (2017). Beyond autism: introducing the dialectical misattunement hypothesis and a Bayesian account of intersubjectivity. *Psychopathology 50*, 355–372. doi: 10. 1159/000484353

Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence, 47*, 139–159.

Brown, R. (1828). XXVII. A brief account of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *The Philosophical Magazine*, 4(21), 161-173.

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

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. https://doi.org/10.1007/s11229-016-1239-1

Bruineberg, J., Rietveld, E., Parr, T., van Maanen, L., & Friston, K. J. (2018). Free-energy minimization in joint agent-environment systems: A niche construction perspective. Journal of Theoretical Biology, 455, 161–178. https://doi.org/10.1016/j.jtbi.2018.07.002

Bruineberg, J., Dolega, K., Dewhurst, J. and Baltieri, M. (2020) *The Emperor's New Markov Blankets* [Preprint]. *PhilSci Archiv* http://philsci-archive.pitt.edu/id/eprint/18467

Bruineberg, J. and Rietveld, E. (2014) Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience 8* (599), pp. 1-14.

Burge, T. (2010). *Origins of objectivity*. Oxford University Press.

Calvo, P., and Friston, K. (2017). Predicting green: really radical (plant) predictive processing. Journal of the Royal Society Interface, 14(131): 20170096. Doi: 10.1098/rsif.2017.0096

Campbell, J.O., (2016) Universal Darwinism as a process of Bayesian inference. *Front. Syst. Neurosci. 10*(49). DOI: 10.3389/fnsys.2016.00049

Carr, A. P. (1969) Size grading along a pebble beach: Chesil beach, England. J. *Sedimentary Petrol. 39*, 297–311. doi:10.1306/74D71C3A-2B21-11D7-8648000102C1865D

Chapman, R. (2020). The reality of autism: on the metaphysics of disorder and diversity. *Philos. Psychol. 33*, 799–819. doi: 10.1080/09515089.2020.17 51103

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

Clark, A. (2013) Whatever Next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences 36*, 181–253

Clark, A. (2015). Predicting peace: The end of the representation wars. In Metzinger, T. and Windt, J. (Eds.). *Philosophy and Predictive Processing: 7*. MIND Group, 1–7

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Clark, A. (2017) How to Knit Your Own Markov Blanket: Resisting the Second Law with Metamorphic Minds in Metzinger, T. and Wiese, W. (eds.). Philosophy and predictive processing. Frankfurt am Main : MIND Group.

Colombo, M. and Wright, C. (2018) First principles in the life-sciences: the free-energy principle, organicism and mechanism. Synthese, 1-26. https://doi.org/10.1007/s11229-018-01932-w

Colombo, M. Elkin, L. and Hartman, S. (2018) Being Realist about Bayes, and Predictive Processing. The British Journal for the Philosophy of Science, axy059, https://doi.org/10.1093/bjps/axy059

Conant, R. C., and Ashby, W. R. (1970) Every good regulator of a system must be a model of that system, *Int. J. Systems Sci., 1*(2), pp. 89–97

Constant, A., Bervoets, J. Hens, K., and Van de Cruys, S. (2018). Precise Worlds for Certain Minds: An Ecological Perspective on the Relational Self in Autism. *Topoi*. https://doi.org/10.1007/s11245-018-9546-4.

Constant, A., Ramstead, M., Veissière, S., Campbell, J., Friston, K. (2018) A variational approach to niche construction. J. R. Soc. Interface 15: 20170685. http://dx.doi.org/10.1098/rsif.2017.0685

Corcoran, A. W., Pezzula, G., and Hohwy, J. (2020) From Allostatic Agents to Counterfactual Cognisers: Active Inference, Biological Regulation, and The Origins of Cognition. *Biology and Philosophy, 35*(3). https://doi.org/10.1007/s10539-020-09746-2

Crauel, H., Flandoli, F. (1994) Attractors for random dynamical systems. *Probab. Th. Rel. Fields 100***,** 365–393. https://doi.org/10.1007/BF01193705

Curry, G. and Ravenscroft, I. (2002). *Recreative Minds*. Oxford University Press, Oxford. doi:10.1093/acprof:oso/9780198238089.001.0001

Daunizeau, J., den Ouden, H. E. M., Pessiglione, M., Kiebel, S. J., Stephan, K. E., Friston, K. J. (2010) Observing the Observer (I): Meta-Bayesian Models of Learning and Decision-Making. *PLoS ONE 5*(12): e15554. https://doi.org/10.1371/journal.pone.0015554

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The helmholtz machine. Neural computation, 7(5), 889-904.

de Bruin, L and Michael, J. (2017) Prediction error minimization: Implications for Embodied Cognition and the Extended Mind Hypothesis. Brain and Cognition 112:58-63

Degenaar, J., & Myin, E. (2014). Representation-hunger reconsidered. *Synthese*, *191*(15), 3639-3648.

