Synthese (2016) 193:559–582
DOI 10.1007/s11229-015-0762-9

Predictive coding and representationalism

Paweł Gładziejewski[1]

Received: 11 March 2015 / Accepted: 23 April 2015 / Published online: 10 May 2015

Abstract According to the predictive coding theory of cognition (PCT), brains are predictive machines that use perception and action to minimize prediction error, i.e. the discrepancy between bottom–up, externally-generated sensory signals and top–down, internally-generated sensory predictions. Many consider PCT to have an explanatory scope that is unparalleled in contemporary cognitive science and see in it a framework that could potentially provide us with a unified account of cognition. It is also commonly assumed that PCT is a representational theory of sorts, in the sense that it postulates that our cognitive contact with the world is mediated by internal representations. However, the *exact* sense in which PCT is representational remains unclear; neither is it clear that it deserves such status—that is, whether it really invokes structures that are truly and nontrivially representational in nature. In the present article, I argue that the representational pretensions of PCT are completely justified. This is because the theory postulates cognitive structures—namely action-guiding, detachable, structural models that afford representational error detection—that play genuinely representational functions within the cognitive system.

Keywords Cartographic maps · Generative models · Job description challenge · Mental representation · Predictive coding · Structural representation

1 Introduction

Historically speaking, the relationship between the development of cognitive science and our changing views on the nature and existence of mental representations

✉ Paweł Gładziejewski
pawel_gla@o2.pl

[1] Institute of Philosophy and Sociology, Polish Academy of Sciences, ul. Nowy Świat 72, 00-330 Warsaw, Poland

could be characterized as a co-evolution. On the one hand, the old philosophical idea that cognition makes use of internal representations has influenced the way in which cognitive scientists understood—and often still understand—the project of "furnishing" the behaviorist's black box. After all, the plan was to fill behaviorist's box with naturalistically-construed representations of the world. The same pattern of influence can be found in the case of antirepresentationalism, as antirepresentationalist positions in philosophy have inspired the development of antirepresentationalist frameworks in cognitive science, such as enactivism or (radical versions of) the embodied cognition approach. On the other hand, theoretical developments in cognitive science have influenced the way in which researchers understand the nature of representations, as well as their positions on whether cognition involves representations at all. For example, it seems that as the popularity of the rule-based, symbolic view of cognition waned, cognitive scientists also moved away from viewing internal representations as (functionally and semantically) analogous to beliefs, desires, and other propositional attitudes; whereas the emergence of dynamic modeling in cognitive science has inspired new attempts to formulate and defend antirepresentationalism.

The relationship between cognitive science and the notion of representation becomes even more complicated if we take into account a problem forcefully expressed by Ramsey in his book *Representation Reconsidered* (2007). There, Ramsey argues that many models and theories which are heralded as representation-invoking turn out to be representational in name only. This is because in their theorizing, cognitive scientists often use the term "representation" in such a liberal and unconstrained way that it no longer denotes structures that could be regarded as representational in any recognizable or explanatorily useful manner. These structures fail to meet what Ramsey calls the "job description challenge": it seems impossible to show that the functional roles they play in a cognitive system are truly *representational*. In such circumstances, the representational terminology too often serves as an empty and misleading ornament, devoid of any real explanatory value—a mere representational gloss on what is at its core a non-representational story about cognition.

Recently, a new and ambitious account of cognition—the predictive coding theory (henceforth PCT)—has emerged in the theoretical landscape of cognitive science, and is steadily growing in prominence. According to PCT, the brain is constantly involved in prediction-error minimization, that is, in minimizing the mismatch between internally-generated, top–down sensory signals and bottom–up sensory signals caused by the external environment. Many authors have high hopes in PCT (see Clark 2013b; Friston 2010; Hohwy 2013; Huang 2008). The theory not only promises to explain, in a mathematically elegant way, perception and action along with their interrelationship, but also to unify cognitive science as such by showing that prediction-error minimization constitutes the brain's main, or even only function. If these promises could be at least partially fulfilled, we could be on the verge of entering a new, "predictive" era in the history of cognitive science.

In the present article my aim is to investigate how this new predictive coding story relates to the ongoing debate about the nature and existence of mental representations. Granted, it is usually assumed that PCT presents us with a thoroughly representational view of the mind (see Clark 2013b; Hohwy 2013). The mind/brain is thought to owe its ability to minimize prediction error to its being equipped with a rich internal representation of the causal structure of the external world. However, the relationship between PCT and representationalism has not yet been subjected to closer, more systematic scrutiny.[1] Here I will pose two questions that, to my knowledge, the literature on PCT has not yet addressed in (enough) detail. First, I want to ask what representations actually are, according to PCT; or, what conception of representation is PCT really committed to? Second, I intend to adopt Ramsey's perspective and investigate whether representational talk is justified in PCT's case, or, whether representations as construed by PCT meet the job description challenge.

The article is structured as follows. In Sect. 2, I briefly recapitulate some of the main claims of the predictive coding account of cognition, selectively focusing on those aspects of the theory that are directly relevant to my present agenda. In Sect. 3, I flesh out Ramsey's idea of the job description challenge in more detail. In the same section, I discuss a cluster of functional properties that, I claim, characterize a prototypical non-mental representation, namely a cartographic map (in other words, I provide a job description of representations as cartographic maps). In Sect. 4, I argue that PCT postulates internal representations whose functional profile matches, to a nontrivial degree, the functional profile of cartographic maps. Put straightforwardly, according to the view on offer here, PCT postulates action-guiding, detachable, structural representations that afford representational error detection. In defending this position, I not only show how PCT views the nature of representations, but also that the representations it postulates meet the job description challenge and thus fully deserve their status. In fact, if the proposal I put forward here is right, then PCT might be as representational as cognitive-scientific theories get. Section 5 briefly summarizes the present discussion and draws some broad conclusions from it.

2 Predictive coding: a brief and selective overview

PCT rests on the general idea that in order to successfully control action, the cognitive system (the brain) has to be able to infer "from the inside" the most likely worldly causes of incoming sensory signals (Clark 2013a, b; Friston 2010; Friston and Kiebel 2009; Friston and Stephan 2007; Hohwy 2013; Huang and Rao 2011; Lee and Mumford 2003; Rao and Ballard 1999). For example, to generate an appropriate reaction, the system needs to distinguish whether an incoming flow of sensory stimulation has been caused by a tiger or by a suggestive tiger-resembling plush toy. Such a task is unavoidably burdened with uncertainty, as no one-to-one mapping from worldly causes to sensory signals exists. Depending on contextual factors, many different things in the world can cause similar sensory signals; and one and the same thing can cause a variety of different signals. According to PCT, the brain deals with this uncertainty by implementing or realizing (approximate) Bayesian reasoning. That is, incoming sensory signals are treated as data and the cognitive system attempts to settle on a "hypothesis" about their causal origins, namely that which has the highest posterior

[1] Two notable exceptions to this are: Hohwy (2013) and Rescorla (2015), although the discussion in the latter does not concern PCT specifically, but is framed in more general, Bayesian terms. When developing my own proposal later in the present paper, I will draw on some of Hohwy's (2013) ideas, specifically his views on how to apply the predictive coding framework to explaining perceptual content and misperception.

probability (i.e. probability of being true in light of the data) among alternatives, given its likelihood (i.e. the probability of obtaining the data if the hypothesis were true) and prior probability (i.e. the probability of its being true regardless of the current data).

PCT postulates that the brain performs this Bayesian reasoning by employing a generative model of the sensory signal (Clark 2013a, b; Hohwy 2013; Huang and Rao 2011; Lee and Mumford 2003; Rao and Ballard 1999). The generative model "recapitulates" the causal–probabilistic structure of the external environment that impinges on the organism's sensory apparatus. It constantly generates, in a top–down manner, a flow of virtual or mock sensory signals that predicts the unfolding of sensory signals generated by external causes. This process is supposed to minimize prediction error, defined as the degree of mismatch between the two signals (where larger mismatch means larger error). Intuitively speaking, prediction error is most successfully minimized if the system selects a correct hypothesis about the causal etiology of incoming sensory data and uses this hypothesis as the basis for its top–down predictions. For example, assuming that what one is in fact observing is a tiger, then the hypothesis that the inflow of sensory information has been caused by a tiger will generate (on average) smaller prediction error than alternative hypotheses—including any that attribute the causal origins of the incoming signal to a plush toy or a domestic cat. Importantly, the size of the prediction error is always modulated by the precision that the system attributes, within a given context, to sensory signals coming from the world (Clark 2013b; Hohwy 2013). Crudely speaking, the higher the estimated precision, the more reliable the signal is treated to be, and the more reliable it is treated, the greater the (precision weighted) prediction error that is generated (and so the more likely it is that the system will initiate the hypothesis revision process).

