# Moderate Predictive Processing

Krzysztof Dolega

Recent developments in the predictive processing literature have led to the emergence of two opposing positions regarding the representational commitments of the framework (Hohwy 2013; Clark 2015; Gładziejewski 2016; Orlandi 2015). Proponents of the *conservative* approach to predictive processing claim that the explanatory power of the framework comes from postulating a rich nesting of genuine representational structures which can account for many cognitive functions (Gładziejewski 2016). Supporters of the more *radical* interpretation of predictive processing, on the other hand, postulate that not all elements of the computational architecture should be interpreted as full-blown representations, stressing the framework's connection to ecological and embodied approaches to cognition instead (Clark 2015, Orlandi 2015). Surprisingly, despite defending opposing positions, both camps seem to adopt William Ramsey's representational 'job description challenge' (Ramsey 2007) as a standard for genuine ascriptions of representational function.

The aim of this paper is to evaluate competing approaches and show that both sides of the debate must overcome additional challenges with regard to determining the representational commitments of the predictive processing framework. Following the discussion of the opposing views and the way they employ the representational job description in their arguments, I raise a worry that Ramsey's criterion may be ill suited for establishing a strong distinction between them. In light of this I propose to frame the debate between the two camps as a disagreement over the contents of generative models, rather than the functional roles fulfilled by the structures under investigation. Finally, presented with the problem of content determination and the lack of framework constrains strong enough to help resolve the disagreement, I call for moderation in making claims about predictive processing's representational status.

**Keywords**

Conservative predictive processing | Internal models | Markov blanket | Predictive coding | Radical predictive processing | Representational job description | Representations | Structural representations

Probabilistic modeling of perception and cognition is quickly becoming the leading trend in cognitive psychology and neuroscience. Hierarchical predictive coding or processing (henceforth predictive processing or PP for short) is one of the more prominent and widely discussed frameworks emerging from the recent developments in applying statistical methods to machine learning and computational neuroscience (Rao and Ballard 1999; Friston 2008; Friston 2010; Hohwy 2013; Clark 2013; Clark 2015; Clark 2016).

The framework is an attempt at formulating a unified explanation of the processes underlying perception, cognition, and action, by postulating that the brain's neural populations are organized into multiple hierarchies performing cascading statistical inferences in order to predict future inputs. This is done through the workings of an internal generative model, which aims "to capture the statistical structure of some set of observed inputs by tracking (one might say, by schematically recapitulating) the causal matrix responsible for that very structure" (Clark 2013, p. 185). The model is used to generate hypotheses (predictions or expectations) about the states of the external world and the corresponding activations of sensory peripheries, which are then tested against the actual states of the sensorium, driving perception and action in a top-down manner.

Dolega, K. (2017). Moderate Predictive Processing.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 10. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573116

In this paper I discuss two leading positions regarding the representational commitments of the predictive processing approach to cognition. So called conservative predictive processing (cPP) (Clark 2015), claims that the explanatory power of the framework comes from postulating a rich nesting of genuine representational structures which come to serve as a model of the organism's external and internal milieu (Hohwy 2016; Gładziejewski 2016). Radical predictive processing (rPP), on the other hand, postulates that not all elements of the computational architecture should be interpreted as full-blown representations (Clark 2015). Instead, proponents of this approach argue that the relevant computational level descriptions (Marr 1982) serve merely as abstract schemata aimed at capturing the dynamics of processes, which are embedded into neural structures by evolutionary selection (Orlandi 2013; Orlandi 2014; Orlandi 2015) and do not depend on manipulating genuine representations (Clark 2016; Downey 2017; Bruineberg 2017).

Surprisingly, both approaches are motivated by the adoption of William Ramsey's representational 'job description challenge' (Ramsey 2007), according to which appeals to representational posits must be supported by a convincing demonstration of the relevant elements or structures playing a representational role within the functioning of a wider cognitive system. Gładziejewski follows Ramsey in order to show that PP's generative models perform a representational function by acting as detachable, information bearing structures which stand-in for the features of the environment, in order to enable capacities such as action-guidance and error detection. Clark and Orlandi, on the other hand, claim that only higher levels of the PP hierarchy exhibit such capacities, while lower (e.g. perceptual) levels should be construed as *model-free* structures governed by biases acquired through reinforcement learning or phylogenetic development.

The main aim of this paper is to evaluate the competing approaches in relation to the underlying computational architecture and show that the disagreement between the two sides of the debate does not cut as deep as it has been presented in the literature. The main reason for this is that both sides agree on how to understand Ramsey's representational challenge and the functionally distinct notions of representation it introduces. However, an investigation into the roles played by the structures posited by PP reveals that the job description challenge may be insufficient to differentiate the competing views. Although it does not support the radical claim that peripheral layers of PP systems consist in solely non-representational elements, it also poses a serious problem for the conservative side of the debate by inviting ambiguities with regard to the ascription of representational function to more nested structures. Proponents of cPP are tasked with distinguishing internal models targeting external, environmental features from structures functioning as meta-representations modelling the behavior of other parts of the system. I propose that the difference between the competing positions must come down to the question about the content of PP's internal models. Supporters of rPP are faced with the task of introducing additional criteria which would help distinguish their position from the conservative one. Members of the cPP camp, on the other hand, can secure their interpretation by providing conditions for determining the contents of representational and meta-representational elements. Still, this is not an easy task due to the informational encapsulation of different layers of the system and the unclear conditions for identifying cases of misrepresentation.

Because this paper is part of a larger collection beginning with a primer aimed at elucidating the main tenets of PP to an uninitiated audience (Wiese and Metzinger 2017), I will skip a typical introduction to the framework. Instead, I will start by articulating Ramsey's representational job description requirement for non-vacuous ascription of representational function (section 1), followed by a brief presentation of the two competing interpretations of the framework (section 2). Having delineated the available positions, I will move on to evaluate which of the elements posited by PP should be the target for the debate over the framework's representational status (section 3.1). From there, I will argue that the representational job description challenge does not offer sufficient ground for distinguishing rPP from its conservative counterpart, by showing that it fails to secure a non-representational interpretation of the system's peripheral layers (section 3.2). This, however, does not mean that cPP cannot be

Dolega, K. (2017). Moderate Predictive Processing.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 10. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573116

contested, as proponents of this reading must face the opposite problem of functional indeterminacy in models removed from the sensory periphery (section 3.3). While, in principle, it is possible to resolve these issues by appealing to the contents of such posited representations, in practice this solution faces further difficulties, relating to a lack of clear conditions for content determination (section 3.4). I close the paper with a call for moderation in making claims regarding PP's representational status, and point to two strategies for solving the problem at the heart of the debate (section 4).

## 1 William Ramsey on the Confusion About the two Aspects of Mental Representations

The representational job description challenge (Ramsey 2007) is central to much of the recent literature concerning the representational status of probabilistic approaches to cognition. The aim of the job description challenge is simple — to provide a condition for genuinely explanatory ascriptions of representational function in cognitive science. Ramsey motivates the need for such a criterion by pointing out that our notion of mental representation must somehow be rooted in the everyday conception of what a representation is (otherwise calling such theoretical posits 'representations' would make no sense), but that our common usage fails to provide us with a clear grasp of how to individuate such entities or how to identify the role they should play in naturalistic theories of cognition. Thus, the goal is to formulate a condition that yields a scientifically valuable, yet intuitively recognizable notion of representation, one which will offer an explanation of how, or in virtue of what, particular posits function as representations. However, it is important to stress that it is not a challenge to define or describe what content is in naturalistic terms.[1]

According to Ramsey, there has been a long standing confusion between "understanding how a physical structure actually functions as a representation" as opposed to "[…] understanding the nature of the relationship in virtue of which it represents one sort of thing and not something else" (Ramsey 2016, p. 5). To clarify this confusion, he proposes to distinguish two dimensions, or aspects, of mental representation, both of which have to be fulfilled "for something to actually qualify as a full blown representation […]" (p. 6). The representational job description is set out to elucidate the first dimension — one of a representation's functional role, by providing a "set of conditions that make it the case that something is functioning as a representational state. In other words, it is the set of relations or properties that bestow upon some structure the role of representing" (p. 4). The other aspect of representations concerns *what* such structures or states represent. This dimension can be understood as "[…] a set of relations or features that bestow upon some representational structure its specific representational content" (p. 4). With this distinction in place, I will now focus on the question of what the functional dimension challenge exactly *is*.