De Jaegher, H. (2013). Embodiment and sense-making in autism. *Front. Integr. Neurosci. 7*:15. doi: 10.3389/fnint.2013.00015

Dennett, D. C. (2013). *Intuition pumps and other tools for thinking*. W.W. Norton & Company.

de Oliveira, G.S. (2018) Representationalism is a dead end. Synthese, 1-21. https://doi.org/10.1007/s11229-018-01995-9

Di Paolo, E., & Thompson, E. (2014). *The enactive approach.* In L. Shapiro (Ed.), *Routledge handbooks in philosophy. The Routledge handbook of embodied cognition* (p. 68–78). Routledge/Taylor & Francis Group

Di Paolo, E. (2005). Autopoiesis, Adaptivity, Teleology, Agency. Phenomenology and the Cognitive Sciences, 4(4), 429. https://doi.org/10.1007/s11097-005-9002-y

Di Paolo, E., and De Jaegher, H. (2012). The interactive brain hypothesis. *Front. Hum. Neurosci. 6*:163. doi: 10.3389/fnhum.2012.00163

Di Paolo, E., Buhrmann, T., & Barandiaran, X. (2017). *Sensorimotor life: An enactive proposal*. Oxford University Press.

Di Paolo, E., Cuffari, E. C., & De Jaegher, H. (2018). *Linguistic Bodies: The Continuity Between Life and Language.* Cambridge, MA, USA: MIT Press.

Domokos, Gibbons (2012) The evolution of pebble size and shape in space and time. *Proc. R. Soc. A 468*, 3059–3079. doi:10.1098/rspa.2011.0562

Dołega, K. (2017). Moderate Predictive Processing. In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:10*. Frankfurt am Main: MIND Group

Einstein, A. (1905). On the movement of small particles suspended in stationary liquids required by the molecularkinetic theory of heat. *Ann. d. Phys, 17*(549-560), 1.

Engel, A. K., Fries, P. & Singer, W. (2001). Dynamic predictions: Oscillations and synchrony in top-down processing. *Nat Rev Neurosci 2*(10), 704–716.

Fabry, R. (2017) Transcending the evidentiary boundary: Prediction error minimization, embodied interaction, and explanatory pluralism, *Philosophical Psychology, 30*:4, 395-414

Fields, C., & Levin, M. (2020). Scale-Free Biology: Integrating Evolutionary and Developmental Thinking. *BioEssays*, 42(8), 1900228.

Fodor, J. A. (1975). *The language of thoug*ht. Harvard university press.

Fodor, J. A. (1987). *Psychosemantics*. MIT Press.

Frigg, R. and Hartmann, S., (2018) Models in Science, in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy (Summer 2018 Edition)*, URL = <https://plato.stanford.edu/archives/sum2018/entries/models-science/>.

Friston, K. (2002) Functional integration and inference in the brain. Progress in Neurobiology 590, 1–31

Friston, K. (2003) Learning and Inference in the Brain. Neural Networks 16, 1325–1352

Friston, K. (2005) A Theory of Cortical Response. Phil. Trans. R. Soc. B 360, 815–836. doi:10.1098/rstb.2005.1622

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) Embodied inference: or "I think therefore I am, if I am what I think". In: Tschacher, W., Bergomi C. (eds) The implications of embodiment: cognition and communication. Imprint Academic.

Friston, K. (2012) A Free Energy Principle For Biological Systems. Entropy 14, 2100-2121. doi:10.3390/e14112100

Friston, K. (2012) The history of the future of the Bayesian brain. *NeuroImage*, *62*(2), 1230-1233.

Friston, K. (2013). Active inference and free energy. *Behavioral and Brain Sciences*, *36*, 212–213.

Friston, K. (2013). Life as we know it. *Journal of the Royal Society, Interface, 10*, 20130475. https://doi.org/10.1098/rsif.2013.0475

Friston, K. (2019). *A free energy principle for a particular physics.* Unpublished manuscript.

Friston, K., and Stephan, K. E. (2007). Free energy and the brain. Synthese, 159, 417–458.

Friston, K., Ao, P. (2012) Free-energy, value and attractors.*Computational and mathematical methods in medicine*.

Friston, K., J. Daunizeau, J. Kilner and S. Kiebel (2010). Action and behavior: A free–energy formulation. *Biological Cybernetics 102*(3): 227–260.

Friston, K., Kilner, J. & Harrison, L. A (2006) free energy principle for the brain. Journal of Physiology-Paris 100, 70–87, https://doi.org/10.1016/j.jphysparis.2006.10.001.

Friston, K., Parr, T., Yufik, Y., Sajid, N., Price, C. J., & Holmes, E. (2020). Generative models, language and active inference. *PsyArXiv*. DOI: 10.31234/osf.io/4j2k6

Friston, K., S. Samothrakis and R. Montague (2012). Active inference and agency: Optimal control without cost functions. *Biological Cybernetics 106*(8): 523–541.