It is assumed that there are two ways for a cognitive system to minimize prediction error (Clark 2013a, b; Friston 2010; Friston and Kiebel 2009; Friston and Stephan 2007; Hohwy 2013). The first simply consists in selecting a hypothesis (or revising a previously selected hypothesis) that is as effective as possible in minimizing the error. The second strategy is to *act* in a way that leads to a better fit between the incoming signal and internally-generated predictions (although this connection to action was missing in the original formulations of the predictive coding story and was only added later on; see Clark 2013b). This could, for example, be achieved by engaging in an action that makes a given hypothesis come true. The first strategy is called "perceptual inference", while the second is called "active inference". As such, perceiving the world (perceptual inference) and acting on it (active inference) turn out to be two sides of the same coin—the same process of minimizing prediction error.[2] Our cognitive

[2] One thing needs to be noted in order to avoid terminological (and potentially conceptual) confusion. In the present article, I follow philosophical literature (e.g. Hohwy 2013, 2014; Clark 2013a) and identify active inference with action, construed as one of two possible strategies of minimizing prediction error. However, in some foundational and prominent formulations of the Bayesian/predictive approach to cognition (particularly those rooted in statistical physics), the term "active inference" is also used in a wider sense (see this point raised in: Friston 2013a). There, the notion of active inference is introduced to refer to a more fundamental imperative for self-organization (based on minimizing variational free energy or the entropy of actively sampled sensory states). Active inference thus understood subsumes both action and perception, and the processing scheme that PCT introduces to explain cognition is treated as its corollary. Importantly, even though here I use the term "active inference" in a more restricted sense, I am sympathetic to the general

engagement with the environment rests on a constant oscillation between perceptual and active inference.

Here I will discuss what is perhaps the most theoretically ambitious version of PCT, namely that which situates predictive coding within the larger context of theoretical biology (Friston 2010, 2013a, b; Friston and Kiebel 2009; Friston and Stephan 2007; Hohwy 2013). From this point of view, prediction error minimization is, essentially, a tool for self-organisation. Minimizing prediction error serves as an indirect way of controlling actions so that the organism avoids circumstances that are surprising (or, more technically, have a high surprisal) relative to its phenotype. These are the circumstances that bring the organism nearer to thermodynamic equilibrium, i.e. increase its thermodynamic entropy (disorder). Prolonged food deprivation or sustaining a severe bodily injury are good examples of such situations. On this version of PCT, the generative model is equipped with high-level prior expectations (priors) that bound the organism's predicted sensory states. The system thus expects to find itself within a limited set of circumstances, namely those that have low surprisal. This way, the size of the prediction error serves as an approximation of how surprising a given situation is—the larger the error, the larger the surprisal. Perception and action are two strategies for making sure that the sensory signal coming from the world stays within the expected bounds, enabling the organism to avoid disorganization and retain integrity over time. In simpler terms, the way the whole prediction error minimization machinery works is not neutral from the point of view of the "interests" of an organism as a self-organising entity. Agents expect to retain their integrity and engage in active inferences (guided by perceptual inference) that aim at making those expectations true. For example, if we assume that being in close proximity to a large predator such as a tiger is surprising relative to a human phenotype, then selecting the hypothesis (through perceptual inference) that a tiger is nearby could trigger the prediction that one will *not* find oneself in such a situation—a prediction that could be fulfilled by engaging in active inference of the appropriate kind. (This, of course, is a fancy, predictive-coding-based description of a situation in which someone starts running away upon seeing a tiger.) Fundamentally, then, minimizing prediction error serves (broadly speaking) adaptive action. As I will try to show later in this article, this link between minimizing prediction error and adaptively acting in the world is significant if we want to understand how PCT views representations.

3 Representations, predictive coding, and the job description challenge

3.1 Ramsey's job description challenge and the compare-to-prototype strategy

As mentioned at the beginning of this paper, my main aim here is to investigate the connection between PCT and the idea that cognition is mediated by internal representations. My approach will be largely inspired by Ramsey's (2007) critical assessment of the way in which the concept of representation functions in cognitive-scientific

Footnote 2 continued
idea about the primacy of action in explaining why cognitive systems use the predictive processing scheme as described by PCT (see main text).

theorizing. Ramsey builds on the general observation that the notion of representation is currently used in a highly unconstrained way, such that many researchers are willing to treat as a representation virtually any structure that mediates between environmental stimuli and behavior. He argues that if we want to avoid over-application (and the resulting trivialization) of the concept of representation, then invoking representations as explanatory or theoretical posits should not come cheap. This is where the job description challenge comes in. Roughly speaking, the challenge is to show, in detail, how or in what sense a structure or state that features as a representation in a given cognitive model or theory serves a genuinely representational role inside the cognitive system; for example, in what sense its "job" in a cognitive system consists in *standing-in* for external states of affairs, as opposed to playing some other, non-representational function. If the job description challenge is not met, then according to Ramsey we are simply dealing with a non-representational structure undeservedly and confusedly called "representation" by cognitive scientists.

Throughout his book, Ramsey (2007) submits a variety of notions of representation routinely used by cognitive scientists to the aforementioned challenge. To this end, he often employs an argumentative strategy that for present purposes we may call the "compare-to-prototype" strategy. Here, one starts out by pointing out a type of structure that can be *pretheoretically* categorized as a representation in an uncontroversial way. In particular, one concentrates on the functions served by the structure in question—on what it *does* for its users that makes it a representation. This is our representational prototype. Subsequently, one concentrates on a particular concept of representation used in cognitive science and verifies whether structures that fall under this concept have a functional profile that matches, to a non-trivial degree, the functional profile of the pretheoretical prototype. In other words, one asks whether a given type of representation posited by cognitive scientists plays a functional role that is similar enough to the role played by the prototype that the former can be naturally regarded as (truly, non-trivially, genuinely, etc.) representational in nature. If it does play this role, then the job description challenge is successfully met.[3]

Interestingly, more often than not deployment of the compare-to-prototype strategy leads Ramsey to negative conclusions with respect to the different conceptions of

[3] It might be objected that the compare-to-prototype strategy is too conceptually conservative, since it gives excessive credit to our commonsense, pretheoretical intuitions about the nature of representation. Isn't it the case that this prevents cognitive scientists from developing truly novel and strictly technical ways of understanding representation, ones that are alien to our pretheoretical intuitions? This worry can be answered in two ways (see also, for an in-depth discussion of this issue, the opening chapter of Ramsey 2007). First, the prototype-based approach does not have to be the *only* way of evaluating cognitive-scientific notions of representation. In fact, it does not appear to be the only strategy employed by Ramsey (2007) himself (see e.g. his discussion of input–output- or IO-representations). *I* certainly do not wish to claim or imply that this is the only strategy available. Thus, while being functionally similar to a representational prototype might be *sufficient* for meeting the job description challenge, it is not *necessary*. Second, the compare-to-prototype strategy does not preclude that cognitive-scientific representations may—in addition to properties that they share with their prototypical analogues—have interesting properties that are *not* exemplified by any structures pretheoretically recognized as representations. Furthermore, whatever functional properties representations in the cognitive-scientific sense share with the prototypes, they may exemplify those properties in a somewhat different manner or sense (see Sect. 4). This way, while our pretheoretical intuitions may anchor the scientific use of representational notions, by no means do they fully determine or put strong limitations upon it.

representation found in cognitive science. For example, Ramsey (2007) goes to great lengths to show that there is a discrepancy between the way prototypical covariance-based representations (such as fuel gauges or compass needles) work and the workings of internal, cognitive-scientific (purported) representations construed as receptors or detectors (such us the famous bug detectors in a frog's visual cortex; see Lettvin et al. 1959). In a painstakingly detailed analysis, Ramsey argues that while the former play a role that consists in *informing* someone about something, the latter turn out to be nothing more than causal mediators between a worldly state of affairs and some behavioral reaction (for example, between the presence of a bug at a particular point in space and the frog snapping its tongue in that direction). Given that causal mediation is surely not the same as representing something—at least not in any natural or explanatorily useful sense of "representing"—it follows that internal receptors do not meet the job description challenge and as such turn out to be representations in name only (see also a somewhat similar point raised in: Hutto and Myin 2013).

