Review of Philosophy and Psychology
https://doi.org/10.1007/s13164-022-00648-8

A Hybrid Account of Concepts Within the Predictive Processing Paradigm

Christian Michel[1]

Accepted: 17 June 2022

Abstract

We seem to learn and use concepts in a variety of heterogenous "formats", including exemplars, prototypes, and theories. Different strategies have been proposed to account for this diversity. *Hybridists* consider instances in different formats to be instances of a single concept. *Pluralists* think that each instance in a different format is a different concept. *Eliminativists* deny that the different instances in different formats pertain to a scientifically fruitful kind and recommend eliminating the notion of a "concept" entirely. In recent years, hybridism has received the most attention and support. However, we are still lacking a cognitive-computational model for concept representation and processing that would underpin hybridism. The aim of this paper is to advance the understanding of concepts by grounding hybridism in a neuroscientific model within the Predictive Processing framework. In the suggested view, the different formats are not distinct parts of a concept but arise from different ways of processing a functionally unified representational structure.

Keywords Concept · Concept eliminativism · Concept pluralism · Concept hybridism · Predictive Processing · Coactivation package account of concepts

✉ Christian Michel
chris.michel08@gmail.com

[1] Department of Philosophy, University of Edinburgh, Edinburgh, Scotland

Published online: 21 July 2022

1 Introduction

We seem to learn and process concepts[1] in different and heterogenous "formats"[2], like exemplars (e.g., Medin and Schaffer 1978; Nosofsky 1986), prototypes (e.g., Posner and Keele 1968; Rosch 1978; Hampton 2006) and theories (e.g., Keil 1989; Murphy and Medin 1985; Gopnik and Wellman 2012). Exemplar theory holds that concepts are represented as a set of *exemplars* stored under a category label. *Prototypes* are abstracted summary representations, for instance, in the form of a list of features with typicality ratings. And theory-theory describes concepts as embedded in *theory-like* structures or as little *theories* themselves. Other formats are sometimes hypothesized: for instance, definitions (a set of necessary and sufficient characteristics), scripts (procedural knowledge) or ideals (a description of an ideal member of a category). However, exemplars, prototypes and theories are the formats that are generally accepted; for this reason, here I will focus on those three.

Those formats were posited to account for a large range of empirical, mostly behavioral, data related to conceptual development and conceptual tasks (some of which I will discuss later). But none of the aforementioned accounts turns out to be able to accommodate the wealth of empirical data (e.g., Kruschke 2005:188, 190; Machery 2009). Therefore, format variety is now generally recognized as an unavoidable conclusion (e.g., Bloch-Mullins 2018; Hampton 2015; Voorspoels et al. 2011) and has been discussed in depth by Machery (2009).

This heterogeneity of formats sparked many early hybrid proposals, most of them combining two formats (e.g., Osherson and Smith 1981; Nosofsky et al. 1994; Erickson and Kruschke 1998; Anderson and Betz 2001). Given the limited scope and other defects of those initial hybrids, Machery (2009) concluded that each format corresponds to a different fundamental type, and we should dispose of the notion of a concept because the formats have nothing scientifically interesting in common.

Notwithstanding this, many researchers find eliminativism implausible and have continued to propose hybrid solutions in defence of the notion of a concept (e.g., Bloch-Mullins 2018; Keil 2010; Margolis and Laurence 1999, 2010; Rice 2016; Vicente and Martínez Manrique 2016), searched for unity behind the diversity of concept formats (e.g., Danks 2014) or endorsed conceptual pluralism (e.g., Piccinini and Scott 2006; Weiskopf 2009).

Arguably, hybridism is the approach that has received most attention and support in recent years. Therefore, here I will leave pluralism and eliminativism aside and focus only on hybrid accounts. My overall goal is not to defend hybrid approaches. Rather I want to provide a novel way to spell out a hybrid account in the spirit of Vicente

[1] I take concepts to be certain *bodies of information* (see Machery 2009) that are used in many higher cognitive tasks, i.e., abilities like categorization, inductive and deductive reasoning, planning or analogy making. The focus here is on the *psychological* notion of concepts (see Machery 2009, 2020), which is concerned with their *cognitive-computational* significance.