Friston, K. & Kiebel, S. (2009) Cortical circuits for perceptual inference. *Neural Networks 22*:1093–104.

Friston, K. and Buzsáki, G. (2016) The functional anatomy of time: What and when in the brain. Trends in Cognitive Science, 20(7): 500-511.

Friston, K. and Siebel, S. (2009). Predictive Coding under the Free-Energy Principle. *Phil. Trans. R. Soc. B 364*, 1211-1221.

Friston, K. J., & Frith, C. D. (2015). Active inference, communication and hermeneutics. *Cortex; a journal devoted to the study of the nervous system and behavior*, *68*, 129–143. https://doi.org/10.1016/j.cortex.2015.03.025

Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, *159*(3), 417-458.

Friston, K. J., Daunizeau, J. & Kiebel, S. J. (2009) Reinforcement learning or active inference? PLoS (Public Library of Science) One 4(7): e6421.

Friston, K. J., Fagerholm, E. D., Zarghami, T. S., Parr, T., Hipólito, I., Magrou, L., & Razi, A. (2020). Parcels and particles: Markov blankets in the brain. *arXiv preprint arXiv:2007.09704*.

Friston, K. J., Fortier, M., & Friedman, D. A. (2018). *Of woodlice and men: A Bayesian account of cognition, life and consciousness. An interview with Karl Friston*. ALIUS Bulletin, 2, 17-43.

Friston, K. J., Parr, T., and de Vries, B. (2017a). The graphical brain: belief propagation and active inference. *Netw. Neurosci.* 1, 381–414. doi: 10.1162/NETN_a_00018

Friston, K. J., Wiese, W., Hobson, J. A., (2020) Sentience and the Origins of Consciousness: From Cartesian Duality to Markovian Monism. *Entropy 22*(5), 516; https://doi.org/10.3390/e22050516

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

Frith, U. (1996). Cognitive explanations of autism. *Acta Paediatr. 416*, 63–68. doi: 10.1111/j.1651-2227.1996.tb14280.x

Fuchs, T., and De Jaegher, H. (2009). Enactive intersubjectivity: participatory sense-making and mutual incorporation. *Phenomenol. Cogn. Sci. 8*, 465–486. doi: 10.1007/s11097-009-9136-4

Gagniuc, Paul A. (2017). *Markov Chains: From Theory to Implementation and Experimentation*. USA, NJ: John Wiley & Sons.

Gallagher, S. (2004). Understanding Interpersonal Problems in Autism. *Philosophy, Psychiatry, and Psychology 11* (3):199-217.

Gallagher, S. (2017) *Enactivist Interventions: Rethinking the Mind*. Oxford University Press

Gallagher, S. (2020). Action and interaction. Oxford University Press.

Gallagher, S. and Allen, M. (2018) Active inference, enactivism and the hermeneutics of social cognition. *Synthese 195*(6), 2627-2648

Gammaitoni, L., Hänggi, P., Jung, P., & Marchesoni, F. (1998). Stochastic resonance. *Reviews of modern physics*, *70*(1), 223.

Gibson, J. J. (1979), *The Perception of the Visual World*. Lawrence Erlbaum.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Gibson, J. J. (1986). *The Ecological Approach to Visual Perception*. Psychology Press , London.

Giere, R. (2010). An agent-based conception of models and scientific representation. *Synthese, 172*(2), 269–281.

Giere, R. N. (1988). Explaining science: A cognitive approach. Chicago: University of Chicago Press.

Godfrey-Smith P (1996) *Complexity and the function of mind in nature*. Cambridge University Press.

Godfrey-Smith, P., and Sterelny, K. (2016) Biological Information in Zalta, E. N. (ed) *Stanford Encyclopedia of Philosophy (Summer 2016 Edition.*

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences, 290*(1038), 181-197.

Gulli, R. A. (2019). Beyond metaphors and semantics: A framework for causal inference in neuroscience. Behavioral and Brain Sciences, 42.

Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: causally relevant and different from detectors. *Biology & philosophy, 32*(3), 337–355. https://doi.org/10.1007/s10539-017-9562-6

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese, 193*(2), 559–582.

Happé, F., and Frith, U. (2006). The weak coherence account: detail-focused cognitive style in autism spectrum disorders. *J. Autism Dev. Disord. 36*, 5–25. doi: 10.1007/s10803-005-0039-0

Hens, K., and Langenberg, R. (2018). *Experiences of Adults Following an Autism Diagnosis*. Cham: Springer. doi: 10.1007/978-3-319-97973-1

Hesp, C., Ramstead, M., Constant, A., Badcock, P., Kirchhoff, M., & Friston, K. (2019). A multi-scale view of the emergent complexity of life: A free-energy proposal. In G. Georgiev, J. Smart, C. L. Flores Martinez, & M. Price (Eds.), *Evolution, development, and complexity: Multiscale models in complex adaptive systems* (pp. 195–227). Springer.