For Ramsey, the 'job' of a representational state or entity is to *stand-in* for something external to the wider consumer system in which it is employed. Therefore, the challenge for any scientific theory is to explain how particular posits fulfill the functional role of standing-in for external features of the world within that theory or model.[2] Although this account of representational function may seem simple, it proves to be a powerful tool for assessing applications of the notion within cognitive science and philosophy.

As Gładziejewski (Gładziejewski 2016) points out, the way Ramsey develops his challenge relies heavily (though not exclusively, see section 4) on an argumentative strategy of 'comparing-to-prototype' — one examines the everyday, pre-theoretical uses of the notion in order to find a widely ac-

1 In a similar manner, the central issue of Ramsey's investigation should not be confused with the question about the possibility of describing a purely physical process as intentional (this would be a question about the possibility of adopting the intentional stance — Dennett 1987), nor should it be confused with the question concerning the indispensability of intentional ascriptions (since, following Dennett, every physical system can be described in purely causal/structural terms — a description or stance which may be more relevant for some tasks).

2 It should be noted that this is only a sufficient condition for ascriptions of the representational role to cognitive systems. See Gładziejewski's discussion of the challenge for more detail.

cepted and uncontroversial application, which then serves as a prototype to which particular uses of that notion in scientific literature are compared. For Ramsey, cartographical maps serve as an intuitive prototype for ascriptions of representational function. Following his lead, Gładziejewski (Gładziejewski 2015, Gładziejewski 2016) sets out to distinguish several features of all map and map-like devices (e.g. GPS navigation), which make them suitable for fulfilling the representational role on Ramsey's account. Firstly, maps are deemed useful in virtue of the relation (a mapping function — see footnote 3) obtaining between the physical properties or features of the map and the properties of the environment which is its representational target. It is this relation that allows for the physical artifact to stand-in for the environment during navigation and action planning, even when the depicted location is not immediately present to the user (detachability). The mapping relation also makes it possible for the map user to control her performance and adjust her actions in accordance with the available information (action guidance). Finally, it also accounts for the possibility of the representational item failing to fulfill its role, e.g. in cases of low fidelity or error in the map's structure, therefore allowing for misrepresentation and the presence of user detectable error.

Having established the prototypical case for non-vacuous representational ascription, Ramsey proceeds with the second step of his argumentative strategy — comparing different scientific applications of the concept to that prototype. It is here that his strategy proves its mettle by offering interesting and unexpected results, challenging some of the established applications of the notion of representation in philosophy and cognitive science. For the sake of brevity, I will focus on just two key examples.

The so-called 'cognitive maps' discovered in the rat's hippocampus (O'Keefe and Dostrovsky 1971; O'Keefe and Nadel 1978) and entorhinal cortex (Hafting et al. 2005) are among the favorite study cases of the proponents of the job description challenge (see e.g. Miłkowski 2015). The view that the activations of place cell neurons encode information about the two-dimensional structure of a rat's environment has been largely accepted by the scientific community, earning the scientist who made the discovery a Nobel prize. This claim has been strengthened by recent findings showing that the same neural mechanism is also processing information about the animal's future location in the environment, thus pointing to the place cells' involvement in action anticipation and planning (Van der Meer and Redish 2010; Gupta et al. 2013). Together, these developments suggest that such 'maps' are the right kind of structures to pass the representational job description challenge, as they not only have elements that correspond to features of the world, but also allow for exploiting this relation in action guidance and error correction (Pfeiffer and Foster 2013).

The notion of representation which emerges from the job description challenge and the discussion of cognitive maps is one where representation can act as a model of the world in virtue of having an internal structure which maps onto the structure of the world[3] — a notion of structural representation or S-representation.[4] As Ramsey elaborates:

> What S-representation has going for it […] is a distinctive role within a cognitive system that is recognizably representational in nature and where the state's content is relevant to that role. […] With S-representation, the fact that a given state stands for something else explains how it functions as part of a model or simulation, which in turn explains how the system performs a given cognitive task (Ramsey 2007, p. 126).

3 Importantly, this does *not* mean that the target properties must be represented by the same kind of properties in the mental representation, i.e. that the representation must be first-order isomorphic with its target. Structural representations, in the sense discussed above, are in fact standing in a second-order, functionally isomorphic relation to their representational targets (see Palmer 1978, for a detailed discussion), meaning that, for example, distance can be represented by frequency, frequency by magnitude, etc. Similarly, the analogy with cartographical maps (which *are* first-order representations) does not mean that structural mental representation must be static and cannot change in accordance with changes in the world or in the needs of the consumer system. See, for example, Wiese 2016, for a discussion of Rick Grush's emulator representations (Grush 2004) as structural representations (see also Bartels 2005, Ch. 3.).

4 Note that S-representation is not the only notion of representation that passes the representational job description challenge, see also the discussion of Wiese 2016, in section 3.4.

However, not all established uses of the notion of representation fare equally well when confronted with Ramsey's challenge. On the tradition of information theoretic accounts (most notably Dretske 1981), which were later supplemented with teleological accounts of function (e.g. Millikan 1984; Dretske 1988), biological mechanisms can be considered representational when they have a function of responding to certain environmental conditions and there is a lawful-like dependence between the signaling of the system and the behavior of the organism. This account allows for error and misrepresentation because such indicator systems can be triggered by environmental factors or properties different from the ones they are supposed to react to. It also "[…] provides Dretske with a way of showing how informational content can be explanatorily relevant. Structures are recruited as causes of motor output because they indicate certain conditions" (Ramsey 2007, p. 130). Despite its appeal, the notion of representation employed by teleofunctionalists does not meet the representational job description challenge. Ramsey points out that "there are several non-representational internal states that must, in their proper functioning, reliably respond to various states of the world […]". For example, the immune system reacts to infections, "yet no one suggests that any given immunological response (such as the production of antibodies) has the functional role of representing these infections. While nomic dependency may be an important element of a workable concept of representation, it clearly is not, by itself, sufficient to warrant viewing an internal state as a representation" (Ramsey 2007, p. 125).

Thus, Ramsey argues the notion of *detector* or *indicator* representations employed by Dretske and his followers, when treated as a criterion for ascribing the functional role for representations, threatens to trivialize the representational theory of mind. Instead he proposes (Ramsey 2016) to treat the teleological project as an account of representational content. He notes that:

> […] there is a notorious problem of content indeterminacy for any account of representation based upon structural similarity. The problem is that isomorphisms are cheap — any given map or model is going to be structurally similar to a very wide range of different things. While elements of models may function as representational proxies during various sorts of cognitive operations, exactly what they represent is impossible to determine by merely focusing on the 'structural' properties of the model or map itself (Ramsey 2016, p. 8).

This means that, even though structural properties are sufficient to determine whether a state or part of a system *could* function as a representation, a mere mapping relation is not enough to determine whether or not it carries any content. Ramsey suggests that, in order to postulate that theoretical posits are, in fact, full blown representations, constraints additional to the representational job description challenge must be presented and fulfilled.

## 2 Two Flavours of PP

The distinction between *conservative* and *radical* predictive processing has recently been introduced in to the literature by Clark (Clark 2015), who aims to distinguish between two different approaches to the framework's philosophical and scientific significance.

### 2.1 Conservative Predictive Processing

According to the conservative understanding, PP is similar to *reconstructive* views of perception, which "[…] depict our cognitive contact with the world as rooted in a kind of neuronally-encoded rich inner recapitulation of an observer-independent reality" (Clark 2015, p. 12). In this take on the framework, the cognitive system is reliant on an inner model that encodes the structure of the complex relationship between stimulations of the sensory peripheries and their distal causes, effectively becoming what Hohwy has called "an internal mirror of nature" (Hohwy 2013, p. 220). Clark describes the inner models of cPP as being employed by the system "[…] to stand-in for the external world for

the purposes of planning, reasoning, and the guidance of action" (Clark 2015, p. 12). This is similar to Gładziejewski's defense of the representational reading of predictive processing as relying on S-representations.