[2] I use the term "format" as a placeholder for whatever protypes, exemplars and theories turn out to be (representational structures, types of knowledge, ways of processing, etc.). Thanks to an anonymous reviewer for suggesting this way of using the term. Also note that "format" is sometimes used in connection with concepts to distinguish amodal and modality-specific representations. This is not the way I use the term here.

& Martínez Manrique's "coactivation package" account (2016). Vicente & Martínez Manrique have forcefully argued that hybrids that do not consider "functional integration" of the formats are hopelessly flawed. While I endorse this view, I nevertheless argue that their approach deserves further development and improvements.

I do not develop a full theory of concepts here. Rather, I focus on the aspect of how a concept needs to be structured as a representational device so that it can serve the roles that the different formats (exemplars, prototypes, and theories) are supposed to play in conceptual cognition. A full theory of concepts would need to address a host of additional desiderata, for instance, how concepts compose to more complex concepts, how they can be shared among members of a language community, etc. (see, e.g., Prinz 2002).

The rest of the paper is structured as follows. In Sect. 2, I discuss hybrid accounts and examine in some more detail Vicente & Martínez Manrique's coactivation package hybrid proposal. I identify two aspects that need further development. In Sect. 3, I introduce a model of concepts that is emerging from neuroscience. In Sect. 4, I introduce Predictive Processing (PP), a cognitive computational framework, and show how the concept model from Sect. 3 can be embedded in it. In Sect. 5, I suggest how the different formats of concepts might arise and how this approach improves the coactivation package account.

2 Hybrid accounts of concepts

I focus on Vicente and Martínez Manrique (2016) (V&MM) which is one of the most recent hybrids[3]. Their account, which I call a "functional hybrid", is a reaction to previously dominating "mereological hybrids". To better appreciate the strengths and weaknesses of V&MM's account, and motivate needed improvements, let me set the stage by briefly discussing mereological hybrids.

2.1 Mereological hybrids

Mereological hybrids treat instances of concepts in different formats as numerically distinct entities that are combined to create a hybrid entity. For most such hybrids, their proponents do not emphasize and provide principles for a deeper functional integration of the parts. This is not to say that mereological hybrids do not provide some integrating principle, of course, but the characterization of how and why the components are integrated is rather minimal and "thin." That, however, makes them vulnerable to various anti-hybrid arguments put forward by eliminativists and pluralists (see, e.g., Vicente and Martínez Manrique 2016, for a discussion). In a nutshell, mereological hybrids have difficulty explaining what keeps the components together, beyond some minimal description, and hence what justifies calling the cluster of formats a concept. Furthermore, it is unclear what explanatory advantage hybridism would have over pluralism and eliminativism. Secondly, mereological hybrids can-

[3] Another account that could be considered a "*functional* hybrid", in the sense defined here, is Bloch-Mullins (2018), which I will briefly discuss in Sect. 5.3.

not say much about what formats are possible, how they hang together and interact, and how they are acquired. They do not seek to reveal an underlying principle from which different formats might naturally arise. Hence, they have an ad-hoc air and lack deeper unity.

As an example, in Margolis & Laurence's (2010) account the different formats are "bound to the same mental symbol". However, no constraints are provided for what formats can be bound to a symbol. Also, nothing is said about how exactly the formats are represented and processed, in particular how different formats are selected on some use occasion. Rice's "pluralist hybrid" (2016) is a further instance of a mereological hybrid. In his proposal, we store information in different formats in long term memory. Information chunks in different formats are retrieved and combined dynamically to create a concept, which is then processed, depending on the task, context, and category. Each combination of different formats corresponds to a different concept. This proposal has the advantage that it does justice to the highly dynamic and flexible processes in concept retrieval. But Rice does not provide constraints for what kind of formats are possible. He also does not explain how those formats are represented and how the selection and assembly mechanisms work.