Hipólito, I. (2019). A simple theory of every 'thing.' *Physics of Life Reviews*, *31*, 79–85. https://doi.org/10.1016/j.plrev.2019.10.006

Hipólito, I., Baltieri, M., Friston J, K., &

Hipólito, I., Hutto, D, & Chown, N. (2020). Understanding autistic individuals. Cognitive diversity not theoretical deficit. In Rosqvist, H. B., Chown, N., and Stenning, A (eds), *Neurodiversity Studies. A New Critical Paradigm.* Routledge, London. https://doi.org/10.4324/9780429322297

Ramstead, M. J. (2020). Embodied Skillful Performance: Where the Action Is. *Synthese*.

Hipólito, I., Ramstead, M., Constant, A., & Friston, K. J. (2020). Cognition coming about: Self-organisation and free-energy: Commentary on "The growth of cognition: Free energy minimization and the embryogenesis of cortical computation" by Wright and Bourke (2020). *Physics of Life Reviews*.

Hipólito, I., Ramstead, M., Convertino, L., Bhat, A., Friston, K., & Parr, T. (2020b). Markov blankets in the brain. *arXiv preprint arXiv:2006.02741*.

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Hohwy, J. (2016) The self-evidencing brain. *Noûs 50*(2), 259–285. doi: 10.1111/nous.12062

Hohwy, J. (2017). How to Entrain Your Evil Demon. In T. Metzinger & W. Wiese (Eds.). Philosophy and Predictive Processing: 2. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573048

Hohwy, J. (2018). The predictive processing hypothesis. *The Oxford handbook of 4E cognition*, 129-146.

Hohwy, J. (2020). Self-supervision, normativity and the free energy principle. *Synthese*, 1-25.

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

Hurley, S. (2001). Perception and action: Alternative views. *Synthese, 129*(1), 3-40.

Hutto, D. D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Mit Press.

Hutto, D. D., & Myin, E. (2017). *Evolving enactivism: Basic minds meet content*. MIT press.

Hutto, D.D. (2017). Getting into predictive processing's great guessing game: bootstrap heaven or hell? *Synthese*, 1–14. doi: 10.1007/s11229-017-1385-0

Hutto, D. D. and Satne, G. (2015) The Natural Origins of Content. *Philosophia 43*(3), 521-536

Hutto, D. D. (2003). Folk psychological narratives and the case of autism. _Philosophical Papers_ 32 (3):345-361.

Järvilehto, T. (1998) The Theory of the Organism-Environment: System: I. Description of the Theory, *Integrative Physiological and Behavioral Science, 33*(4), 321-334.

Jaswal, V., & Akhtar, N. (2019). Being versus appearing socially uninterested: Challenging assumptions about social motivation in autism. *Behavioral and Brain Sciences, 42,* E82. doi:10.1017/S0140525X18001826

Jones, R. M., Southerland, A., Hamo, A., Carberry, C., Bridges, C., Nay, S., Stubbs, E., Komarow, E., Washington, C., Rehg, J. M., Lord, C., & Rozga, A. (2017). Increased Eye Contact During Conversation Compared to Play in Children With Autism. *Journal of Autism and Developmental Disorders*, *47*(3), 607. https://doi.org/10.1007/s10803-016-2981-4

Jung, P. (1993). Periodically driven stochastic systems. *Physics Reports*, *234*(4-5), 175-295.

Kiebel S. J., Daunizeau J., Friston K. J. (2008) A Hierarchy of Time-Scales and the Brain. *PLoS Comput Biol 4*(11): e1000209. https://doi.org/10.1371/journal.pcbi.1000209

Kiebel, S. J., Daunizeau, J., Friston, K. J. (2010). Perception and hierarchical dynamics. *Frontiers in Neuroinformatics 4*(12). doi: 10.3389/neuro.11.020.2009.

Kiefer, A., Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. Synthese, 195(6), 2387–2415. https://doi.org/10.1007/s11229-017-1435-7

Kiefer, A., Hohwy, J. (2019). Representation in the prediction error minimization framework. *Routledge handbook to the philosophy of psychology*, 2nd ed. Oxford, UK: Routledge.

Kirchhoff, M., (2018) 'The body in action: Predictive processing and the embodiment thesis' in Newen, A., De Bruin, L., and Gallagher, S. (eds) *Oxford Handbook of Cognition: Embodied, Extended and Enactive*. Oxford University Press.

Kirchhoff, M. (2015) Extended cognition & the causal-constitutive fallacy: In search for a diachronic and dynamical conception of constitution *Philosophy and Phenomenological Research, 90*(2): 320-360.