In his 2016 article, Gładziejewski employs Ramsey's compare-to-prototype strategy in order to argue that the generative models employed in architectures postulated by PP play a role akin to probabilistic maps of the world. By analyzing the role such models must fulfill in order to coordinate perception and behavior in ways that will minimize prediction error and keep the organism within homeostatic bounds, he comes to the conclusion that parts of the PP system responsible for generation of predictions must function as information bearing structures with features which correspond to the structure of the external environment.

To support this claim, Gładziejewski imagines a toy example of a two level PP system, consisting of a sensory periphery and a generative model. Because the system is assumed to approximate Bayesian reasoning, the prior probabilities of certain environmental occurrences taking place (e.g. of encountering a particular kind of object) should be stored in the generative model. Gładziejewski points out that the model in question must encode a set of, so called, 'hidden' or 'latent' variables which act as parameters for generating predictions about the lower stage of the system by corresponding "[…] to different likelihoods of potential patterns of activity at the lower sensory level" (Gładziejewski 2016, p. 571). Moreover, because "[…] the hidden variables are not only related to lower-level, sensory patterns, but to each other (intra-level) as well, their […] values evolve over time in mutually-interdependent ways […]" (Gładziejewski 2016, p. 572), which allows them to have a structure mirroring the dynamics of their target domain.

Together, these properties allow the PP structures to fulfill the functional role of genuine representations which work as detachable models of the environment during action guidance and error detection. Such generative models must be detachable since, on this framework, action involves an off-line computation of multiple possible strategies for minimizing error (see also section 3.1 and 3.2). This process, in turn, leads to a deployment of predictions which can drive behavior. Finally, predictions which are sub-optimal for a given context produce error signals that can be corrected in the next cycle of hypothesis testing. All this leads Gładziejewski to conclude that generative models qualify as S-representations and that the framework belongs to the tradition of computational-representational theories of mind.

## 2.2 Radical Predictive Processing

Proponents of the radical approach are not opposed to the claim that *some* levels of the probabilistic architecture do fulfill a genuinely representational role. They are, however, against viewing the framework as another iteration of the computational-representational theory of mind by stressing the embodied and embedded nature of many cognitive processes.

Once again, this is most visible in Clark's 2015 article where he claims that predictive processing should not be construed as being solely dependent on internal models of the environment. Although rPP proponents do not aim to deny that higher cognitive functions (e.g. abstract reasoning, planning, language) most likely depend on manipulating some kind of mental representations, they do stress that the appeal of the PP framework lies in reconciling such structures with 'fast and frugal' heuristics for action which emerge from the agent's dynamical coupling with the environment. Thus, the radical version of the framework is meant to highlight that "[…] sensing delivers an action-based grip upon the world, rather than a rich reconstruction apt for detached reasoning […]" (Clark 2015, p. 15). Similarly, Clark opposes conceptualizing successful behavior only as an "[…] outcome of reasoning defined over a kind of inner replica of the external world" and prefers to see it as an "[…] outcome of perception/action cycles that operate by keeping sensory stimulations within certain bounds" (Clark 2015, p. 15).

In his 2016 book, Clark adds detail to the rPP position by appealing to the distinction between 'model-based' and 'model-free' strategies for behavior selection and guidance (Dayan 2012; Dayan and Daw 2008; Wolpert et al. 2003). On this picture, which resembles the distinction present in dual-system/process literature (Frankish 2010), 'model-based' reasoning involves "[…] the acquisition and the (computationally challenging) deployment of fairly rich bodies of information concerning the structure of the task-domain, while 'model-free' approaches […] implement pre-computed 'policies' that associate actions directly with rewards, and that typically exploit simple cues and regularities while nonetheless delivering fluent, often rapid, response" (Clark 2016, p. 252). When applied to PP, this distinction boils down to the difference between the kind of posits defended by Gładziejewski, as capable of simulating and comparing multiple action plans, and reflex-like responses acquired through reinforcement learning.

In clarifying the relationship between these two kinds of processes, Clark follows (Daw et al. 2011) and suggests that different strategies can be flexibly combined together in accordance with contextual information. Acquisition and deployment of more rigid routines can be guided by information rich internal models, such as the expected precision optimization scheme associated with attention (Feldman and Friston 2010; Hohwy 2012). In such a case, "[…] a kind of meta-model (one rich in precision expectations) would be used to determine and deploy what ever resource is best in the current situation […]" (Clark 2016, p. 253).

Unfortunately, Clark remains vague about the criterion by which genuine representational processes are to be differentiated from the 'fast and frugal' ones. He stipulates that model-free processing should be associated with larger reliance on the bottom-up information, while model-dependent one would rely on top-down influence of prior knowledge. This suggests a kind of gradation from response based processes to genuinely representational ones. Unfortunately, this interesting proposal remains underdeveloped. By claiming that some, but not all, PP structures act as representations, Clark raises a challenge for proponents of the radical interpretation, who are now charged with presenting a criterion by which representational structures can be distinguished from non-representational ones. Failing to provide such a criterion, threatens the rPP view with collapsing into the conservative one.

This issue has recently been taken up by Nico Orlandi (Orlandi 2013; Orlandi 2014; Orlandi 2015), who builds on Clark's proposal by advocating the use of Ramsey's representational job description challenge for the purpose of demarcating embodied and embedded processes from representational ones. Though her focus is placed mostly on the long-standing disagreement between the ecological (Gibson 1979) and inferential (reconstructive) theories of vision (Gregory 1980; Marr 1982; Friston et al. 2012), the representational status of probabilistic models of perception occupies a prominent place in her treatment of that debate.

The main claim of Orlandi's *embedded seeing* (Orlandi 2013; Orlandi 2014) project is that vision is not an inferential process relying on manipulating intermediate states or tokens that qualify as representations. Rather, vision is a process embedded in to the biological structure of the organisms' visual apparatus, which has been molded to reliably respond to the presence of certain environmental properties through evolution and development. Computational theories of vision are merely re-descriptions aimed at capturing the dynamics of causal interactions between the physical elements of the visual system. As Orlandi explains, what follows from her Gibsonian assumptions is that probabilistic theories of vision mistakenly re-describe "[…] biased processes that operate over non-representational states" (Orlandi 2015, p. 1) as inferential. She further clarifies that on her interpretation "priors and likelihoods rather look like built-in or evolved causal intermediaries of perception that incline visual systems toward certain neuronal configurations (or certain 'hypotheses')" (Orlandi 2015, p. 25).

Orlandi motivates her radical departure from the probabilistic orthodoxy in several ways (Orlandi 2015), at least two of which seem to be relevant for understanding what her position is and how it relates to Clark's. The crux of her argument relies on questioning the restricted role of bottom-up inputs in PP schemes. By presenting a simple example of an ambiguous stimulus, such as a circle which

can appear convex or concave depending on the assumed position of the light source, she builds the case that that priors and hyper-priors alone are unable to restrict the hypothesis space of probabilistic models of perception to a single prediction. Hyper-priors, here understood as domain-general assumptions guiding the acquisition and deployment of more domain-specific hypotheses, are presented as too general (Orlandi 2015, p. 9). An appeal to a likelihood function (the probability of evidence — here the inverse of error — given the hypothesis) is similarly ruled out as, according to Orlandi, it would not restrict the hypotheses space enough to favor a single prior. This reasoning leads her to conclude that it is the sensory signal which drives and constrains the system by conveying information about the statistics of natural scenes, activating or pre-selecting the "[…] priors that should be employed even in contexts of high noise where the signal is compatible with multiple hypotheses" (p.10).

An important caveat to the above points is that they do not line up with most of the empirical evidence supporting PP's success in modeling perceptual and cognitive phenomena (see e.g. Rao and Ballard 1999; Spratling 2016).[5] Orlandi fails to appreciate the fact that many versions of the framework assume that the system's internal states come to mimic the statistics and dynamics of the environment (see also section 3.4), precisely for the purpose of disambiguating stimuli in conditions where the sensory signal is corrupted by noise (Hohwy 2012). By not paying attention to this crucial assumption, Orlandi effectively rejects the problem of perceptual inference, echoing Clark's idea that 'model-free' strategies are heavily reliant on prediction errors, which can act as a constraint on the process of hypothesis selection.