2.2 A functional hybrid account

I now discuss how V&MM respond to the problems that afflict the mereological hybrid accounts. I argue that while their response focuses on, and advances in terms of a solution to the first problem, they still face issues, including the second problem of mereological hybrids just discussed.

V&MM suggest that *functional integration* is what holds the different formats of a concept together. Contrary to the above-mentioned mereological hybrids, V&MM put the issue of the functional integration into the spotlight. For this reason, I suggest calling their approach a "functional hybrid." Their proposal is then that the unity of a hybrid rests on the "functional stable coactivation" of the formats:

> In a nutshell, the idea is that different structures can be regarded as constituting a common representation when they are activated concurrently, in a way that is functionally significant for the task at hand, and in patterns that remain substantially stable along different tasks related to the same category. (Vicente & Martínez Manrique, 2016:61)

A concept is, roughly, a "coactivation package" that makes information of different formats available. Different formats are different parts of the concept that are context-sensitively selected:

> Depending on the task at hand, and on background factors, one part or another of this complex structure receives more activation and plays the leading functional role. Taken separately, prototypes, theories, and so on may be not concepts, but they are *components of concepts.* (Vicente & Martínez Manrique, 2016:72, emphasis added)

Note that the authors still speak of formats as "components of concepts". But they use "component" in a rather loose sense, not necessarily implying that formats are strictly "separate modules" (p.73).

I agree with the idea that formats should be integrated in such a way that for a given use of a concept the different formats should simultaneously play some functional role. Only some form of functional interdependence guarantees integration. And without integration it is difficult to see why we need hybrids rather than formats as standalone entities, as pluralists and eliminativists claim. Functional integration makes the hybrid resistant to the above-mentioned anti-hybrid arguments, moreover, it undermines eliminativism, because a functionally integrated unit certainly is a scientifically interesting kind that gives rise to generalizations.

However, I see two issues with V&MM's account.

First, what exactly is "functional significance"? V&MM have not spelled out in detail what this notion amounts to. They only provide a minimal characterization:

> The idea behind the functionality condition is that only representational components that *make a positive contribution to select the appropriate tokening of the concept* count as part of such a concept. (p.69, emphasis added)

According to V&MM, the concept components are "functional" in so far as they make a "positive contribution" to the selection of the "appropriate tokening of the concept". I assume here that V&MM mean that "appropriate tokening" involves two elements. Firstly, the "correct" concept should be selected (e.g., DOG instead of HORSE) and, secondly, it should be tokened in an appropriate format (each concept can be tokened in different ways by selecting different "representational components", which I understand correspond to different formats). The interesting question then is: what does this contribution consist of exactly? An answer to this question crucially requires an account of how the context-sensitive selection of formats works, which is not provided by V&MM.

A second issue with the coactivation package account is that it provides no constraints for possible formats. Should we include, for instance, ideals, scripts, and definitions in the coactivation package? The account is simply silent on this question. Formats are given and then merely added to the coactivation package as a range of possible formats. While V&MM strongly emphasize functional integration, without further details about what exactly this consists in and without further constraints on admissible formats, their account risks remaining a programmatic desideratum about functional integration.

I suggest that we can further develop and improve V&MM´s account by adding a level of description from below, i.e., by being more specific about aspects of neural-level implementation. Rather than starting with a set of independently given formats, we should start from a general neurocognitive architecture that is motivated independently of the question of format variety. From this we can then derive the formats.

As such a general neurocognitive framework, I will use Predictive Processing (PP). But before describing it in Sect. 4, I will first provide a sketch of a current neuroscientific picture of how concepts might be represented in the brain.

3 A neuroscientific model of concepts

The hybrid account I propose builds on a model of the neural realization of conceptual representations that, so I suggest, crystalizes out of an increasing body of current empirical and theoretical neuroscience. This model can be articulated in the form of three core claims.