Now, I do not want to dwell on whether or not Ramsey's critique of what he calls the "receptor notion" of representation is conclusive. I mention it for other reasons. First, in order to show how the compare-to-prototype strategy might lead one to rather revisionist conclusions—that is, how one might use it to argue that it is at the very least problematic whether particular ways of understanding representations meet the job description challenge. Second, I intended to illustrate an argumentative strategy that I am about to apply to the notion of representation that, I think, is present in PCT.

3.2 Representations in PCT and the things that maps do for us

Let us return to PCT, then. As I have already indicated, it is usually implicitly assumed that the theory presents us with a representationalist outlook on the nature of mind or cognition. Authors who write about predictive coding are clearly inclined to use representational terminology, as exemplified by claims that PCT postulates "probabilistic representations" of the world; that the neurons purportedly involved in generating the stream of top–down predictions are "representation units"; or that the process of prediction error minimization aims to minimize the mismatch between how things are and how the brain/mind "represents" them as being (see e.g. Clark 2013b; Hohwy 2013, 2014; Friston and Stephan 2007). Unfortunately, as a rule philosophers and cognitive scientist do not explicate what exactly they mean by "representation" in such contexts (however, see note 1), let alone dissect whether the use of such terminology is even justified. As things stand, it might still turn out that representational talk is not used here to convey any deep theses about the nature of the mind–world relation, but instead plays the role of a mere heuristic device—admittedly helpful when it comes to illuminating the basics of PCT for newcomers, albeit nothing more.

My aim now is to apply Ramsey-style methodological suspiciousness to the representational pretensions of PCT. I will argue for two claims. First, I want to put forward a proposal about how we ought to interpret the representational commitments of predictive coding. I will argue that PCT provides us with a conception of mental representations as a particular sort of structural or S-representations, whereby the representation itself and what is represented are related in terms of structural similarity.

I will lay this idea out in detail in the next section. Second, I want to argue that representations thus understood do not share the fate that Ramsey attributes to internal receptors. In other words, PCT postulates cognitive structures that *do* meet the job description challenge.

Let me start by shedding more light on the argument for the second of the two aforementioned theses. In line with the compare-to-prototype strategy, I want to draw an analogy between (what I claim to be) the representational posits of PCT and a particular type of prototypical (and non-mental) representational structure. Namely, I claim that representations as PCT sees them earn their status because the functional roles they play non-trivially resemble the roles played by *cartographic maps*. In other words, the representational posits of PCT fit the same job description as cartographic maps. Thus, maps serve as a representational prototype in my argumentation—a sort of 'golden standard' of representation. For this reason, it will be beneficial to focus on maps and *their* job description—on what they do for us, their users, *qua* representations. I propose that cartographic maps have four functional properties that are constitutive of them playing a role of representations. Maps are (1) structural representations that (2) guide the actions of their users, (3) do so in a detachable way and (4) allow their users to detect representational errors.[4] I will now look into each of those four properties in turn.

Maps belong to the category of structural or S-representations (see Cummins 1989; O'Brien and Opie 2004; Ramsey 2007; Shea 2014; Swoyer 1991). Representations of this sort are characterized by the fact that they represent in virtue of sharing, to at least some degree, their relational structure with whatever they represent. Here, I will draw on Gerard O'Brien and Jon Opie's definition of structural similarity:

> Suppose $SV = (V, \Re V)$ is a system comprising a set V of objects, and a set $\Re V$ of relations defined on the members of V. [...] We will say that there is a second-order [structural – PG] resemblance between two systems $SV = (V, \Re V)$ and $SO = (O, \Re O)$ if, for at least some objects in V and some relations in $\Re V$, there is a one-to-one mapping from V to O and a one-to-one mapping from $\Re V$ to $\Re O$ such that when a relation in $\Re V$ holds of objects in V, the corresponding relation in $\Re O$ holds of the corresponding objects in O. (O'Brien and Opie 2004, p. 11)

In the cartographic case, the pattern of (at least some) spatial (e.g. metric) relations holding between (at least some) constituents of the map preserves the pattern of corresponding spatial relations holding between corresponding constituents of the represented terrain.[5] Assume that we are dealing with a more or less adequate map, in which points A', B' and C' correspond, respectively, to buildings A, B, and C on a given terrain. Based on this, we might for example conclude that if A' is located

[4] I do not mean to suggest that maps *exhaust* the category of action-guiding, detachable, structural representations that afford error detection. It seems that other representational artifacts—such as scale models or diagrams—can belong to this category as well. My choice of maps in particular here is largely dictated by rhetorical purposes. I simply suspect that most people are so familiar with the cultural practice of map-based-navigation that cartographic maps can serve as a particularly clear and simple example in the present discussion.

[5] Here I am simplifying a little for the sake of the discussion, as maps can also represent spatial relations (e.g. elevation above sea level) through non-spatial relations (e.g. using color gradients).

closer to B' than to C', then building A is closer to building B than C; if A' is situated *between* B' and C', then A is located between B and C; and so on (see O'Brien and Opie 2004).

The following issue might be raised in the present context. I promised to discuss the *functional* properties of maps. However, while structural similarity is a relational property of a map, it does not seem to be a properly functional property. To answer this worry, it needs to be stressed that by structural similarity I do not simply mean a relation that may or may not hold between the map and the represented terrain. Rather, I mean a relation that is, in addition, *relevant to whether a map can properly perform its representational function*. In other words, the idea is that in the case of maps, structural similarity is an *exploitable relation* (see Shea 2007, 2013a, 2014). By this I mean that the map's proper functioning as a representation nonacidentally depends upon whether the structural similarity holds—or on the degree to which it holds—between the map itself and the represented terrain.

To get a better grip on the above idea, we should focus now on the second functional property that I attribute to maps, namely the fact that they serve their users by providing them with guidance for actions (see Anderson and Rosenberg 2008; Bickhard 1999, 2004). On this hopefully quite uncontroversial view, a cartographic map acts as a surrogate or stand-in, enabling its user to successfully act with respect to the terrain for which it stands in. When we traverse an unknown city, a map can guide our decisions about which path to choose, where to turn and when, which directions to avoid, etc. This saves us the inconvenience (if not the pain) of trying to achieve our practical aims through blind luck, by chaotically trying out different possible paths. In addition, a map can guide our *cognitive* actions, as is the case, for example, when we consult one in order to form true judgments about relative distances between elements of the terrain, or to try out alternative paths to a given destination in a purely counterfactual manner (as when we engage in a sort of armchair travelling).

Now, when I say that structural similarity is relevant for a map's proper functioning, I mean that map-users exploit or, perhaps more precisely, *depend on* this relation when they use maps to guide their actions. Whenever we navigate, make decisions, or form judgments on the basis of a map, our *success* nonacidentally depends on whether the map structurally resembles the terrain, or on the degree to which it resembles it.[6] A map that actually mirrors, to the required degree, the terrain it represents enables us to meet our destinations, avoid obstacles, or form true beliefs about the spatial layout of the terrain. An outdated map whose structure no longer resembles the structure of the terrain (or resembles it only very poorly) will usually lead us astray.

The third functional property of a cartographic map consists in the fact that it can perform its action-guiding duties off-line or in a detached way, i.e. even when the

[6] This is not to suggest that a map is valuable as a representation only under the condition that it mirrors the terrain in all of its tiniest structural detail. Good maps are usually full of idealizations and slight distortions. The extent to which a map should resemble the terrain it represents depends on its practical application; or, in other words, on the sort of action for which it is supposed to provide guidance. An architect planning the spatial layout of a future housing estate will require a map much more detailed than that needed by a tourist as she walks through an unknown city. This does not imply, however, that the architect's map is "better" in some absolute sense. Rather, it is better only relative to the engineer's practical interests, as many of its details would constitute confusing and useless noise for the tourist.

represented terrain (or certain parts thereof) is not directly present or available to the map's user (see Clark and Grush 1999; Grush 1997; Haugeland 1991). This detachment can be either *partial* or *complete*. As an example of partial detachment, think of an electronic map connected to a GPS system that you use to navigate your way through an alien city. I propose that in a case like this, the functioning of the map can be characterized as detached for two reasons. First, your ongoing navigational decisions (about which path to choose, when and in which direction to turn, etc.) are not directly controlled by some sort of causal coupling with the terrain itself, but rather depend on the contents of the map construed as a representational stand-in. Second, this process is anticipatory or proactive in nature, as it enables you to perform "mental leaps" into the future in order to react to turns or obstacles that are not presently visible, but which will become present to you within a relatively short period of time (sometimes just a couple of seconds). As advertised, this is a rather minimalistic sort of detachment. After all, the whole process of map-use takes place in the larger context of a direct, ongoing interaction with the city that one is traversing. Also, although our navigational decisions crucially depend on what the map tells us, they are continuously verified and, if needs be, corrected in light of what we see in front of us.