What is crucial for the present treatment is that Orlandi uses these points to motivate distancing her position from Gładziejewski's. Interestingly, she applies the same criterion he has previously used to defend the representational status of PP's generative models. According to her the early stages of the visual system are best conceived of as "mere detectors or mediators" (Orlandi 2015, pp. 23-24), meaning that they do not pass Ramsey's representational job description challenge. For example, she argues that levels directly involved in predictive coding of information about the states of the retina (e.g. cells in primary visual cortex sensitive to discontinuities in patterns of retinal stimulation traditionally associated with edge perception) are not sufficiently detached from their inputs and do not play a robust role in driving and guiding action.[6] She defends the position that, in the case of such low levels, neither predictions, nor error signals seem to have the function of carrying information about things external to the system. Errors convey only information regarding "[…] the need to adjust its own states to reach an error-free equilibrium", while predictions "are states produced for checking the level below them", which "exhausts their function" (Orlandi 2015, pp. 23-24). Following this reasoning, Orlandi claims that only the outputs of the whole visual system can pass the representational job description challenge, because they are the only part of the visual system that can be said to play the role of standing-in for the world in consumer systems realizing higher cognitive functions. This and other claims held by the rPP camp, however, may not be supported by the underlying assumptions of the framework.

5 Admittedly, Orlandi draws attention to the fact that PP models differ from the traditional constructivist views of perception in that they do not postulate early and intermediate visual states to explicitly track well-defined elements (Jehee and Ballard 2009). This point, however, does not undermine a representational reading of PP in any way. PP is a probabilistic modeling framework and it is consistent with the idea that, for example, lower-levels of the system track different targets, depending on the predictions they receive from the higher layers of the hierarchy. Orlandi fails to appreciate such a possibility as she does not engage with a large body of empirical work on top-down modulation in early visual areas (e.g., Petro et al. 2014). Nothing in the present article hinges on this point and it will not be discussed further.

6 The editors have pointed out the possibility of augmenting the rPP view by developing a graded account of representation in which the degree to which something serves as a model or a representation is, at least partially, determined by the degree of its detachment from the target. This is an intriguing suggestion, but it faces significant obstacles which deserve a separate full length treatment. For example, one of the requirements for such a view is to provide a set of conditions that would allow for defining the degree of detachability independently from the level of analysis at which the system is decomposed. Moreover, I would like to point out that this would not absolve the proponents of rPP from the task of providing a clear and unambiguous distinction between the representational and non-representational ends of the spectrum.

## 3 Three Obstacles on the Way to Determining PPs Representational Status

By now the tension between the conservative and radical treatments of PP should be visible. What is of special interest here is that the conflict at hand stems from the employment of the same conceptual tool — the representational job description challenge, in order to defend opposite positions regarding the representational commitments of the framework. One way to resolve the qualm between cPP and rPP would be to show that one of the sides misunderstands or misconstrues Ramsey's challenge. However, little in the discussion of these positions suggests such a solution, especially since the members of both camps seem to agree on how the distinction between S-representations and detector representations should be understood. Rather, it seems that the point of contention consists either in the way the challenge is applied to PP (i.e. which parts of the frameworks' computational description are submitted to Ramsey's test), or in some disagreement independent from the issue of the representations' functional role, such as the kind of contents that the relevant parts of the PP system trade in (e.g. rich or poor).[7]

In what follows, I begin by focusing on the functional dimension of PP, starting with the problem of clarifying what structures pass the representational job description. However, as the discussion progresses I will show that Ramsey's challenge is not sufficient for establishing the two positions as competing alternatives. The difference between these two views must come down to the question about the content of posited structures, rather than the question about their functional role.

### 3.1 Which Elements of PP Fulfill the Representational Job Description?

From the discussion in section 2 it can be seen that each side of the debate puts stress on different aspects of PP's computational architecture, opening possibilities for miscommunication. Clark and Gładziejewski focus on the functional role played by internal models, but it is not always clear which parts of system constitute a model and whether or not such models are coextensive with levels. Friston and Hohwy apply the notion quite loosely, easily changing between talking about a single, over-arching model of the world spanning the whole predictive hierarchy and multiple, restricted models of different cognitive domains. Finally, Orlandi seems to be preoccupied with the representational status of top-down and bottom-up messaging pathways communicating priors and errors respectively. This is why it is important to clarify which parts of the PP system are the target of the discussion.

As has been mentioned in the introduction and the supplied exposition of PP (Wiese and Metzinger 2017), the framework posits probabilistic systems organized in a hierarchical manner. Since these architectures are supposed to model the behavior of brain structures, they are usually implemented as artificial neural networks. Although competing implementations may have different assumptions about the wiring and number of functionally distinct types of nodes (compare e.g. Rao and Ballard 1999, with Spratling 2008), the general blueprint is shared by all versions of PP. It is the schema of a 'stacked' hierarchy with two distinct feed-back and feed-forward passageways connecting 'prediction estimation' (PE) groups which are assumed to correspond to specific cortical regions (Spratling forthcoming). It is these PEs that are usually referred to as 'levels' in the subject literature, and are assumed to encode the parameters of predictive models.

The PE levels are collections of different types of units themselves, including ones encoding the hidden variables used for generating predictions. Notably, when describing the make-up of these modules, Rao and Ballard include the input and output units as their components. Each level in the

7 It is also possible that the disagreement is about issues orthogonal to the problem of representation altogether. For example, Hohwy and Clark seem to disagree about the normative commitments of the framework. The first author is genuinely interested in an epistemic interpretation of PP, according to which its Bayesian roots offer a cognitive system aimed at truth, whereas the latter views the system as aimed at ecologically and evolutionarily bounded optimality. In the present treatment I am steering away from this debate for several reasons: a) the relationship between the cognitive (PP) and normative (Bayesian rationality) components of the framework is not entirely clear (see the treatment of Fink and Zednik 2017, for some ideas about that); and b) the issues discussed in this paper do not hinge on whether one assumes that the representations manipulated by the PP system are truth-capturing or just truth-approximating.

hierarchy (let's call it the *n-th* level in an ordering starting from the lowest to the highest level) is specified as receiving two kinds of inputs. The first are the descending predictions from the level above or upstream (*n + 1*), which constrain the activity on the level in question (*n*) by fulfilling a role which is equivalent to that of priors in Bayesian inference. The second kind of input received by any level is the bottom-up prediction error signal, which carries the information about the difference between the actual and predicted state of the level below or downstream (*n – 1*). Each of the PE levels is also producing two kinds of output: constraining predictions about the expected activity on the lower level (*n – 1*), and an unresolved residue of the error signal fed to the level above (*n + 1*).[8]

It is possible to restrict the discussion of PP's representational commitment to the question about the functioning of particular units. However, this would not be productive, as placing too much focus on the components of PEs can lead to a false conclusion about the functioning of the wider architecture. First of all, network implementations of PP do not differ from other artificial neural nets, which have been extensively discussed as being composed of simple detector-like elements rather than full blown representations (Ramsey 2007, p. 145). Since the nodes encoding causes of the lower level patterns of activation must reliably react to incoming inputs as well as be able to trigger appropriate activations in populations which communicate predictions, it is possible to label them as mere causal relays. Orlandi is correct in arguing that, on their own, error and prediction units do not fulfill the representational job description challenge. It does not, however, mean that more complex, representational systems cannot be constructed out of such simple elements. For example, the employment of S-representations in explanations of cognitive maps' functioning is not threatened by such maps being composed of detector-like elements. The fact that the firing rates of place cell neurons co-vary with the rat's location does not undermine the functioning of the whole structure as a detachable representation employed in controlling behavior and navigation.