C1. Conceptual representations are realized as *extended networks of nodes*: A conceptual representation is neurally realized as the activation of a set of neuron assemblies (nodes) in the form of a distributed network that can cover different brain areas, from higher cortical areas down to lower-level sensorimotor ones.

C2. Concepts are *hierarchically organized networks*: Different subassemblies (nodes) of the network structure of a concept represent information with different degrees of abstraction/schematicity. The network forms a hierarchy of nodes with an abstraction gradient. Very roughly, higher layers of nodes are sensitive to lower-level node patterns, or in other words, they *compress* lower-level information. The lowest level in the hierarchy corresponds to the sensory periphery, where representations are maximally *modality specific*. As we go higher in the hierarchy, information represented by the nodes gets not only increasingly abstracted/compressed, but also *convolved*, i.e., different modalities (visual, acoustic, proprioceptive, affective, etc.) get mixed (see also Eliasmith 2013).

C3. *Context-sensitive and flexible* conceptual processing: On different occasions different parts of the network of a concept are activated in a task- and context-sensitive manner. The tokening of the same concept on different occasions can reach into lower levels of the hierarchy to different degrees.

C1 and C3 closely follow the view of the neural realization of concepts suggested by Kiefer and Pulvermüller (2012). They characterize concepts as "flexible, distributed representations comprised of modality-specific conceptual features". Furthermore, with regard to C2, it is well established that the brain is hierarchically organized; neural layers and areas correspond to different levels of abstraction/compression (e.g., Raut et al. 2020, Hilgetag and Goulas 2020). This suggests that the extended network structure reaching from higher cortical levels down to sensorimotor areas plausibly has an abstraction/compression gradient.

Kuhnke et al. (2021) have put forward a model and empirical evidence that characterizes the hierarchical structure in more detail by mapping the different hierarchy levels on specific brain regions. Lower-level monomodal representations are compressed in layers in so-called unimodal convergence zones. Those feed into layers in multimodal convergence zones. The highest level is an amodal[4] layer that compresses multimodal input. We have here a double gradient in the hierarchy. On the one hand, the higher the level, the more abstract and compressed the information is. Secondly, in multimodal convergence zones we have a mixing (or convolution) of different modalities. That is, neuron assemblies are sensitive to patterns that involve

[4] The authors call the highest level in the hierarchy "amodal". However, it seems also appropriate to call it "multimodal", given that in that layer we abstract across a maximally broad range of modalities, so it is just one more step in the abstraction/convolution hierarchy, not a qualitatively different step (see also Michel 2020b).

various modalities. The different layers can be identified with different brain areas (e.g., being the "amodal" layer the ATL). Kuhnke et al. (2021) also show that the connectivity between the layers is strongly task-dependent (claim C3).

C1, C2 and C3 are closely interrelated and empirical evidence for them is increasing. Modality-specific (action, visual, gustatory, olfactory, sound, but also interoceptive) representations often activate complex extended neural networks including modality-specific lower-level brain areas (e.g., Hoenig et al. 2008; see also the overview by Harpaintner et al. 2018). What is debated however, is whether a concept includes sensorimotor areas *each time* it is tokened, and whether *abstract concepts* like democracy or freedom also include sensorimotor information.

It is safe to say that lower-level sensorimotor areas are not necessarily activated on each occasion even for concrete concepts (Barsalou 2016; Kemmerer 2015; Pecher 2018). Van Dam, van Dijk, Bekkering and Rueschemeyer (2012) argue for the flexibility and context-dependency of the activation of lower-level modality-specific areas in the case of lexical concepts. Yee and Thompson-Schill (2016) conclude that concepts are highly fluid and their activations depend on the context, including the individual short and long-term experience.