Importantly, a map can also work off-line in a much stronger sense—this is the complete detachment mentioned above. Complete detachment includes cases where one uses a map outside of any direct interactions with the represented terrain. While comfortably sitting in an apartment in a European city, I could reach for a map of Tokyo. Obviously, in such circumstances the map is of no use when it comes to satisfying my immediate navigational needs—after all, I am not currently taking a stroll through the streets of Tokyo. However, the map may still guide my *cognitive* actions. I could investigate it in order to learn about the distance between Tokyo's districts, about the layout of its streets, the locations of important buildings, and so on. This way, the map could also prepare me for *future* travels through the city of Tokyo, should I ever happen to visit.

The last functional property of maps *qua* representational devices is that they enable their users to *detect* representational errors. Maps, unsurprisingly, happen to be inaccurate from time to time. Now, what I want point to here is the fact that it is possible for us, the users of maps, to detect cartographic misrepresentation. Of course, it is usually far beyond us to simply hop aboard a helicopter and *directly* compare, from high up, the structure of a map with the structure of a terrain. This does not preclude us, however, from having epistemic access to whether a map misrepresents or not. We obtain this access *indirectly*, in virtue of how maps fit into our *practical* engagements with the world (see Anderson and Rosenberg 2008; Bickhard 1999, 2004). Inaccurate maps prevent us from meeting our destinations, prolong our journeys, and, in general terms, prove to be more of an obstacle than a help. *These* practical failures we seem to have access to. As such, it is possible to identify an inaccurate map not by comparing it to a terrain, but by recognizing failures of *actions* that are guided by it. Metaphorically speaking, we can estimate a map's inaccuracies by the number of times we bump into things while using it as a guide for our actions.

Before I move on, let me make the following observation: it seems that two different sorts of representational error are possible when it comes to cartographic maps. The first of these arises when a map as such misrepresents the terrain, i.e. when it is so structurally dissimilar to a terrain that it fails as a provider of action guidance. The second type of error arises when we, as users, inaccurately locate our own position within the terrain represented by the map. Just imagine a person trying to find her way through New York using a perfectly accurate map, but mistakenly thinking she is walking through Upper Manhattan, when in fact she is located in Brooklyn. Both types of error can lead to action failures, and both can be detected this way. As a result, however, action failure as such cannot discriminate regarding which type of error has been committed on a given occasion—this is an interpretational dilemma with which we have to deal with as map-users.

4 Representations in PCT as action-guiding, detachable, structural models that afford representational error detection

4.1 Preliminary remarks

I propose that PCT postulates internal structures whose functioning inside a cognitive system closely resembles the functioning of cartographic maps. It might be said that on the proposed interpretation of the theory, cognitive systems navigate their actions through the use of a sort of causal–probabilistic "maps" of the world. These maps play the role of representations within the theory. Specifically, this map-like role is played by the generative models. It is generative models that, similarly to maps, constitute action-guiding, detachable, structural representations that afford representational error detection.

As indicated before, I propose that if the abovementioned claim is on point, then this also answers the question of whether or not the representational pretensions of PCT are justified. Because (what I claim to be) the representational posits of PCT non-trivially resemble, at a functional level, the workings of a prototypical representation such as a cartographic map, the former meet the job description challenge. Put bluntly, we are dealing with full-blown internal representations here, and not just cheap imitations thereof.

Of course, the accuracy of the second point mentioned above rests on the accuracy of the first, i.e. on the claim that PCT postulates structural representations that guide action in a way that is detachable and affords representational error detection. The aim of this section is to elucidate and motivate this thesis. Before I proceed to doing this, however, two clarifications of my position are in order.

First, it should be clear that I do not wish to suggest that maps and the generative models postulated by PCT work in exactly the same way. While generative models have each of the four functional properties that I earlier attributed to cartographic maps, they exemplify those properties in a slightly different way or sense. For example, even though on my view both maps and generative models represent in virtue of a structural similarity, the relata of the similarity relation are construed differently in each case. Needless to say, contrary to maps, generative models clearly do not mirror in their *spatial* structure the *spatial* structure of the environment. Other, more subtle differences will transpire in the discussion to follow. However, I think that discrepancies of this kind never run deep enough to undercut the general analogy to which I am pointing here.

Second, let me stress how important it is for my case that generative models are characterized by *all four* of the functional properties in question. I suspect that if I put forward an argument that only attributed a subset of those features to generative models, then my proposal might be open to trivializing counterexamples (viz. cases where a clearly non-representational structure nevertheless satisfies my proposed functional description, thus undercutting the representational interpretation of PCT). For example, an exploitable structural similarity will not do by itself when it comes to making something into a representation. No one would consider a key to represent a lock simply because in order to properly perform its function, the former needs to be isomorphic to the latter (Tonneau 2012). Detachability combined with action-guidance does not seem to fare much better. Off-line control of adaptive engagements with environment can be achieved by structures that cannot be considered representational in any explanatorily useful and non-trivial sense (Chemero 2009; see also Calvo Garzón and García Rodríguez 2010). Now, again, on the view on offer here, what constitutes the representational status of generative models is not that they exemplify some subset of the properties that are constitutive of maps as representations, but rather that they exemplify all four of them. This, I think, renders the notion of representation embedded in PCT robust enough to be immune—at least provisionally, until proven otherwise—to counterexamples or trivializing arguments found in the literature.

With these preliminary remarks out of the way, let me get to the point and explain how the generative models postulated by PCT exemplify the four prototypically representational functional properties. I will discuss each property in turn.

4.2 Structural similarity

To see how PCT postulates structural representations (S-representations), let us consider the following simplified but illustrative example of an imaginary cognitive system equipped with a two-level generative model.[7] The system in question possesses certain sensory apparatus and inhabits a world populated by medium-sized physical objects that undergo patterned causal interactions. Importantly, these causal patterns are probabilistic: whether and how an object X will causally affect an object Y always has a

[7] That is, the model considered here will be comprised of two levels: (1) the level of hidden variables and parameters and (2) the sensory level that registers sensory input. The hidden level minimizes prediction error with respect to sensory input generated by the world at the sensory level. This is a drastic simplification of how things work in real life. The generative models postulated by PCT are richly hierarchically structured, with each level exclusively engaged in minimizing prediction error at the level directly below it (Clark 2013b; Hohwy 2013; Lee and Mumford 2003). This way, the distinction between the "sensory level" and the "hidden level" is relative to each pair of levels located directly over/beneath each other (a level that works as sensory relative to the level above it will provide hidden predictive variables for the level directly below it). In addition, in such a scheme, different layers of the model track causal patterns that appear at different temporal scales. To illustrate that Spring will come after Winter is predicted by a level placed much higher in the hierarchy than the level(s) engaged in predicting rapid changes pertaining to, for example, contours or color hues of currently perceived objects. It is through the use of such hierarchical processing scheme that prediction error minimization becomes computationally tractable, as the brain is not required to directly predict minute perceptual details on the basis of high-level, abstract hypotheses about the world.

certain context-dependent probability (even tigers eat people only with certain probability). Now, these causal–probabilistic regularities in the external world determine statistical patterns in the system's sensory input. The statistics of the incoming sensory signals serve as the only system-available "trace" of the causal–probabilistic structure of the external world. From the point of view of the predictive coding framework, our cognitive system uses the statistics of the input in order to build up an internal, skull-bound model of the external world that produces this input. Put otherwise, assuming there is a function that probabilistically maps states of the world (external causes of the sensory signal, i.e. objects with their patterned interactions) to states of the system's sensorium, then the job is to internally recreate this function by constructing a model-surrogate of the external environment. This model is generative in that it works as a sort of experience simulator or, as Andy Clark (2013a, p. 476) poetically puts it, a "virtual reality generator". That is, the internal model of the world constantly activates the system's sensory level top–down in a way that is supposed to (predictively) simulate the sensory activity produced by the external world. The better the simulation, the smaller the prediction error.