Kirchhoff, M. (2018). Autopoiesis, free energy, and the life-mind continuity thesis. *Synthese*, 195(6), 2519-2540.

Kirchhoff, M. (2018). Predictive brains and embodied, enactive cognition: an introduction to the special issue. *Synthese*.

Kirchhoff, M., (2018) Predictive processing, perceiving and imagining: Is to perceive to imagine, or something close to it? *Philos Stud 175*:751–767

Kirchhoff, M., Froese, T. (2017). Where There Is Life There Is Mind: In Support of a Strong Life-Mind Continuity Thesis. *Entropy, 19*(4). https://doi.org/10.3390/e19040169

Kirchhoff, M., Kiverstein, J (2019). How to determine the boundaries of the mind: a Markov blanket proposal. *Synthese*. https://doi.org/10.1007/s11229-019-02370-y

Kirchhoff M, Parr T, Palacios E, Friston K, Kiverstein J. (2018) The Markov blankets of life: autonomy, active inference and the free energy principle. J. R. Soc. Interface 15: 20170792. http://dx.doi.org/10.1098/rsif.2017.0792

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

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

Krueger, Joel & Maiese, Michelle (2018). Mental institutions, habits of mind, and an extended approach to autism. *Thaumàzein 6*:10-41.

Kuchling F, Friston K, Georgiev G, Levin M. (2019) Morphogenesis as Bayesian inference: A variational approach to pattern formation and control in complex biological systems. *Phys Life Rev.* doi: 10.1016/j.plrev.2019.06.001.

Kuenen, Ph. H. 1964 Experimental abrasion of pebbles. VI. Surf action. *Sedimentology 3*, 29–43. doi:10.1111/j.1365-3091.1964.tb00273.x

Laland, K., Matthews, B., and Feldman, W. (2016). An introduction to niche construction theory. *Evolutionary Ecology*, 30, 191-202.

Landon, R. E. 1930 An analysis of beach pebble abrasion and transportation. *J. Geol. 38*, 437–446. doi:10.1086/623739

Lee (2018) Structural representation and the two problems of content. *Mind and language,* 1-21

Lemons, D. S., Gythiel, A., & Langevin's, P. (1908). "Sur la théorie du mouvement brownien [On the theory of Brownian motion]". *CR Acad. Sci.(Paris)*, 146, 530-533.

Levin, M. (2020, July). Robot Cancer: what the bioelectrics of embryogenesis and regeneration can teach us about unconventional computing, cognition, and the software of life. *In Artificial Life Conference Proceedings* (pp. 5-5). One Rogers Street, Cambridge, MA 02142-1209 USA

Lindner, B., Garcıa-Ojalvo, J., Neiman, A., & Schimansky-Geier, L. (2004). Effects of noise in excitable systems. *Physics reports*, *392*(6), 321-424.

Linson, A., Clark, A., Ramamoorthy, S., and Friston, K. (2018) The Active Inference Approach to Ecological Perception: General Information Dynamics for Natural and Artificial Embodied Cognition. *Frontiers in Robotics and AI*, *5* (21). doi: 10.3389/frobt.2018.00021

Longo, G., Montévil, M., & Kauffman, S. (2012). No entailing laws, but enablement in the evolution of the biosphere. In Proceedings of the 14th international conference on genetic and evolutionary computation conference companion (pp. 1379–1392).

Lyons V., and Fitzgerald M., (2012). Critical Evaluation of the Concept of Autistic Creativity, In Fitzgerald, M. (ed) *Recent Advances in Autism Spectrum Disorders - Volume I*, IntechOpen. doi: 10.5772/54465

Manicka S, Levin M. (2019) Modeling somatic computation with non-neural bioelectric networks. *Scientific reports. 9*(1):18612. doi: 10.1038/s41598-019-54859-8

Markram, K., and Markram, H. (2010). The intense world theory – a unifying theory of the neurobiology of autism. *Front. Hum. Neurosci. 4*:224. doi: 10.3389/ fnhum.2010.00224

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information.* Freeman.

Marr, D. and Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London. Series B; Biological Sciences, 207*(1167): 187–217.

Martyushev, L. M., & Seleznev, V. D. (2006). Maximum entropy production principle in physics, chemistry and biology. *Physics reports*, *426*(1), 1-45.

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and Cognition—The Realization of the Living*, ser. Boston Studies on the Philosophy of Science. Dordrecht, the Netherlands.

McGeer, V. (2007). Why neuroscience matters to cognitive neuropsychology. *Synthese 159*, 347–371. doi: 10.1007/s11229-007-9234-1

McGeer, V. (2018)Scaffolding agency: A proleptic account of the reactive attitudes. *Eur J Philos.; 27*: 301– 323. https://doi.org/10.1111/ejop.12408

Milton, D. (2017). *A Mismatch of Salience: Explorations of the Nature of Autism from Theory to Practice*. West Sussex: Pavillion.