Somewhat similarly, the success of certain connectionist architectures employing generative models stems exactly from the organization of their simple elements (Hinton 2007). These systems exhibit complex behavior, which cannot be fully accounted for by an appeal to mere biasing. The aptly named Helmholtz machine (Hinton et al. 1995), for example, is said to construct a generative model of its own input in a manner analogous[9] to the PP systems — i.e. by approximating Bayesian inference and learning the complex statistical regularities in a current data set in order to predict possible future inputs of the same kind. By using the so called 'wake-sleep' algorithm (Dayan et al. 1995), it operates in two alternating modes — processing inputs in a bottom-up manner during the wake phase, and optimizing its internal dynamics by a top-down generation of possible inputs in the off-line 'fantasy' phase. While it is possible to give a purely causal, adaptationist account of a passive learning process in connectionist models (Ramsey 1997), the ability to generate inputs off-line for the purpose of adjusting the internal parameters and optimizing performance seems to go beyond such simple accounts. It implies that the system can not only recapitulate the internal structure of the target domain, but also deploy it in a manner which is detached from the inputs, additionally assessing and correcting its own performance.

A non-representational account falls short in cases where top-levels of a hierarchical architecture come to encode different models of regularities present in its input sets, which are then used to guide lower levels to produce outputs that are best at capturing the structure of these input sets. A good

8 In most hierarchical predictive coding algorithms (e.g. Rao and Ballard 1999; Friston 2008) this means that the bottom-up signal communicates only the difference between the actual and predicted states of the target level, ignoring the redundant information. However, it is worth noting that in Spratling's PC/BC-DIM (Predictive Coding/Biased Competition using Divisive Input Modulation) algorithm, due to a different mapping of levels onto cortical regions, bottom-up signal carries estimates about the predicted input's causes (see also footnote 5 and Spratling forthcoming, p.5).

9 It is important to note that, although both approaches depend on the use of generative models, there are some significant differences between them. Clark 2016, p.309, stresses that Hinton's algorithms employ self-generated prediction during learning and/or optimization, not during online processing. It is also significant that Hinton's work represents the connectionist tradition of back propagation algorithms which, despite belonging to the same wider category of recurrent networks as predictive coding algorithms, can differ significantly in computational detail (even if both can be seen as approximating Bayesian inference). However, there is a growing interest in bringing these two approaches together, see e.g. Rezende et al. 2014, and Domingos 2015.

example of this is Charles Kemp & Joshua Tenenbaum's unsupervised, hierarchical Bayesian model, which can learn different forms of two dimensional graphs and then produce outputs in a form best matching the relations between the members of a given data set (see Kemp and Tenenbaum 2008, and Griffiths et al. 2010, for more details). If we agree that such outputs are structural maps of the input data, then it seems that the same should be said about the system's acquired internal models consisting of extracted input regularities and graph forming rules used to generate such outputs. After all, the main difference between them is that of format and not of function — the internal model guides the behavior of the lower levels of the system in the same way that the external graph drives the behavior of human users. Even though Kemp and Tenenbaum's model lacks the feedback-like error correction capacity that is exhibited by the Helmholtz machine and the PP models, it puts stress on views aiming to explain the behavior of probabilistic models in terms of mere biasing relations between their elements. While such a perspective can be useful in trying to decompose the system into its parts, placing sole focus on particular components of PEs could obscure how the interactions between different types of units and sub-systems contribute to the functioning of the wider system.

What is crucial for the present discussion of PP's status is that, from a formal perspective, each PE level in the hierarchy is performing the same kind of basic function — carrying out probabilistic inferences aimed at producing hypotheses, which are best at accommodating the currently available data (incorporating the information fed from the level downstream into the next estimation about the activity on that level). This is done by modelling (via hidden variables) of possible causes responsible for the obtained data, which, together with top-down information, is used for generating the most probable states of hidden variables on the level below (downward projecting prediction units), and are updated in response to the actual states of the modeled variables (backward error connections). What is crucial in this picture is that each PE is predicting the activity on the level below by building an estimate of the causes of that activity. By interacting in such a way, each level of the hierarchy is minimizing prediction error by effectively acting as a model of the level below, although this does not mean that the system *only* represents itself as Anderson 2017, suggests.

## 3.2 The Functional Role of Low-Level Prediction Estimators — A Puzzle for rPP

The above discussion of PEs calls into question Orlandi and Clark's arguments against treating low levels of the hierarchy as fulfilling a representational role.

Firstly, although composed of simple elements, PEs gain their computational prowess from the interactions of their parts and the fact that they are estimating possible causes of the states which they are supposed to model. This is especially problematic for Orlandi's claim about the non-representational nature of early perceptual stages. Levels directly predicting the states of the sensory organs do so by employing latent variables, which act as estimates of external states responsible for activations of sensory periphery. These variables play the role of model parameters for generating predictions (or 'mock inputs' as Orlandi calls them) which are tested against the actual states of the sensorium. Peripheral levels of the prediction error minimization hierarchy cannot act as mere error detectors, even if they employ simple non-representational units. To function properly they must be able to update their internal estimates in response to the error signals and the information from levels above. In other words, they must be capable of error correction in the same way that levels further up the hierarchy are — by updating a model of the hidden causes from which predictions are to be generated. This means that, despite their proximity to the periphery, PEs exploit estimates of states which are not immediately available to them and which are removed from the structures that are doing the modeling. Therefore, even if the task of the lowest levels consists in generating patterns of 'mock-stimulus' activations which will be compared against the states of sensory detectors (e.g. rods and cones), the manner in which this is fulfilled goes beyond mere error detection. In other words, the lowest levels act as models, not because they can generate 'mock inputs', but in virtue of how such inputs are generated.

Furthermore, though predictions on the lowest perceptual levels may not play a direct role in causing behavior, similar parts of the hierarchy terminating in the ventral horn of the spinal cord play a crucial role in initiating and controlling bodily actions. For example, Friston and colleagues propose that such prediction passageways carry information about the expected proprioceptive inputs (Friston et al. 2010). In cases of discrepancy between actual and predicted bodily states such predictions can, upon further processing in the spinal cord and the peripheral nervous system, effectively act as motor commands which bring bodily actuators (such as muscle spindles) into their expected states. Importantly, they do not have such an effect in virtue of being simple activation commands themselves, but by carrying information about the desired state of particular actuators, which is then translated into commands that can bring it about (this proposal can explain somatic reflex arcs and offers an outlook on developing it into a full-fledged story about goal-directed action; see also Burr 2017; Vance 2017; Limanowski 2017, for a more detailed discussion of PP's treatment of action).[10]

A similar point holds for Clark's claim regarding model-free processing. Since early processing stages are supposed to act, not only as models of stages downstream, but also as models of the environment (not to mention relying on upstream stages for the control of their internal estimates), there can never be a truly model-free process within the hierarchy. The initial proposal that some models may rely heavily on bottom-up input is equally difficult to defend, since all PP algorithms assume that bottom-up channels carry information about the difference between the predicted and actual states on the level below[11]. This signal is informative only in context of current estimates of the causes used to generate the predictions. Admittedly, Clark does motivate his position by pointing out that PP systems strive for efficient coding by simplifying and reducing the complexity of their models (FitzGerald et al. 2014). However, this can be understood, for example, in terms of generating predictions using less hidden variables or variables with fewer degrees of freedom. Such a solution does not undermine the fact that PP's processing stages are functioning as models. Instead, it would relegate the problem of differentiating rPP from cPP to the side of implementation details or questions about content of different PEs. For example, the issue would now be to explicate the number of hidden variables at lower levels or to explain which environmental features they correspond to, rather than to describe the way in which they are employed by the system. However, as I will try to show in the penultimate section of this paper, the issue of content determination in PP can be a problem even for proponents of cPP.

### 3.3 The Functional Role of High-Level Prediction Estimators — A Puzzle for cPP

The discussion of PEs, as presented so far, has confirmed Gładziejewski's claim that the PP systems' generative models do fulfill a representational role, thereby meeting the representational job description challenge. But the cPP camp should not celebrate victory just yet, as this is not the end of problems for a representational reading of PP.