This generative model can be understood as a sort of brain-implemented statistical or Bayesian network (see Pearl 2000) whose structure resembles the causal–probabilistic structure of our system's environment.[8] More specifically, the top level of the model contains a structure of hidden variables and parameters that maps onto the causal–probabilistic structure of the environment. The model structurally resembles the world in three dimensions.

First, hidden or latent variables in the model encode the probability distributions of obtaining lower-level observations. In other words, different variables correspond to different *likelihoods* of potential patterns of activity at the lower, sensory level. The idea is that there is a structure-preserving mapping from hidden variables to causes in the world in terms of those likelihoods (see e.g. Kemp and Tenenbaum 2008; Tenenbaum et al. 2011). Worldly causes are thus represented in terms of the likelihoods of producing different sensory patterns in the system. For example, from our imaginary cognitive system's perspective, tigers might be characterized by the probability of causing, in a system specific, correlated patterns of sensory activity (hearing roaring sounds, seeing sharp teeth and black stripes on an orange background, etc.). The internal generative model captures this world–sensorium relationship in virtue of its having a hidden variable that is correspondingly related (through model parameter values) to lower-level sensory patterns.[9]

[8] The structural similarity relation is understood here in accordance with the O'Brien and Opie's (2004) definition, mentioned earlier. The thing about structural resemblance is that it is weaker than homomorphism or isomorphism. This, I think, is an advantage, as it seems that characterizing the relation between generative models and the world in terms of one of the latter relations would be too restrictive (this, by the way, also applies to cartographic maps; see O'Brien and Opie 2004). It is reasonable to assume that generative models are selective and idealized—they are not literally *mirrors* of the causal–probabilistic structures of the world. Rather, they are restricted to recapitulating the action-relevant Umwelt of a given (type of) system (Clark 2013b). And those selective and idealized models can be effective at minimizing prediction error, just as selective and idealized maps can be effective at providing action guidance.

[9] Notice that in more biologically realistic generative models that are comprised of not one, but multiple levels of hidden variables built on top of each other (see note 7), the situation will be more complicated. Namely, elements of the world's causal–probabilistic structure will not only be characterized by how they are

Second, 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, which are defined by the model parameters. This is how the *dynamics* of causal–probabilistic relations between worldly objects are encoded in the model. In other words, the pattern of diachronic relationships in the hidden layer of the model resembles the pattern of diachronic relationships in the external world. Again, take our imaginary simple system that makes use of a two-level generative model. If there is some causal–probabilistic relation between tigers and humans (say, such that upon contact, the former tend to engage in certain behaviors and the latter react by engaging in a different sorts of behaviors), then some corresponding relation should hold between corresponding hidden model variables. This way, tigers are represented not only in terms of the sensory states they are likely to produce, but also in terms of their causal interactions with other things in the world. Also notice how the fact that the top–down sensory signal predicts the unfolding of incoming sensory signal is made possible by the fact that the dynamics of the model's hidden layer (which underlie predictions) capture the dynamics of causal interactions in the world (which underlie incoming signals).[10]

Third, since our system is supposed to realize Bayesian reasoning, there is one other aspect that the model structure and the environment structure should have in common. Namely, *prior* probabilities of worldly causes should be encoded in the generative model (see Clark 2013b; Friston and Kiebel 2009; Hohwy 2013). That is, the model's hidden variables should be distinguished from one another in a way that maps onto distinctions between prior probabilities that characterize worldly causes. Take our illustrative simple cognitive system. If it is to minimize prediction error by realizing Bayesian reasoning, its generative model should be sensitive to the probability of the system's stumbling upon a given type of object, *regardless* of the current context or current incoming sensory signal. Supposing that, in the world that our imaginary system inhabits, it is far more likely to see a tiger than a plush tiger-resembling toy, the generative model should exhibit a corresponding preference for the "tiger" hypothesis (the hidden variable value that corresponds to this hypothesis) over the "tiger-resembling-toy" hypothesis.

The general point, then, is that representational posits of PCT turn out to be internal S-representations. The present treatment is obviously very sketchy. It postulates that representations as PCT views them are related to what they represent by way of structural similarity. We have a general understanding of what the represented structure is. It is the causal–probabilistic structure of the environment. What we need in order for this picture to be complete are specifics about the relevant structure of the *representation itself*. How exactly are the likelihoods, dynamics, and priors encoded or implemented in the nervous tissue? This is a vast and, to some degree, open sub-

Footnote 9 continued
probabilistically related to lower-level causal regularities, but also by how they fit into *higher*-level, slower regularities. As Karl Friston (2003, p. 1343) puts it: "[…] if the causal structure of generative processes is hierarchical, this will be reflected, literally, by the hierarchical architectures trying to minimise prediction error, not just at the level of sensory input but at all levels".

[10] For a sophisticated and biologically plausible example of how a generative model can recreate birdsongs by simulating their underlying causal dynamics, see (Friston and Kiebel 2009).

ject. Suffice it to say that there is a growing body of empirical and theoretical work that (directly or indirectly) addresses the question of neural implementation of PCT, both when it comes to more macro-level issues regarding hierarchical message passing within the brain (see e.g. Bastos et al. 2012; Kanai et al. 2015), as well as more micro-level (and non-hierarchical) issues regarding how the brain encodes probability distributions (see e.g. Pouget et al. 2013; Vilares et al. 2012). I will not delve into those detailed discussions regarding neural implementation, as they are beyond the scope of present discussion. My aim here is simply to elucidate PCT's commitments as far as the nature of mental representations goes. And, to repeat, it seems clear that PCT is committed to the claim that the brain implements a structural model of the world.[11]

Before I move on to discuss other functional features of generative models, let me address a following worry for the S-representational reading of PCT. According to information-theoretic formulations of the predictive coding framework, prediction error minimization is a process that increases the *mutual information* between internal states of the cognitive system and the states of the external world (Hohwy 2013). This means, informally speaking, that the better the system is at minimizing error, the stronger the *correlation* or *co-variance* between its internal states (viz. perceptual hypotheses or hidden variables that generate predictions) and the external causes of its sensory signals. Now, this fact could well inspire a reading of the representational commitments of PCT that might be considered orthogonal to that for which I am arguing (or one that might at least render my treatment incomplete). That is, one could say that PCT invokes co-variance-based *receptor* representations. After all, if a cognitive system works by having its internal machinery reliably co-vary with environmental states of affairs, should we not conclude that it makes use of the internal receptors of those states of affairs? And, if so, is there any reason whatsoever to favor the interpretation that appeals to S-representations instead of receptors? Or maybe PCT really posits both?

I think that this "receptor" reading of the representational posits of PCT is untenable. While PCT does postulate a co-variance between the states of the system and the states of the world, the crucial thing to notice here is *how this correlation gets established*. In the case of receptor representations, the assumption is that the co-variance in question arises in virtue of a causal relation between the (purported) representation and some worldly state of affairs, where the latter reliably causally affects the former, all things being equal. This, however, is *not* how organisms maximize mutual information on the predictive coding view of things. According to PCT, the perceptual hypothesis selection/revision process does not result from a bottom–up receptor-style causal connection to the world. Rather, it involves a top–down *endogenous* activity of

[11] In fact, it might be suggested that one could strengthen the argument for the claim that PCT posits S-representations by giving it *a priori* flavor. As an anonymous reviewer of this paper pointed out, mathematically, prediction error that is minimized is the same as (negative) model evidence, where model evidence is defined as the probability of sensory samples (inputs) given an agent's model of its world. The process of prediction error minimization leads to selection of models that have the greatest evidence, in terms of their statistical structure, in light of sensory samples of the world. This means that the very imperative to minimize prediction error *mandates* a structural mapping (similarity) between the generative model and the causal–probabilistic structure of the world. The model's structure is "sculpted" by the causal–probabilistic structure of the world—by minimizing prediction error, the former is (necessarily) made to match the latter.

simulating future sensory inflow. Furthermore, the effectiveness of generative models at internally predicting the sensory signal depends on how well the model's structure resembles—in the aforementioned sense—the environment (see the next subsection for an elucidation of this point). It is this internally-driven process of exploiting the structural similarity between the internal model and the environment that subserves the maximization of mutual information (i.e. co-variance between hypotheses and states of the world). Representations *qua* receptors or detectors simply do not enter this story. More fundamentally, it seems like the receptor reading of the representational commitments of PCT would miss the whole point of how the theory views the nature of perception. PCT presents us with a view of perception as a Kantian in spirit, "spontaneous" interpretative activity, and not a process of passively building up percepts from inputs. In fact, on a popular formulation of PCT, the bottom–up signal that is propagated up the processing hierarchy does not encode environmental stimuli, but *only* signifies the size of the discrepancy between the predicted and actual input (Clark 2013b; Hohwy 2013). On such a view, although the world itself surely affects perception—after all, the size of the bottom–up error signal is partly dependent on sensory stimulation—its influence is merely corrective and consists in indirectly "motivating" the system, if needed, to revise its perceptual hypotheses.