Mirski, R., Bickhard, M. H. (2019). Encodingism is not just a bad metaphor. *Behavioral and Brain Sciences*, 42.

Mirski, R., Bickhard, M. H., Eck, D., & Gut, A. (2020). Encultured minds, not error reduction minds. *Behavioral and Brain Sciences*, 43.

Mole, C. and Zhao, J. (2016) Vision and abstraction: an empirical refutation of Nico Orlandi's non-cognitivism. *Philosophical Psychology, 29*:3, 365-373

Mossio, M. and Moreno, A., (2010) Organisational Closure in Biological Organisms. *Hist. Phil. Life Sci.*, 32 (2010), 269-288

Mottron, L., Dawson, M., Soulières, I., Hubert, B., and Burack, J. (2006). Enhanced perceptual functioning in autism: an update, and eight principles of autistic perception. *J. Autism Dev. Disord. 36*, 27–43. doi: 10.1007/s10803-005- 0040-7

Moyal-Sharrock, D. (2019) From deed to word: gapless and kink-free enactivism. *Synthese,* 1-21

Myin, E. (2016). Perception as something we do. *Journal of Consciousness Studies, 23*(5–6), 80–104.

Myin, E., & Degenaar*, J*. (2014). Enactive vision. In *The Routledge handbook of embodied cognition/*Shapiro, Lawrence [edit.] (pp. 90-98).

Noë, A. (2004). *Action in perception*. MIT press.

Noë, A. (2006) Experience of the World in Time. *Analysis* 66 (1), 26-32.

Nunn, T. P. (1909-1910). Are secondary qualities independent of perception? *Proceedings of the Aristotelian Society, 10,* 191-218.

Olivotos, S., & Economou-Eliopoulos, M. (2016). Gibbs Free Energy of Formation for Selected Platinum Group Minerals (PGM). *Geosciences*, *6*(1), 2. doi:10.3390/geosciences6010002

Orlandi, N. (2012) Embedded seeing-as: Multi-stable visual perception without interpretation. *Philosophical Psychology, 25*:4, 555-573

Orlandi, N. (2013) Embedded Seeing: Vision in the Natural World. *Noûs (47)*4 727–747

Orlandi, N. (2014) *The Innocent Eye: Vision is not a cognitive process*. Oxford University Publishing

Orlandi, N. & Lee, G. (2018). How Radical is Predictive Processing? in Eds., Colombo, Irvine, & Stapleton, *Andy Clark & Critics.* Oxford University Press

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

Palacios, E. R., Razi, A., Parr, T., Kirchhoff, M., & Friston, K. (2020). On Markov blankets and hierarchical self-organisation. Journal of Theoretical Biology, 486, 110089. https://doi.org/10.1016/j.jtbi.2019.110089

Palmer, C. J., Paton, B., Kirkovski, M., Enticott, P. G., & Hohwy, J. (2015). Context sensitivity in action decreases along the autism spectrum: a predictive processing perspective. *Proceedings of the Royal Society B: Biological Sciences, 282*(1802), 20141557

Parr, T. (2020). Inferring What to Do (And What Not to). *Entropy, 22*(5), 536.

Parr, T., and Friston, K. (2017). Working memory, attention, and salience in active inference. *Scientific Reports*, 7: 14678 | DOI:10.1038/s41598-017-15249-0

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann.

Pearl, J. (2001). Bayesian networks, causal inference and knowledge discovery. *UCLA Cognitive Systems Laboratory, Technical Report*.

Pellicano, E. (2013). Sensory symptoms in autism: a blooming, buzzing confusion? *Child Dev. Perspect. 7,* 143–148. doi: 10.1111/cdep.12031

Pezzulo, G., Donnarumma, F., Iodice, P., Maisto, D. and Stoianov, I. (2017) Model-Based Approaches to Active Perception and Control. *Entropy, 19*(6), 266

Pitt, D. (2017) Mental Representation. In Zalta, E. N. (Ed.) *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition).

Polger, T. (2015) The Innocent Eye: Why Vision Is Not a Cognitive Process, by Nico Orlandi. [Review] *Analysis Reviews 75*(2), 343-345

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature neuroscience*, *16*(9), 1170–1178. https://doi.org/10.1038/nn.3495

Pourhasan, R. (2016) Non-equilibrium steady state in the hydro regime. *J. High Energ. Phys. 2016,* 5. https://doi.org/10.1007/JHEP02(2016)005

Ramsey, W. M. (2007). *Representation reconsidered*. Cambridge University Press.

Ramstead, M., Badcock, P., and Friston, K. (2018) Answering Schrödinger's question: a free energy formulation. Physics of Life Reviews, 24, 1-16

Ramstead, M., Badcock, P., and Friston, K. (2019) Variational neuroethology: Answering further questions: Reply to comments on "Answering Schrödinger's question: A free-energy formulation" Physics of Life Reviews, 24, 59-66

Ramstead, M., Kirchhoff, M. D., Constant, A., and Friston, K. (2019). Multiscale Integration: Beyond Internalism and Externalism. *Synthese*, 10.1007/s11229-019-02115-x.