Recall, that what follows from the modeling details of PP is that, from a formal perspective, each level is acting as a model of the level below. Meaning that, while producing predictions about the behavior of the nearest stage downstream, each level is also modeled by the level immediately above. This is supposed to be accomplished by each PE generating predictions from estimates of causes responsible for the behavior of the stage below. In the previous section we saw that this creates a problem

10 Orlandi claims that equating action with error correction for the purpose of explaining how internal models can fulfill the function of action guidance required by S-representations is question-begging (Orlandi 2015, p. 23). This accusation seems to lean heavily on a different claim made by Orlandi, namely, that low-level PEs are not sufficiently detached from their representational target — the body. I hope that this and the previous section are successful at elucidating why such sentiment is misguided. I would also like to point out that the view of motor control offered by PP can be traced back to popular forward-model/efference copy accounts on which the motor control system optimizes motor performance using the predicted sensory outcomes of action commands (e.g. Wolpert and Miall 1996, and Grush 2004). As Clark himself has argued, PP accounts of action production belong to the wider category of forward-models (Pickering and Clark 2014). The main difference being, that most forward-model based theories assume the models to be implemented in separate prediction modules, rather than be integrated into the parts of the system issuing motor commands.

11 Note that, although Spratling's PC/BC-DIM does forward propagate estimations of causes, its lowest levels are still in business of comparing predictions to states of sensory peripheries (see Spratling 2008, Fig.2).

for proponents of rPP as it suggests that, in levels terminating in the sensory peripheries, PEs model external causes responsible for activation patterns of sensory receptors. However, this feature of PP architectures may also be a problem for proponents of cPP. Consider, once again, an *n-th* level of the system, one which is at least 2 steps away from the periphery ($n > 2$). The PE on the *n-th* level is supposed to model the behavior of the level below in virtue of estimating the causes behind the 'observed' activity patterns of the lower level's estimator units. The question here is: what exactly are the causes of $n - 1$'s behavior, which the hidden variables of $n$ are estimations of? Do the hidden variables of $n$ compute distal environmental causes, just like in the levels terminating in the sensory receptors, or do they estimate the input $n - 1$ receives from $n - 2$?

This issue has recently surfaced in the subject literature. According to Spratling's review of PP algorithms, only some of the computational models "[…] are concerned with finding the coefficients which encode the underlying causes of the sensory data […], while others are concerned with finding these coefficients only for the purpose of calculating, and transmitting, the residual error […]" (Spratling forthcoming p.5). The reason behind Spratling's assessment seems to be that the processing stages of some PP algorithms are said to be covered by a 'Markov blanket' (Pearl 1988), where each level is said to be 'inferentially encapsulated' from all but the immediately neighboring stages, the informational states of which completely determine the total informational state of the level under investigation (see Friston 2008, for a detailed overview of the mathematical properties of his models, and Hohwy 2016, for a philosophical discussion of the consequences of such encapsulation). In cases of algorithms with such properties, discovering what exactly is being tracked by levels deep in the hierarchy can be problematic, since determining their states does not require taking any system external context into consideration. This, in turn, implies an ambiguity between levels which perform the function of standing-in for features of the environment and those which act only as meta-representations by standing-in for the features of levels downstream.

This is a serious problem for proponents of the S-representational reading of PP, since the very notion rests on the assumption that the function of representational structures is to exploit a mapping relation obtaining between said structures and the world. Proponents of cPP could attempt to relax this condition and simply claim that features of a representational structure have to map onto any other physical structure, regardless of whether it is internal or external to the larger system in which the representation is employed.[12] However, taking this route would significantly weaken the structural view and would not solve the problem at hand. It would obscure how S-representations fulfill their function, since we could no longer refer to prototypes in order to guide the ascription of representational roles (unless someone can point to an uncontroversial, commonsense example of such a structure performing a map-like function). Moreover, such a solution would inflate the problem of content indeterminacy, as it would extend the representational structures' target domain to include not only external, but also internal states of the system.

## 3.4 Structural Ambiguity of Prediction Estimators — A Puzzle for PP

Proponents of cPP may push back against the accusation of functional indeterminacy by calling upon the mapping relation ingrained into the notion of S-representation. After all, the newly presented problem of functional ambiguity questions what the levels (or rather the currently active models encoded in relevant PEs) stand-in for and not whether they perform the function of standing in for anything at all. The cPP proponents can, therefore, try to solve this problem by refocusing on the issue of content and showing that PEs do refer to the worldly causes and therefore satisfy the requirement for being S-representations. One way of accomplishing this is by pointing out that functional ambiguity

12 Interestingly, this move would bring the account of representation espoused by proponents of cPP closer to Grush's notion of emulator representation. Still, it is worth noting that supporters of S-representation take the representational job description to be merely a sufficient condition while Grush holds it (or something very close to it) as a jointly necessary and sufficient one: "[…] something is a representation if and only if it is used by some system to stand for something else, and the "stand for" is explained in terms of use" (Grush 2004, p.428).

does not mean that the levels do not come to represent some statistically relevant features of the world in a transitive way. Any level can be said to process information about the system's peripheral inputs in virtue of being fed prediction error from a lower level $n - 1$, which receives inputs from a yet lower level, and so on until level $n - x$ which receives inputs directly from the sensory periphery (which under current taxonomy should not be considered a 'level' of the predicative hierarchy). Therefore, a possible line of defense is that, each and every level is ultimately modelling changes in the sensory states. In order to predict these changes the processing stages must come to resemble the sources of the occurring stimulations. This seems to be an assumption held by Friston and Hohwy, who claim that the dynamics of the higher levels in the hierarchy come to mirror the causal dynamics of their target domain by employing the empirical Bayes method, which allows the system to extract estimates of prior probabilities from the input data through learning (Friston et al. 2012).

However, this reply seems to face at least two major problems. The first, more general worry, is that Friston's answer to the problem of content determination does not clarify how representations modeling different worldly regularities can be distinguished. The typical answer here seems to be that the structure of such PEs will somehow resemble the structure of the represented domain (this is, for example, how Gładziejewski interprets this claim, but see also Friston 2013). However, as has been pointed out by Ramsey, resemblance is a very weak constrain which cannot solve the problem of content determination on its own.

This issue has been recently addressed by Wanja Wiese who suggested that, by focusing on particular computational models and algorithms, it is possible to identify unique constrains they place on the kinds of cognitive contents that are employed in PP (Wiese 2016). As Wiese points out, computational models make specific predictions about the mathematical contents (Egan 2014) they operate on (e.g. employing continuous vs. discrete values), the kind of relations obtaining between them (e.g. the relationship between different levels in the hierarchy), and the ways they interact with each other (e.g. in the expected precision system, which modulates the bandwidth or gain on the error signal pathways, see Hohwy 2012). This allows for a higher degree of specificity with regards to the structures which are supposed to underpin different cognitive processes, introducing constraints specific to PP, which can help in identifying cognitive representations. Additionally, since the models themselves are meant to serve as functional descriptions of such processes, they must offer some initial specification and constraints on the kinds of cognitive contents involved. Finally, some PP theorists have postulated that the computational descriptions on offer have implications for subjects' phenomenology. By placing constraints on the viable phenomenological descriptions, Wiese argues, PP introduces yet more constraints on the system's contents, e.g. by making predictions about the dynamics and possible manipulations of perceptual experience.

Owing to constraints in space, I cannot do justice to Wiese's nuanced position, I will merely gesture at one possible objection to his proposal, before moving onto a more burning problem facing all of PP. The initial argument that by focusing on the mathematical details of particular algorithms, PP can offer a nuanced story about the implementation and individuation of representational structures, is compelling, However, it is not entirely clear that this will provide a grip on the environmental properties to which different parts of the system correspond. Cognitive contents used in descriptions relating cognitive functions to their underlying computational processes can be seen as serving the role of identifying the mapping between the inputs/outputs of cognitive functions and the inputs/outputs of an abstract computational description. William Ramsey has argued that such input-output representations or IO-representations pass the representational job description by playing a crucial explanatory role in computational theories of cognition. Firstly, they are used for defining the theory-independent explananda, with different competing computational descriptions acting as their explanans (Ramsey labels such IO-representations as 'external'). Secondly, they are meant to guide the process of task decomposition by specifying the start- and end-points of intermediate processing stages postulated by a theory or computational description (in which case they describe what Ramsey calls 'internal' IO-rep-

resentations, specific to the computational description under investigation). What is important here is that this notion of representation is different from S-representation, as it is not meant to accord to our everyday use of the concept, and the explanatory role it plays in our theories does not presuppose the existence of a mapping relation between the representational structure and the world (rather the mapping is supposed to obtain between the computational description and its physical vehicles manipulated by the computational mechanism). Therefore, a possible worry is that it is not clear whether this type of representation can offer an account of how S-representational contents are to be individuated, as opposed to presenting an account of implementation, explaining how we can adjudicate between adequate and inadequate computational descriptions of cognitive functions. Wiese seems to postulate that the constraints placed by the mathematical contents of the system's internal IO-representations will be sufficient to guide ascriptions of cognitive contents to corresponding S-representations. This is certainly an interesting proposal that warrants further investigation. Still, it is not clear whether it will help us determine the cognitive contents of particular structures beyond providing a general description of the kinds or types of content that portions of the hierarchy are supposed to process in order to fulfill their function (e.g. by labeling several levels as responsible for shape perception).