4.3 Action guidance

Analogously to cartographic maps, structural similarity is an *exploitable* relation when it comes to generative models (see Shea 2007, 2013a, 2014). This means that the generative model's proper functioning is nonacidentally dependent on the degree to which it resembles the causal–probabilistic structure of the world. What is this function which is so dependent on the similarity relation? I propose that it consists—again, by analogy to maps—in providing *guidance for action*. Of course, it is customary to attribute to generative models the function of minimizing prediction error, and not guiding actions. Let us remember, however, that minimizing prediction error is not an aim in itself—at least not in the version of PCT on which the present discussion is based (see Sect. 2). Organisms do not minimize the discrepancy between incoming and predicted sensory signals just for the sake of it. The whole error-minimization ordeal serves a more fundamental purpose, namely that of steering the organism's activity in the world in a way that allows it to avoid situations that have high surprisal and thus endanger its thermodynamic integrity (Friston 2010, 2013a, b; Friston and Kiebel 2009; Friston and Stephan 2007; Hohwy 2013).

Generative models, then, guide action in the following technical sense. "Action" here is equated with active inference.[12] Thus, acting consists in actively intervening

[12] Notice that this is a different sense of "action" than the one to which we refer when we say that cartographic maps guide action. In the latter sense, actions are understood commonsensically, as, very roughly speaking, activities resulting from a subject's desires or intentions. In the context of PCT, however, actions are identified with a technical category of active inference, i.e. with endogenously-controlled activities that aim at prediction-error minimization. This way of understating action does not presuppose personal-level intentional or representational categories. This is good news given the fact that we need to avoid homuncular fallacy.

in the world in a way that is supposed to minimize prediction error, and this way keep the organism within the expected bounds. Now, as mentioned in Sect. 2, there is constant interaction between active and perceptual inference. The cognitive system continuously predicts—on the basis of perceptual inference—how the incoming signal will unfold depending on the actions taken, and then engages in those actions (in active inference) to confirm its perception-based predictions. Crucially, the dependence between active and perceptual inference is not only causal, but *functional* as well. The *success* of active inference at minimizing prediction error nonacidentally and reliably depends on the accuracy of perceptual inference. Notice that the accuracy of perceptual inference, in turn, depends on how well the generative model captures—in terms of likelihoods, dynamics, and priors—the causal–probabilistic structure of the external world; the more accurate the model, the more accurate the hypotheses it produces. Just imagine a (human) cognitive system that—due to the fact that it uses a less-than-accurate generative model of the world—settles on the hypothesis that it is seeing a plush imitation of a tiger (viz. selects a model variable that corresponds to this hypothesis as a basis for its sensory predictions), when what it in fact faces is a live tiger. This sort of situation will most likely lead the system to engage in active inference that is drastically inefficient at minimizing prediction error—and, by extension, at keeping the system within its expected states.

To summarize this part, as Jakob Hohwy (2013, p. 91) puts it, “engaging in active inference on the basis of an inaccurate conception of what the world is like is unlikely to land the organism in unsurprising states in the long run”. The tighter the structural similarity between the model and the causal–probabilistic structure of the world, the smaller the prediction error that the active inference generates, on average and in the long run[13]; and so, by extension, the less likely it is that the organism will find itself in circumstances characterized by high surprisal. On such a view, adaptively acting in the world requires having a good model (S-representation) of the world. This is analogous to how the chances of our actions succeeding depend on whether (or, rather, on the degree to which) the map that we have at our disposal resembles the terrain.[14]

[13] This caveat is significant. As observed by Hohwy (2013), inaccurate hypotheses sometimes happen to be quite effective at minimizing prediction error. However, the probability of this being the case decreases with a larger number of samples. For example, even though miscategorization of a tiger as a plush tiger-toy can once in a while minimize error effectively, the average prediction error it will produce over more interactions with tigers is large (or, larger than the average error produced by an accurate categorization).

[14] Even though it explicitly ties representations to action, the present treatment is nonetheless concerned with indicative representations, understood as representations with mind-to-world direction of fit. This is why I associate deployment of representations with perceptual inference, i.e. the process of figuring out what the causes of incoming sensory stimuli actually are. However, it seems that the representational interpretation of PCT could be extended to encompass imperative representations with world-to-mind direction of fit (see also Hohwy 2013; Shea 2013b). Rather than fixing a hypothesis about an actual cause of the sensory signal (in perceptual inference), the cognitive system could select a (false) hypothesis and then engage in active inference that minimizes prediction error induced by this hypothesis. For example, upon seeing a tiger, the system could activate a hypothesis that it finds itself in safe, tiger-free circumstances, and then engage in active inference that minimizes prediction error by making this hypothesis true (of course, succeeding at this would presumably require a lot of guidance from perceptual inference on the way). In such a case, expecting that the system is in safe circumstances would be analogous to placing a cross on a specific point on a map not with an intention to mark one’s own actual position within the terrain (“I am here”), but rather to set a destination point (“I need to go there”).

4.4 Detachability

Similarly to maps, generative models work in a way that is detachable from the environment. This idea lies at the very core of the PCT, given that the theory views the nature of generative-model-based navigation as an endogenously controlled and inherently forward-looking process. From this perspective, generative models guide actions off-line rather analogously to the way in which we navigate a car using an electronic map connected to a GPS system. First, similarly to such maps, it is generative models—and not the world itself—that serve as a locus of control for the actions in which cognitive systems engage. That is, on PCT view of things, active inferences are dictated by endogenously-generated hypotheses about causes in the external world, and not by some direct coupling with the environment itself. Second, analogously to GPS navigation, the whole process of model-based action guidance is future-oriented. When we walk about the world, our generative models constantly generate a cascade of sensory predictions about what will happen *next*—about as-yet unactualized possibilities located in the future (say, about how the tiger's fur will feel when stroked). Generative models guide action by constituting map-like internal stand-ins for worldly states of affairs.

However, the above form of detachment is only partial—again in a sense similar to how the maps that underlie GPS-based navigation work only partly off-line. Action guidance through generative models is inherently attuned to actual states of the world. After all, perceptual hypotheses are constantly and swiftly verified in action. That is, even though the system guides its actions on the basis of internally-driven predictions, it constantly uses the actual sensory signal coming from the world in order to correct and, if needs be, revise its hypotheses (and modify actions accordingly). Generative-model-based interactions with the world require this sort of constant corrective feedback from outside. Our internal models may look into the future, but their functioning is nonetheless bound to on-line interactions with the environment.

It is only natural to ask whether the analogy to cartographic maps runs even deeper here. Can generative model work *completely* off-line, outside of any direct interaction with the environment? Can the internal models that PCT postulates free cognitive systems from the confines of their immediate practical dealings with world, enabling them to represent states of affairs located in the past, distant future, or even purely counterfactual? This is where things become more speculative. As things stand, PCT is a theory of perception and action, and the extent to which it scales up to domains of cognition that are completely off-line remains an open matter. Nonetheless, I think there are some preliminary reasons for thinking that the representational posits of PCT could turn out to work in a completely detached way.