Ramstead, M., Veissière, S., and Kirmayer, L. (2016) Cultural Affordances: Scaffolding Local Worlds Through Shared Intentionality and Regimes of Attention. Front. Psychol. 7:1090. Doi: 10.3389/fpsyg.2016.01090

Ramstead, M. Badcock, P., Friston, K. (2018) Answering Schrödinger'squestion- A free-energy formulation. *Physics of Life Reviews 24*, 1–16

Ramstead, M. J., Friston, K. J., & Hipólito, I. (2020). Is the free-energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy*, *22*(8), 889.

Ramstead, M. J. D., Constant, A., Badcock, P. B., & Friston, K. J. (2019). Variational ecology and the physics of sentient systems. *Physics of Life Reviews*, *31*, 188–205. https://doi.org/10.1016/j.plrev.2018.12.002

Ramstead, M. J. D., Kirchhoff, M. D., & Friston, K. J. (2019). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior, 28*(4), 225-239.

Rao and Ballard, (1999) Predictive Coding in the Visual Cortex: a Functional Interpretation of Some Extra-classical Receptive-field Effects. Nature Neuroscience, 2(1):79-87

Razi, A., & Friston, K. J. (2016). The connected brain: causality, models, and intrinsic dynamics. IEEE Signal Processing Magazine, 33(3), 14-35.

Reeke, G. N. (2019). Not just a bad metaphor, but a little piece of a big bad metaphor. *Behavioral and Brain Sciences*, *42*.

Rescorla, M. (2016). Bayesian sensorimotor psychology. *Mind & Language, 31*(1), 3–36.

Richman, KA, Bidshahri, R. Autism, theory of mind, and the reactive attitudes. *Bioethics. 32*: 43– 49. https://doi.org/10.1111/bioe.12370

Rock, I. (1983). *The logic of perception*. MIT Press.

Rosenberg, A. (2015) The Genealogy of Content or the Future of an Illusion, *Philosophia 43*(3), 537-547.

Ruiz-Mirao, Mavelli (2007) On the way towards 'basic autonomous agents': Stochastic simulations of minimal lipid–peptide cells. *BioSystems, 91*, 374-387

Satne, G., & Hutto, D. (2015). The Natural Origins of Content. *Philosophia*, *43*(3), 521–536. https://doi.org/10.1007/s11406-015-9644-0

Schrödinger, E., (1944) *What is life?* Cambridge: Cambridge University Press.

Segal, G. (1989). Seeing what is not there. *The Philosophical Review, 98*(2):189–214.

Servick, K., (2019, October 1). *Echolocation in blind people reveals the brain's adaptive powers*. Science Mag (retrieved 13th of June 2020), https://www.sciencemag.org/news/2019/10/echolocation-blind-people-reveals-brain-s-adaptive-powers

Seth, A. (2013) Interoceptive inference, emotion and the embodied self. *Trends Cogn Sci. (11)*:565-73

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal, 27*(3):379–423.

Shapiro, L. (2014) Radicalizing Enactivism: Basic Minds without Content, by Daniel D. Hutto and Erik Myin. [Review.] *Mind 123*(489), 213–220.

Shea, N. (2007). Consumers need information: Supplementing teleosemantics with an input condition. *Philosophy and Phenomenological Research, 75*, 404–435.

Shea N (2014) Exploitable isomorphism and structural representation. *Proc Aristot Soc XIV:*77–92. doi:10.1111/j.1467-9264.2014.00367.x

Shoemaker, D. (2015). *Responsibility from the Margins*, Oxford University Press

Skinner, B. F. (1953). *Science and human behavior* (No. 92904). Simon and Schuster.

Skulachev, V .P., (1992) The laws of cell energetics. *Eur. J. Biochem. 208*, 203–209.

Strawson, P. F. (2008). *Freedom and Resentment and Other Essays.* Routledge. https://doi.org/10.4324/9780203882566.

Suarez, M. (2003). Scientific representation: Against similarity and isomorphism. International Studies in the Philosophy of Science, 17(3), 225–244.

Thompson, E. (2007). Mind in life: Biology, phenomenology, and the sciences of mind. Cambridge, MA: Harvard University Press.

Tonneau, F. (2012) Metaphor and truth: A review of *Representation Reconsidered* by W. M. Ramsey, *Behavior and Philosophy, 39/40*, 331-343.

Travis, C. 2004. The silence of the senses. *Mind 113*(449): 57–94.