The second general, but closely related worry, is that it is not clear to what extent the learning process can yield variables with structural features resembling the dynamics of properties existing in the environment. Living organisms can only achieve bounded optimality. It is entirely possible that some organisms could acquire models which are useful under certain environmental and ecological constraints but which do not, in fact, correspond to easily discernible features of the world. Properties which are organism-relative, such as 'attractiveness' or 'tastiness', could serve as an example.[13] Although such properties do depend on the physical make-up of the world and the organism, they do not directly pick out any fixed features of the environment, but rather rely on an abstraction from a set of complex interdependencies obtaining between such features. This is problematic for determining contents of PP models because such properties can vary with changes in the environment or in the organism itself. For example, a model of a desirable mating partner can not only vary significantly within one species, but change due to seemingly unrelated factors (e.g. climate, access of food, presence of predators, etc.), further complicating the task of determining what exactly is the target of the system under investigation.

One way to tackle this problem, as Ramsey himself proposes, is to consider the contents of S-representations to be, at least partially, determined by more than structural similarity. Augmenting his view with a teleological account of function could yield a more restricted account of content by appealing to the representational system's etiology and the relevance of its contents for action guidance. In a similar vein, Gładziejewski suggests that action based accounts of content, such as success semantics (Blackburn 2005) or interactionist accounts of representation (Bickhard 1999), which broadly claim that contents are determined by conditions of successful action, could also provide a way out of this problem. Both of these proposals are interesting and worth further exploration within the PP framework.

However, as Gładziejewski himself notes, there may still exist cases where such augmented account would break down, especially in situations where contents do not affect organismic actions directly, but only guide perceptual processes. Therefore, a more problematic possibility is the existence of models which are adaptive, but operate on variables which do not correspond to any properties in the environment. It is possible that a PP system could acquire a false model of hidden causes which is 'good enough' at minimizing error under certain often encountered conditions. Such pragmatically driven misrepresentations could be a result of reducing internal models' complexity, e.g. in cases where regular co-occurrence of two environmental states or features is wrongly inferred as being caused by a common coefficient. In such scenarios, the model could operate on variables which do not correspond to any worldly counterpart, but are successful at generating error minimizing predictions as long as the organism is under the conditions that shaped them (see, e.g. Pliushch 2017, for a more detailed

13 I am indebted to one of the reviewers for this suggestion.

account of how this can occur in cases of self-deception). Thus, we could talk of the system being stuck in a local prediction error minimum until it is affected by an unexpected perturbation.

The above examples do not present an insurmountable obstacle for proponents of the representational interpretation of PP. As McKay and Dennett conclude in their influential treatment on misbelief: "Although survival is the only hard currency of natural selection, the exchange rate with truth is likely to be fair in most circumstances" (McKay and Dennett 2009, p. 509). Following a similar intuition, some proponents of PP point out that the long term survival chances of any organism seriously wrong about the structure of its environment will be very low. The kind of misrepresentation just discussed is likely to be an exception rather than a wide spread phenomenon. This does not mean, however, that the issue of misrepresentation is not a problem for determining the contents of systems' representations. Without a detailed account of how to differentiate the structures that map onto the features of the environment, from those that do not, cPP fails to secure a completely representational interpretation of the framework.

## 4 Conclusion — A Call for Moderation

The aim of this paper was to compare and analyze two main interpretations of the predictive processing framework with regards to its representational status. Two conclusions emerge from this discussion.

Firstly, although proponents of rPP present the discussion as a disagreement over the functional details of PP by appealing to Ramsey's job description challenge, this condition favors a representational interpretation of the framework. Therefore, if rPP is to be defined in terms of commitments about the functional role played by generative models, the position will collapse into the standard, representational reading of the framework. In order to avoid this fate, the disagreement regarding the representational status of PP should be construed as a disagreement about either the contents of PP's models, or the details of their implementation.

Secondly, applying Ramsey's job description challenge to PP architecture supports the cPP reading only if the problems stemming from its use can be resolved. Thereby, proponents of the representational view are burdened with distinguishing models which have the function of representing the world, from those that function as representations or meta-representations of the system's internal states. Defenders of cPP need to offer a detailed account of content determination, since, as Ramsey points out: an account of content and not only function is needed for something to qualify as a full blown representation.

All this calls for moderation in making claims about the representational status of PP. Yes, the framework offers a compelling account of a cognitive system composed of nested probabilistic structures which do function as models capable of self-generating their inputs. Still, it may be too early to bet money on how accurately (or if at all) different features of these models map onto the external environment. There are currently two solutions emerging in response to problems presented in the paper. The first is to double down on the search for additional (e.g. mechanistic) constrains which would help to anchor the representational claims made by cPP in implementations of particular algorithms, hoping that (as in the case of cognitive maps), once relevant physical structures are known, a strong correspondence relation between their features and some worldly domain can be established. This position is endorsed, among others, by Gładziejewski 2015, and Wiese 2016. However, not everyone is optimistic with regard to the framework's ability to provide such constraints (e.g. Miłkowski 2013, argues that PP models are at best mechanistic schemata). Those questioning the framework's explanatory scope and ability to present necessary empirical details, have called for treating probabilistic models of cognition as instrumental (Colombo and Seriès 2012), implying a similar approach to ascribing contents to entities postulated by such models. Much more work is needed to settle this debate, but regardless of its final outcome, the search for framework specific constrains on mental content is likely to exert a tremendous transformative power across the fields of cognitive studies.

## References


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

Bartels, A. (2005). *Strukturale Repräsentation*. Paderborn: mentis.

Bickhard, M. H. (1999). Interaction and representation. *Theory & Psychology, 9* (4), 435–458. https://dx.doi.org/10.1177/0959354399094001.

Blackburn, S. (2005). Success semantics. In H. Lillehammer & D. H. Mellor (Eds.) *Ramsey's legacy*. Oxford: Oxford University Press.

Bruineberg, J. (2017). Active inference and the primacy of the 'I can'. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Burr, C. (2017). Embodied decisions and the predictive brain. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36* (03), 181–204. https://dx.doi.org/10.1017/S0140525X12000477.

——— (2015). Radical predictive processing. *The Southern Journal of Philosophy, 53*, 3–27. https://dx.doi.org/10.1111/sjp.12120.

——— (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. New York: Oxford University Press.

Colombo, M. & Seriès, P. (2012). Bayes in the brain—on Bayesian modelling in neuroscience. *The British Journal for the Philosophy of Science, 63* (3), 697–723. https://dx.doi.org/10.1093/bjps/axr043.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69* (6), 1204–1215. https://dx.doi.org/10.1016/j.neuron.2011.02.027.

Dayan, P. (2012). Instrumental vigour in punishment and reward. *European Journal of Neuroscience, 35* (7), 1152–1168. https://dx.doi.org/10.1111/j.1460-9568.2012.08026.x.

Dayan, P. & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience, 8* (4), 429–453. https://dx.doi.org/10.3758/CABN.8.4.429.

Dayan, P., Hinton, G. E., Neal, R. M. & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation, 7* (5), 889–904. https://dx.doi.org/10.1162/neco.1995.7.5.889.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Downey, A. (2017). Radical sensorimotor enactivism & predictive processing. Providing a conceptual framework for the scientific study of conscious perception. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, MA: MIT Press.

Dretske, F. I. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.

Egan, F. (2014). How to think about mental content. *Philosophical Studies, 170* (1), 115–135. https://dx.doi.org/10.1007/s11098-013-0172-0.

Feldman, H. & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience, 4*, 215. https://dx.doi.org/10.3389/fnhum.2010.00215.