With regard to abstract concepts, studies show that their activation can also include lower-level sensorimotor areas (e.g., Harpaintner et al. 2020), including interoceptive and areas processing emotions. Harpaintner et al. (2018) highlight the "importance of linguistic, social, introspective and affective experiential information for the representation of abstract concepts." Such modality specific features can be context and task-dependently activated (e.g., Harpaintner 2020). Furthermore, various researchers suggest that abstract concepts are grounded in emotions (e.g., Vigliocco et al. 2014, Lenci et al. 2018), supporting the idea that their neural realizations also potentially extend into sensorimotor and affective[5] areas. All of this is evidence that all concepts might have the same fundamental structure. Also, it is evidence for the claim that concepts are sensorimotor grounded in the sense that they are hierarchical networks of nodes that bottom out at the sensorimotor periphery.

It is important to stress that the neuroscientific model of concepts I have articulated here mainly covers the *structure* of the realization of concepts (C1 and C2), but little research is available about the specific dynamics of the context sensitive activation patterns postulated by C3. Specifically, an account of how the different formats of concepts arise is lacking. In other words, from the available neuroscientific work we cannot yet derive a full neuro-mechanistic account of dynamic concept processing and the format heterogeneity. This is where the Predictive Processing framework comes in.

My strategy going forward is to embed the flexible, layered network model of concepts in the Predictive Processing (PP) framework which I will introduce in the next section. I argue that PP can take on board the three core principles of the model and, more importantly, it can bring the wealth of individual findings under a single comprehensive neuro-mechanistic framework. What PP can then bring uniquely to the table is a model of how concepts are *processed*. This will be central for my proposal

[5] Sensory areas are meant to include both exteroceptive and interoceptive modalities.

that different formats arise from *different ways of processing* the network structure that realizes a concept.

4 Concepts within the Predictive Processing framework

In this section I briefly introduce the Predictive Processing (PP) framework and suggest how the model of the neural realization of concepts just described could be embedded in it.[6]

4.1 The Predictive Processing paradigm

Predictive Processing (or coding) (see Clark 2013, 2016; Hohwy 2013; Friston, 2010; Sprevak, 2021) provides a neuroscientific *framework* or *paradigm* for how the brain works from a cognitive-computational perspective. PP is an ambitious framework as it aims at providing a general and unified view on cognitive agency, i.e., an account of perception, action and cognition. It should be stressed that PP is far from being a mature and worked out theory (Sprevak 2021a; Walsh et al. 2020). However, it is a very popular framework in cognitive science. In recent years, its scope of applications has been extended and is now ranging from low-level sensorimotor phenomena to several psychological phenomena and even consciousness (Hohwy 2020).

As a *paradigm*, PP provides guidance and constrains for the development of more specific theories of cognitive phenomena; PP can be seen as a research program based on some programmatic commitments that are generally but not unanimously accepted by the PP community. In the following part I try to synthesize what I consider to be the core commitments that are most relevant for the purpose of this paper. Most if not all commitments taken in isolation are neither original nor unique to PP (see Sprevak 2021a) and it is rather the combination and integration of the commitments that characterizes PP.

4.1.1 Prediction error minimization of sensory input

In very general terms, PP pictures the brain as an anticipation and expectation organ that constantly fine-tunes a mental model to continually predict its sensory input.

For instance, perception is not passive bottom-up feature aggregation and pattern recognition, as traditionally conceived (e.g., Marr 1982, Hubel and Wiesel 1959). Rather, the brain constantly generates hypotheses of its sensorimotor states (including all extero- and interoceptive modalities) and corrects the model in the case of errors, so next time it does a better prediction job. In a way, the brain constantly hallucinates in a manner that happens (normally) to match reality.

[6] Let me stress that I don't aim here at defending the PP framework, therefore I will not put forward arguments or evidence for it. For that I refer to the mentioned literature.

4.1.2 The mental model: generative, hierarchical, and probabilistic

Predictions are being generated by a mental model that is generative, hierarchical, and probabilistic. The attribute *generative* captures the already mentioned idea that the model serves to generate hypotheses constantly and proactively about sensorimotor states.