First, we have known for a long time by now that brain areas normally engaged in perception and the control of action are activated during wholly off-line processes, such as imagery (Moulton and Kosslyn 2009) and concept use (Martin 2007). At the same time, some researchers working within a broadly Bayesian approach to cognition attempt to model human off-line cognitive activity as a sort of probabilistic simulation of causal processes in the world (Goodman et al. 2015; see also Chater and Oaksford 2013). The simulations in question can be *productive* in that they do not necessarily correspond to any past perception—a trait traditionally considered a hallmark of uniquely conceptual thought. These two strains of evidence can be combined into the following hypothesis. Perhaps off-line activation of perceptual and action areas results from an off-line mental simulation of causal processes in the world realized by the generative model. A simulation of this sort—presumably originating at relatively high or abstract levels of the generative model—could generate a cascade of top–down sensory signals, activating levels relatively low within the model hierarchy. This way, generative models could run simulations of possible scenarios that span multiple levels of the processing hierarchy and bring about patterns of neural activity akin to those that accompany perception and action (see Clark 2013b). From this perspective, then, off-line thinking would consist in modeling worldly causal processes in a way that involves (something like) imagery.

Second, at least two notable attempts at *directly* connecting the predictive coding framework with off-line cognition have been made so far. On the one hand, some authors have introduced an idea that generative models encompass a particular sort of *counterfactual* representations (Friston et al. 2012; Seth 2014). Namely, generative models are said to encode how the sensory signal and its precision *would* change under a range of actions that that are merely possible and do not have to be actually executed. On this view, organisms can engage in a sort of fictive, counterfactual sensory sampling of the world, and this process could underlie the ability to preselect actions; the assumption is that those actions will be executed which are predicted by the model to be most effective at reducing uncertainty about external causes of the sensory signal (although the idea of counterfactual sampling may have other applications as well, for example when it comes to explaining the phenomenal experience of perceptual presence; see Seth 2014). On the other hand, it has been proposed that PCT can explain multiple aspects of REM dreaming (Hobson and Friston 2012; Hobson et al. 2014). Roughly speaking, the claim here is that during REM sleep, the brain uses generative models of the world to construct off-line simulations of counterfactual scenarios in a way that is totally freed from the "sensory enslavement" of signals coming from the world. Dreaming thus construed is thought to play a crucial role in reducing the complexity of the generative model (e.g. by getting rid of redundant parameters), which makes it more efficient at minimizing prediction error during waking.

All in all, although the jury is still out on this, possibility of generative models turning out to be completely detachable representations appears relatively high.

4.5 Representational error detection

Last, generative models postulated by PCT resemble cartographic maps in that they allow for representational error detection. To unpack this idea, one should probably begin by explaining what one means by representational error in the context of PCT. Accounting for representational error in naturalistic terms is a hefty philosophical undertaking. For the sake of the present discussion, let me draw a following sketch, inspired in part by Hohwy's (2013) treatment of the issue. I propose that representational error arises whenever (1) a hypothesis has been selected through perceptual inference whose location within the model's structure does not correspond to the location of the cause of the incoming signal within the causal–probabilistic structure of the environment (i.e. the system selects wrong variables as a basis for its sensory predictions); (2) as a result, the cognitive system engages in active inference that does not, on average and in the long run, minimize prediction error effectively[15], given the type of circumstances that the system is in.[16] Again, take the familiar example: you stand in front of a tiger but a hypothesis is selected (on the basis of perceptual inference) according to which the incoming signal has been caused by a plush imitation of a tiger. Given that this model-based perceptual hypothesis does not correspond to how real tigers are placed within the causal nexus of the world, active inferences guided by it—say, an attempt to hug the presumed toy—will be prone to generating large prediction error; and they would produce large error on average, in the long run.

Interestingly, misrepresentation thus understood can arise in two ways. First, it may originate from the inaccuracy of the model itself. That is, the model gets its likelihoods, dynamics, or priors—or some mixture of the three—wrong, and for this reason it fails as an action guider. For example, this might be the case when the generative model that the system uses does not distinguish between real tigers and artificial tiger-resembling toys. This type of error is analogous to a situation in which one walks around an alien city using an inaccurate (say, drastically outdated or underdeveloped) cartographic map. Second, misrepresentation might arise when the system is in possession of an accurate model, but nonetheless misapplies it, for example due to the fact that the incoming sensory signal is unreliable. In such a situation, the system uses a correct representation of the causal–probabilistic structure of the world, but as a result of situational factors, it wrongly locates its own position within this structure, so to speak. This could happen when one miscategorizes a tiger as a plush imitation due to the degraded or noisy nature of the incoming sensory signal. An error of this sort is analogous to a situation in which a person makes use of a perfectly accurate map, but nonetheless wrongly locates her own position within the represented terrain (i.e. when she misapplies the map).

Now, brains or cognitive systems have no way of adopting a view from the "outside" in order to directly ensure that the generative models they use structurally resemble the causal organization of the world (see Bickhard 1999, 2004). Nonetheless, it seems like PCT allows for the possibility of representational error detection. Again, similarly to cartographic maps, within the predictive coding framework, misrepresentation can be

[15] As should be apparent from this, misrepresentation thus construed may not be an all-or-nothing matter; rather, it probably comes in degrees that are proportional to the size of the prediction error generated by a hypothesis, given the circumstances (see Hohwy 2013). Unsurprisingly at this point, we can draw an analogy to maps here (and perhaps to S-representations in general, see Cummins 1989), as they also seem to be (in)accurate in a way that admits gradation.

[16] In simpler terms, misrepresentation arises when an action fails, and it fails as a result of being misguided by a representation. This approach partly accounts for *representational* error through *pragmatic* error, in a way that is somewhat akin to teleosemantics (see e.g. Millikan 1984; Papineau 1987), so called "success semantics" (see Blackburn 2010; Nanay 2013), and interactivist or action-centric approaches to representation (see Anderson and Rosenberg 2008; Bickhard 1999, 2004). Importantly, this is not to say that that active inference has to be the *only* thing that suffers from misrepresentation. As noted by one of the anonymous reviewers, presumably a case could be made that misrepresentation also negatively affects longer term perceptual inference (i.e. shorter term misrepresentation—especially if not immediately corrected in light of the prediction error—may hurt the accuracy of perceptual inference over longer periods of time).

detected in virtue of how cognitive systems try out their models of the world *in action*. Generative-model-based perceptual hypotheses are constantly tested through active inference. Representational error is recognized in this scheme not by comparing the model with the world, but indirectly, on the basis of the bottom–up signal than signifies the size of the *prediction error* produced by ongoing practical engagements with the environment (Clark 2013b; Friston and Kiebel 2009; Hohwy 2013). The size of the error indirectly establishes how far the system has strayed from the task of maintaining itself in its expected states—that is, how (un)successful its actions (active inferences) have been. Remember, though, that the success of active inference depends on the accuracy of hypotheses formed thorough perceptual inference. Given this, the size of prediction error signifies for the system whether (or to what extent) *it got things wrong representationally*.[17] This is apparent in the fact that when the error signal reaches a certain (precision weighted) threshold, it causes the system to re-engage in perceptual inference so as to ensure that prediction error is minimized more effectively.

Last, notice that—once again, by analogy to cartographic maps—action failure itself (understood here as unsuccessful active inference) cannot discriminate between the two sources of misrepresentation discussed earlier. In other words, cognitive systems face an interpretational dilemma with respect to whether the failure of prediction error minimization is due to model *inaccuracy* or model *misapplication*. Interestingly enough, it seems as if PCT has the conceptual resources to differentiate situations in which cognitive systems, so to speak, decide on one of these interpretations. Recognition of the first type of error—that which results from inaccuracies in the generative model itself—corresponds to situations where the system modifies not its current perceptual hypothesis, but the hypothesis space (the generative model) as such. This seems to be the case when the structure of the model parameters is modified and adjusted in light of the prediction error (a process of Bayesian perceptual learning; see Clark 2013b; Friston 2003) or when the system attempts to minimize the overall complexity of the model (Hobson and Friston 2012; Hobson et al. 2014). Recognition of the second type of error corresponds, straightforwardly, to perceptual inference—an ongoing process in which the system revises its current hypotheses (and not the model itself) about the causal etiology of an incoming sensory signal.