Fink, S. B. & Zednik, C. (2017). Meeting in the dark room: Bayesian rational analysis and hierarchical predictive coding. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

FitzGerald, T. H. B., Dolan, R. J. & Friston, K. J. (2014). Model averaging, optimal inference, and habit formation. *Frontiers in Human Neuroscience, 8*, 457. https://dx.doi.org/10.3389/fnhum.2014.00457.

Frankish, K. (2010). Dual-process and dual-system theories of reasoning. *Philosophy Compass, 5* (10), 914–926. https://dx.doi.org/10.1111/j.1747-9991.2010.00330.x.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology, 11* (4). https://dx.doi.org/doi:10.1371/journal.pcbi.1000211.

——— (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11* (2), 127–138. https://dx.doi.org/10.1038/nrn2787.

———(2013). Life as we know it. *Journal of the Royal Society, Interface, 10* (86), 20130475. https://dx.doi.org/10.1098/rsif.2013.0475.

Friston, K. J., Daunizeau, J., Kilner, J. & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics, 102* (3), 227–260. https://dx.doi.org/10.1007/s00422-010-0364-z.

Friston, K., Adams, R., Perrinet, L. & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments.

*Perception Science, 3*, 151. https://dx.doi.org/10.3389/fpsyg.2012.00151.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Hillsdale, NJ: Erlbaum.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society B: Biological Sciences, 290* (1038), 181–197. https://dx.doi.org/10.1098/rstb.1980.0090.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A. & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences, 14* (8), 357–364. https://dx.doi.org/10.1016/j.tics.2010.05.004.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences, 27* (3), 377–396.

Gupta, K., Erdem, U. M. & Hasselmo, M. E. (2013). Modeling of grid cell activity demonstrates in vivo entorhinal 'look-ahead' properties. *Neuroscience, 247*, 395–411. https://dx.doi.org/10.1016/j.neuroscience.2013.04.056.

Gładziejewski, P. (2015). Explaining cognitive phenomena with internal representations: A mechanistic perspective. *Studies in Logic, Grammar and Rhetoric, 40* (1), 63–90. https://dx.doi.org/10.1515/slgr-2015-0004.

——— (2016). Predictive coding and representationalism. *Synthese, 193* (2), 559–582. https://dx.doi.org/10.1007/s11229-015-0762-9.

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B. & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature, 436* (7052), 801–806. https://dx.doi.org/10.1038/nature03721.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences, 11* (10), 428–434. https://dx.doi.org/10.1016/j.tics.2007.09.004.

Hinton, G. E., Dayan, P., Frey, B. J. & Neal, R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science, 268* (5214), 1158–1161.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology, 3*. https://dx.doi.org/10.3389/fpsyg.2012.00096.

———(2013). *The predictive mind.* Oxford: Oxford University Press.

———(2016). The self-evidencing brain. *Noûs, 50* (2), 259–285. https://dx.doi.org/10.1111/nous.12062.

Jehee, J. F. M. & Ballard, D. H. (2009). Predictive feedback can account for biphasic responses in the lateral geniculate nucleus. *PLoS Computational Biology, 5* (5), 1-10. https://dx.doi.org/10.1371/journal.pcbi.1000373.

Kemp, C. & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, 105* (31), 10687–10692. https://dx.doi.org/10.1073/pnas.0802631105.

Limanowski, J. (2017). (Dis-)attending to the body. Action and self-experience in the active inference framework. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Marr, D. (1982). *Vision: A computational approach.* San Francisco: Freeman & Co.

McKay, R. T. & Dennett, D. C. (2009). The evolution of misbelief. *Behavioral and Brain Sciences, 32* (06), 493–510. https://dx.doi.org/10.1017/S0140525X09990975.

Metzinger, T. (2017). The problem of mental action. Predictive control without sensory sheets. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Millikan, R. (1984). *Language, thought and other biological categories.* Cambridge, MA: MIT Press.

Miłkowski, M. (2013). A mechanistic account of computational explanation in cognitive science. In M. Knauff, M. Pauen, N. Sebanz & I. Wachsmuth (Eds.) *Cooperative minds: Social interaction and group dynamics. Proceedings of the 35th annual meeting of the cognitive science society* (pp. 3050-3055). Austin, Texas: Cognitive Science Society. http://csjarchive.cogsci.rpi.edu/Proceedings/2013/papers/0545/paper0545.pdf.

———(2015). Satisfaction conditions in anticipatory mechanisms. *Biology & Philosophy, 30* (5), 709–728. https://dx.doi.org/10.1007/s10539-015-9481-3.

O'Keefe, J. & Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research, 34* (1), 171–175. https://dx.doi.org/10.1016/0006-8993(71)90358-1.

O'Keefe, J. & Nadel, L. (1978). *The hippocampus as a cognitive map.* New York: Oxford University Press.

Orlandi, N. (2013). Embedded seeing: Vision in the natural world. *Noûs, 47* (4), 727–747. https://dx.doi.org/10.1111/j.1468-0068.2011.00845.x.

———(2014). *The innocent eye: Why vision is not a cognitive process.* New York: Oxford University Press.

———(2015). Bayesian perception is ecological perception. http://mindsonline.philosophyofbrains.com/2015/session2/bayesian-perception-is-ecological-perception/.

Palmer, S. (1978). Fundamental aspects of cognitive representation. In E. Rosch & B. B. Loyd (Eds.) *Cognition and categorization* (pp. 259–303). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Fransisco: Morgan Kaufmann.

Petro, L. S., Vizioli, L. & Muckli, L. (2014). Contributions of cortical feedback to sensory processing in primary visual cortex. *Perception Science, 5*, 1223. https://dx.doi.org/10.3389/fpsyg.2014.01223.

Pfeiffer, B. E. & Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature, 497* (7447), 74–79. https://dx.doi.org/10.1038/nature12112.

Pickering, M. J. & Clark, A. (2014). Getting ahead: Forward models and their place in cognitive architecture. *Trends in Cognitive Sciences, 18* (9), 451–456. https://dx.doi.org/10.1016/j.tics.2014.05.006.

Pliushch, I. (2017). The overtone model of self-deception. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Ramsey, W. M. (1997). Do connectionist representations earn their explanatory keep? *Mind & Language, 12* (1), 34–66. https://dx.doi.org/10.1111/j.1468-0017.1997.tb00061.x.

——— (2007). *Representation reconsidered*. Cambridge: Cambridge University Press.

———(2016). Untangling two questions about mental representation. *New Ideas in Psychology, 40, Part A*. https://dx.doi.org/10.1016/j.newideapsych.2015.01.004.

Rao, R. P. N. & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2* (1), 79–87. https://dx.doi.org/10.1038/4580.

Rezende, D. J., Mohamed, S. & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *arXiv:1401.4082*.

Spratling, M. W. (2008). Reconciling predictive coding and biased competition models of cortical function. *Frontiers in Computational Neuroscience, 2*, 4. https://dx.doi.org/10.3389/neuro.10.004.2008.

——— (2016). Predictive coding as a model of cognition. *Cognitive Processing, 17* (3), 279–305. https://dx.doi.org/10.1007/s10339-016-0765-6.

——— (forthcoming). A review of predictive coding algorithms. *Brain and Cognition*. https://dx.doi.org/10.1016/j.bandc.2015.11.003.

Van der Meer, M. A. A. & Redish, A. D. (2010). Expectancies in decision making, reinforcement learning, and ventral striatum. *Frontiers in Neuroscience, 4*, 6. https://dx.doi.org/10.3389/neuro.01.006.2010.

Vance, J. (2017). Predictive processing and the architecture of action. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Wiese, W. (2016). What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences*, 1–22. https://dx.doi.org/10.1007/s11097-016-9472-0.

Wiese, W. & Metzinger, T. (2017). Vanilla PP for philosophers: A primer on predictive processing. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Wolpert, D. M. & Miall, R. C. (1996). Forward models for physiological motor control. *Neural Networks: The Official Journal of the International Neural Network Society, 9* (8), 1265–1279.

Wolpert, D. M., Doya, K. & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 358* (1431), 593–602. https://dx.doi.org/10.1098/rstb.2002.1238.