The model is *hierarchical* because the predictions are being done through representations on many different levels of abstraction/compression (e.g., Clark 2013). In other words, representations, and hence knowledge, are structured in a hierarchy with an abstraction gradient. Higher levels contain representations that are responsive to larger "receptive fields", i.e., they capture more abstract and coarse-grained patterns represented on lower levels. For instance, while on a very low-level pixels in the retina are represented (which change heavily), higher levels contain representations[7] corresponding to concepts like apple, which abstract over many instances of specific apples (and hence are more stable). In the downward flow of information, the predictions of higher-level layers play the role of priors for the lower-level predictions and, in this way, constrain the predictions on lower levels. Predictions are being carried out all the time and on all levels of the model at the same time.

The model is *probabilistic* because it represents probability distributions over (sub-personal) "hypotheses" about the causes of sensory input. Furthermore, prediction error minimization approximates Bayesian inference as its primary computational mechanism (e.g., Clark 2013:188–189; Hohwy 2013:15–39).

4.1.3 Precision weighting mechanism

The PP system contains a so-called "precision-weighting mechanism" of prediction errors (Clark 2016:53–83). Such a mechanism is necessary as the brain must predict the reliability of its sensory input (or more generally the inputs from lower levels in the hierarchy) to distinguish noise and useful signals. In this way, useless modifications of the model due to noisy signals can be avoided. Weights are assigned to the error signals, which allows the system to control the influence of top-down predictions versus bottom-up driven updates of the model. This modulatory mechanism is implemented as part of the overall PP prediction model as (second order) "knowledge" about the reliability and relevance of features in each context (see Michel 2020a).

4.1.4 Neural architecture

PP also makes some general claims about neural implementation. The smallest unit in the model is a combination of an "error unit" and a "representation unit" which I will call a "prediction unit" or simply a "node". Prediction units or nodes are realized as small neural assemblies or "canonical circuits" (see Kanai et al. 2015, also Bastos et al. 2012, Keller and Mrsic-Flogel 2018, Weilnhammer et al. 2018). The error unit

[7] We will later see that it would be more accurate to say here that higher levels contain the *root nodes* of the representational structure corresponding to concepts.

is connected to prediction units on higher levels and the representation unit is connected downwards. Furthermore, there are modulatory inputs into the error units that allow the above-mentioned precision weighting mechanism to tune the error signal.

This brief sketch of the PP paradigm which emphasizes the elements that will play a role in the rest of the paper, should suffice.[8] In the next section I show how the neural model of concepts from Sect. 3 can be embedded in the PP framework.

4.2 PP and concepts

My proposal for how concepts manifest themselves in different formats relies on Michel (2020a, b) who suggests that concepts are implemented in PP by the prediction units just described. Specifically, a given concept is instantiated by a prediction unit, taken as the root node of an extended tree of other prediction units.

The idea then is that the activation of a concept's root node makes available a body of information, namely the subnetwork depending on that root-node. This subnetwork can be seen to correspond to Vicente & Martínez Manrique's "coactivation package". When a concept unit is activated, it makes available a subnetwork that can cover various brain regions, potentially including higher cortical down to primary sensory or motor areas. Critically, which other sub-nodes apart from the root-node itself, are selected is regulated by a context-sensitive modulation mechanism (see Michel 2020a). The basic idea is that higher order knowledge about the reliability and relevance of the different nodes is also encoded in the world model. This higher order knowledge then regulates how the prediction error signals are modulated (i.e., more or less suppressed). Such a mechanism is equivalent to a mechanism that can switch on and off certain parts or nodes of the network depending on the context.

There are concept root-nodes that correspond to patterns on all levels of complexity and spatial and temporal scales. There are, hence, concept root-nodes that range from simple sensory-based expectations, like RED, passing through intermediate-level ones like FACE, to abstract concepts like DEMOCRACY, up to complex situation representations that we grasp in some gestalt-fashion. Such concept root-nodes do not necessarily correspond to lexicalized concepts but also include a host of sub-conscious ineffable ("sub-symbolic") representations that are used as prediction vehicles.