Tschantz, A., Baltieri, M., Seth, A. K., & Buckley, C. L. (2020, July). Scaling active inference. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

Valiyamattam, G. J., Katti, H., Chaganti, V. K., O'Haire, M. E., & Sachdeva, V. (2020). Do Animals Engage Greater Social Attention in Autism? An Eye Tracking Analysis. *Frontiers in Psychology*, *11*, 1.

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., et al. (2014). Precise minds in uncertain worlds: predictive coding in autism. *Psychol. Rev. 121*, 649–675. doi: 10.1037/a0037665

Van de Cruys, S., Kelsey, P. & Hohwy, J., (2019): Explaining hyper-sensitivity and hypo-responsivity in autism with a common predictive coding-based mechanism, *Cognitive Neuroscience*, DOI: 10.1080/17588928.2019.1594746

van Dijk, L., & Withagen, R. (2016). Temporalizing agency: Moving beyond on-and offline cognition. *Theory & Psychology*, *26*(1), 5-26.

van Es (2019) The Embedded View, its critics, and a radically non-representational solution. *Synthese*.

van Es, T. (2019). Minimizing prediction errors in predictive processing: from inconsistency to non-representationalism. *Phenomenology and the Cognitive Sciences*. https://doi.org/10.1007/s11097-019-09649-y

van Es, T. (2020). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior*. https://doi.org/10.1177/1059712320918678

van Es, T. (2020) Minimizing prediction errors in predictive processing: from inconsistency to non-representationalism. *Phenom Cogn Sci 19*, 997–1017. https://doi.org/10.1007/s11097-019-09649-y

van Es, T. and Myin, E. (2020) Predictive processing and representation: How less can be more. In Mendonça, D., Curado, M., and Gouveia, S. S. (eds) *The philosophy and science of predictive processing*. Bloomsbury.

van Es, T., & Hipólito, I. (2020). Free-Energy Principle, Computationalism and Realism: a Tragedy. *PhilSci Archive.*

van Fraassen, B. C. (1980). The scientific image. Clarendon Press.

van Fraassen, B. C. (2008). Scientific representation: Paradoxes of perspective. Oxford University Press

Van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy, 92 (*7), 345–381.

van Grunsven, J. (2020). Perceiving 'Other' Minds: Autism, 4E Cognition, and the Idea of Neurodiversity. *Journal of Consciousness Studies 27* (7-8):115-143

Varela, F. (1979) Principles of Biological Autonomy. The North-Holland Series in General Systems Research, Vol. 2. Elsevier North-Holland

Varela, F. J., Thompson, E., and Rosch, E. (1991). *The embodied mind: cognitive science and human experience*. Cambridge, MA: MIT Press.

Vitas, M., & Dobovišek, A. (2019). Towards a general definition of life. *Origins of Life and Evolution of Biospheres*, *49*(1-2), 77-88.

von Helmholtz, H. (1962). Handbuch der physiologischen optik. 1860/1962. & Trans by JPC Southall Dover English Edition.

Wallace, R. (2019). The Moral Nexus. PRINCETON; OXFORD: Princeton University Press. doi:10.2307/j.ctv3znwhn

Warren, W. (2005). Direct perception: The view from here. *Philosophical Topics, 33* (1), 335-361.

Wedlich-Söldner, R., & Betz, T. (2018). Self-organization: the fundament of cell biology.

Wiese, W. (2017) What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences 16*: 4, pp 715–736

Wiese, W. and Metzinger, T. (2017) Vanilla PP for Philosophers: A Primer on Predictive Processing in Metzinger, T. and Wiese, W. (eds.). *Philosophy and predictive processing*. Frankfurt am Main : MIND Group.

Williams, D. (2018). Predictive coding and thought. *Synthese*, 197, 1749-1775.

Williams, D. (2018) Predictive Processing and the Representation Wars. *Minds & Machines 28*: 141.

Williams, D. & Colling, L. (2018) From symbols to icons: the return of resemblance in the cognitive neuroscience revolution, *Synthese 195*:5, 1941-1967.

Yon, D., de Lange, F. P., & Press, C. (2019). The predictive brain as a stubborn scientist. Trends in cognitive sciences, 23(1), 6-8.

Zahnoun, F. (2019) On representation hungry cognition (and why we should stop feeding it). *Synthese*. https://doi.org/10.1007/s11229-019-02277-8

Zaidel, A., Goin-Kochel, R. P., & Angelaki, D. E. (2015). Self-motion perception in autism is compromised by visual noise but integrated optimally across multiple senses. *Proceedings of the National Academy of Sciences, 112*(20), 6461-6466

Zarghami, T. S., & Friston, K. J. (2020). Dynamic effective connectivity. *Neuroimage, 207*, 116453.

Ziegler, H. (1963) Some extremum principles in irreversible thermodynamics with application to continuum mechanics. In: Sneddon, I.N., Hill, R. (eds.) *Progress in Solid Mechanics,* North-Holland, Amsterdam, pp. 91–193