5 Conclusions

The aim of this article was to critically examine the relationship between the notion of internal representation and the predictive coding theory of cognition (PCT). More specifically, my goals here were twofold: (1) to flesh out in detail the notion of representation embedded in PCT, i.e. to explicate what representations actually are, according to this theory; (2) to evaluate whether representations, as postulated by PCT, meet the job description challenge, i.e. whether they actually earn their representational status in virtue of the roles they play within the cognitive system. With respect to the first

[17] It is important to note, however, that misrepresentation is not the *only* way that action may fail. For example, even when the world is correctly represented, action may still fail if the divergence between the predicted and actual sensory signal is not calculated correctly (I thank an anonymous reviewer for pointing this out). Because of this, the prediction error is only a *fallible* sign of misrepresentation.

problem, I argued that representations that the PCT postulates should be understood as action-guiding, detachable S-representations that afford representational error detection. With respect to the second problem, I claimed that since the functional profile of representations as posited by PCT closely resembles the functional profile of prototypical representations such as cartographic maps, the former meet the job description challenge and thus fully deserve their status. In other words, PCT explains cognition by postulating internal structures that are genuinely and nontrivially representational in nature.

As mentioned at the beginning of this article, the development of cognitive science and the changing status of the representational view of the mind are historically intertwined. This means, among other things, that changes in the theoretical landscape of cognitive science affect both the way in which researchers construe the nature of representations and the extent to which they view the mind as representational in the first place. In the introduction to this paper, I also mentioned that PCT is considered by many authors to be the "next big thing" in cognitive–scientific theory; one that might dominate, or at least play a major role in the discipline in the years to come. It is only natural to expect that a theory this far-reaching in its explanatory ambitions might affect the fate of the idea that cognition involves internal representations. I think that, from the theses defended here, we can draw two general conclusions about how PCT might shape future debate over the nature and existence of mental representations.

First, what I hope transpires from this discussion is that PCT can inspire new formulations of the representationalist view of the mind and cognition. The theory provides us with what I think is a novel, rich, and remarkably robust notion of representations as a sort of probabilistic maps of the world that the organisms depend on in their practical and cognitive engagements with the environment. This way of understanding representation strikes me not only as fruitful in its own right, but also as immune to trivializing arguments that have been raised against many contemporary versions of representationalism (see Chemero 2009; Hutto and Myin 2013; Ramsey 2007). Second, I suspect that PCT might affect the configuration of power in the contemporary representationalism/antirepresentationalism debate—obviously in a way that favors representationalism. If the predictive coding framework proves successful at unifying our conception of cognition as well as at explaining particular cognitive phenomena, this should be welcomed as good news by proponents of representationalism. To the extent that brains are in fact in the business of minimizing uncertainty through the use of generative models, they are thoroughly representational systems as well.

Acknowledgments Work on this paper was supported by the Polish National Science Centre FUGA 3 grant (UMO-2014/12/S/HS1/00343). I wish to thank three anonymous reviewers for this journal for their helpful comments and suggestions. I also have benefitted from conversations with Marcin Miłkowski and Przemysław Nowakowski, as well as from discussing a previous version of this paper at the *Philosophy of Cognitive Science* seminar at the Institute of Philosophy and Sociology of the Polish Academy of Sciences.

References

Anderson, M. L., & Rosenberg, G. (2008). Content and action: The guidance theory of representation. *Journal of Mind and Behavior*, *29*, 55–86.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*, 695–711.

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

Bickhard, M. H. (2004). The dynamic emergence of representation. In H. Clapin, P. Staines, & P. Slezak (Eds.), *Representation in mind: New approaches to mental representation* (pp. 71–90). Oxford: Elsevier.

Blackburn, S. (2010). Success semantics. In D. H. Mellor & H. Lillehammer (Eds.), *Ramsey's legacy* (pp. 22–36). Oxford: Oxford University Press.

Calvo Garzón, P., & García Rodríguez, A. (2010). Is cognition a matter of representations? Emulation, teleology, and time-keeping in biological systems. *Adaptive Behavior*, *18*, 400–415.

Chater, N., & Oaksford, M. (2013). Programs as causal models: Speculations on mental programs and mental representation. *Cognitive Science*, *37*, 1171–1191.

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

Clark, A. (2013a). Expecting the world: Perception, prediction and the origins of human knowledge. *The Journal of Philosophy, CX*, 469–496.

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

Clark, A., & Grush, R. (1999). Towards a cognitive robotics. *Adaptive Behavior*, *7*, 5–16.

Cummins, R. (1989). *Meaning and mental representation*. Cambridge, MA: The MIT Press.

Friston, K. J. (2003). Learning and inference in the brain. *Neural Networks*, *16*, 1325–1352.

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

Friston, K. J. (2013a). Active inference and free energy. *Behavioral and Brain Sciences*, *36*, 32–33.

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

Friston, K. J., Adams, R. R., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology*, *3*, 151. doi:10.3389/fpsyg.2012.00151.

Friston, K. J., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of Royal Society B*, *364*, 1211–1221.

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

Goodman, N. D., Tenenbaum, J. D., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In E. Margolis & S. Laurence (Eds.), *The conceptual mind: New directions in the study of concepts*. Cambridge, MA: The MIT Press.

Grush, R. (1997). The architecture of representation. *Philosophical Psychology*, *10*, 5–23.

Haugeland, J. (1991). Representational genera. In W. Ramsey, S. Stich, & D. Rumelhart (Eds.), *Philosophy and connectionist theory*. Hillsdale, NJ: Erlbaum.

Hobson, J. A., & Friston, K. J. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology*, *98*, 82–98.

Hobson, J. A., Huang, C. C.-H., & Friston, K. J. (2014). Virtual reality and consciousness inference in dreaming. *Frontiers in Psychology*, *5*, 1133. doi:10.3389/fpsyg.2014.01133.

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

Hohwy, J. (2014). The self-evidencing brain. *Noûs*. doi:10.1111/nous.12062.

Huang, G. T. (2008). Is this a unified brain theory? *New Scientist*, *2658*, 30–33.

Huang, Y., & Rao, R. P. N. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*, 580–593.

Hutto, D. D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Cambridge, MA: The MIT Press.

Kanai, R., Komura, Y., Shipp, S., & Friston, K. J. (2015). Cerebral hierarchies: Predictive processing, precision and the pulvinar. *Philosophical Transactions of the Royal Society B*, *370*, Article ID 20140169. doi:10.1098/rstb.2014.0169.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*, 10687–10692.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of Optical Society of America*, *20*, 1434–1448.

Lettvin, J., Maturana, H., McCulloch, W., & Pitts, W. (1959). What the frog's eye tells the frog's brain. *Proceedings of the Institute of Radio Engineers*, *47*, 1940–1951.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, *58*, 25–45.

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

Moulton, S. T., & Kosslyn, S. M. (2009). Imagining predictions: Mental imagery as mental emulation. *Philosophical Transactions of the Royal Society B*, *364*, 1273–1280.

Nanay, B. (2013). Success semantics: The sequel. *Philosophical Studies*, *165*, 151–165.

O'Brien, G., & Opie, J. (2004). Notes toward a structuralist theory of mental representation. In H. Clapin, P. Staines, & P. Slezak (Eds.), *Representation in mind: New approaches to mental representation* (pp. 1–20). Oxford: Elsevier.

Papineau, D. (1987). *Reality and representation*. Oxford: Blackwell.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Pouget, A., Beck, J. M., Ji Ma, W., & Latham, P. E. (2013). Probabilistic brains: Knowns and unknowns. *Nature Neuroscience*, *16*, 1170–1178.

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

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*, 79–87.

Rescorla, M. (2015). Bayesian perceptual psychology. In: M. Matthen (Ed.), *The oxford handbook of the philosophy of perception*. Oxford: Oxford University Press.

Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synaesthesia. *Cognitive Neuroscience*, *5*, 97–118.

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

Shea, N. (2013a). Millikan's isomorphism requirement. In D. Ryder, J. Kingsbury, & K. Williford (Eds.), *Millikan and her critics* (pp. 63–86). Oxford: Wiley.

Shea, N. (2013b). Perception versus action: The computations may be the same but the direction of fit differs. *Behavioral and Brain Sciences*, *36*, 48–49.

Shea, N. (2014). Exploitable isomorphism and structural representation. *Proceedings of the Aristotelian Society*, *64*, 1–19.

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese*, *87*, 449–508.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure and abstraction. *Science*, *331*, 1279–1285.

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

Vilares, I., Howard, J. D., Fernandes, H. L., Gottfried, J. A., & Kording, K. P. (2012). Differential representations of prior and likelihood uncertainty in the human brain. *Current Biology*, *22*, 1641–1648.