This view of concepts within the PP framework can be put in correspondence with the neural account of concepts as dynamic networks from Sect. 3 in the following way:

C1: The extended network of a given concept corresponds to the sub-network in the PP model that consists of the concept root node and all of its child nodes. (Note that each child node is itself a concept root node).

C2: The sub-network corresponding to a concept is organized hierarchically and has an abstraction gradient in the PP model, exactly like in the neuroscientific model.

[8] My brief exposition of PP is far from complete, and I have omitted many features, e.g., active inference, efficient coding, etc. Virtually every paper related to Predictive Processing contains introductions to the framework. I can recommend, e.g., Wiese (2017); Williams (2018); Sprevak (2021a,b), for a more detailed overview.

Regarding C3, we said that neuroscientific evidence suggests that the concept networks are flexibly and context-dependently activated. According to the PP model the depth with which a concept's tree is activated is flexible, namely task and context-sensitive, driven by the error signal weighting mechanism. Lower-level features can be suppressed by the error weighting mechanism when they are estimated to be unreliable or irrelevant. Activation of a concept can be "shallow" (e.g., a "schematic apple" in which no specific colour is co-activated), in which case only higher-level nodes are activated. Or activations can be "deep", which involves, e.g., a more vivid (modality-specific) mental representation due to the co-activation of nodes that are located lower in the hierarchy (a mental picture of an apple, with a specific colour, form, size, etc.).

The existence and flexibility of concepts can be motivated within the PP framework in a principled way (see Michel 2020a). Concepts are necessary vehicles for prediction making; it is in virtue of prediction units that predictions are made. An efficient prediction economy requires making predictions with an adequate level of detail. When you want to cross a street successfully, your brain's predictions cannot and need not happen on the situation's pixel-level of precision. Rather the predictions need to be more schematic and have a coarser grain. There are two ways to regulate prediction detail. The first is by using prediction units at higher levels in the hierarchy. The higher the nodes, the more schematic and compressed (hence less detailed) their content. The second is by co-activating a varying number of other nodes; those represent more detailed and concrete features of that conceptual representation.

In conclusion, by embedding the neuroscientific model of concepts from Sect. 3 in the PP framework, we get a more comprehensive model of concept representation *and processing*. As we have seen, PP can provide an implementational-level proposal for the network structure (a network of PP prediction units with an abstraction gradient). But what PP can crucially contribute is the processing aspect, which is still underdeveloped in the literature. For instance, PP supplies a self-organizing driving force operative in the node network (prediction error minimization), as well as a mechanism for feature selection (based on the precision weighting mechanism). Furthermore, PP motivates the existence of concepts as prediction vehicles, and the need for the right level of granularity, which in turn motivates the existence of the feature selection mechanism.

5 The manifestation of different concept formats

With a cognitive-computational account of the structure of conceptual representations in place, I will now show that the different formats correspond to how the network of a concept is being context-sensitively processed. The different formats mirror not numerically distinct representational entities, but the processing depth and width of the concept's (and surrounding) network structure. More precisely, exemplar effects correspond to relatively deep vertical downward processing (i.e., towards less abstract nodes), prototype effects to relatively shallower vertical downward processing, and theory effects to additional vertical upwards and horizontal processing (i.e., towards parent and neighbor nodes).

5.1 Exemplars and prototypes

In this subsection I argue that a concept can manifest itself in "exemplar mode" and "prototype mode" when the node tree associated with the concept is processed from more to less abstract nodes (vertically downwards processing). Processing only higher-level nodes corresponds to prototypes. Processing in addition lower-level nodes corresponds to exemplars. I will first unpack this proposal by explaining how exactly to understand exemplars and prototypes and how they are realized in the PP model. Then I will provide some examples of how we can account for the exemplar and prototype effects that motivated those formats in the first place.

5.1.1 What exactly are exemplars and prototypes?

In the standard story of exemplar theory, which aims to address exemplar effects, my concept DOG consists of the memorized collection of representations of specific dogs. They are modality-wise specific as they correspond to instances of dogs. Categorizing some animal as a dog implies using dog exemplar(s) and calculating similarities. Note that the exemplars might have very different levels of specificity, i.e., levels of modality-specific detail or vividity. Sometimes we remember object-exemplars only vaguely with little detail, and sometimes very concretely with a lot of detail.

In the standard story of prototype theory, which aims to address prototype effects, my concept DOG consists of some representation of a typical dog. The representation is more abstract compared to an exemplar. Categorizing some animal as a dog under prototype theory, implies using the dog prototype and calculating the similarity.

Note that the processing, for instance in categorization tasks, of both exemplars and prototypes rely essentially on *similarity* calculations, primarily over relatively superficial features.

Some researchers think that exemplars and prototypes are the ends of a continuum rather than two distinct kinds (e.g., Vanpaemel et al. 2005, or Verbeemen et al. 2007). Authors like Barsalou (1990) and Hampton (2003) think that prototypes and exemplars differ only to the extent to which exemplar information is retained or abstracted away. Smith and Medin (1999:209) characterize exemplars in terms of *a relative lack of abstraction*. Exemplars can be maximally specific object-particulars but are not necessarily; they can also be subsets. For instance, PLANET is a subset of HEAVENLY BODY, and hence an exemplar for it.

Following those authors, I assume that there is no fundamental difference between exemplars and prototypes in terms of the deeper, underlying representational structure in the first place. In both cases, the general structure consists of a set of pairs of features and values. Those features might have different degrees of specificity/schematicity.

5.1.2 Prototypes and exemplars in the PP model

The posited structure of a concept as a hierarchical node tree allows us to account for the exemplar and prototype formats. Concept processing in exemplar mode can

be cashed out as the processing of the concept's node tree with attention towards *relatively* more specific information (without necessarily being *maximally* modally specific), while processing in prototype mode can be cashed out as more shallow processing, i.e., involving nodes with relatively less specific information. In both cases we have more or less deep "vertical downwards" processing of more superficial features. Those features are included in the node tree that origins in the concept's root node.

In PP terms, processing a concept in exemplar mode is processing towards lower-level (i.e., modally more specific) nodes. The tokening of the concept DOG in exemplar mode reaches from the conceptual root node [DOG] down to at least a subordinate node and potentially (but not necessarily) further to lower-level nodes down to the sensorimotor periphery. To conceive of a specific dog, e.g., Hasso, *as a dog*, implies the activation of the abstract [DOG] node and the subordinated [HASSO] node and other subordinate nodes, potentially down to specific shapes, colours, odours, etc. So, a whole node sub-tree from [DOG] might be activated.

To categorize a specific dog exemplar, say Hasso, a hypothesis needs to be generated that matches as well as possible whatever sensory input I receive. If my dog Fido is very similar to Hasso, a salient hypothesis is of course that Fido actually is Hasso. So, the hypothesis that reproduces a memory of Hasso fits well with the bottom-up Fido input, i.e., it produces a small prediction error in relation to other hypotheses.

Categorization might also happen via a prototype of DOG. If you cannot see Fido well (because he moves quickly and is far away and could be a cat as well) but hear loud barks, given that the feature of barking is strongly cue valid (i.e., the probability that something that barks is a dog is high), there is no need (and it would not be very economic) to recur to more specific exemplar information. The barking can be immediately explained by the hypothesis DOG and Fido categorized as a dog.

It is important to stress that, in the proposed view, what is an exemplar and what is a prototype is *task-dependent*. It might happen that in a task a prototype of some concept is represented with more detail than an exemplar of that concept in another task. Consider the following example:[9]

> 1) Suppose that a Bach scholar is played a piece of music and asked whether it is typical of Bach. To answer this question, the scholar may draw upon a very rich mental representation of the typical features of Bach pieces, which encodes very specific information about sensorimotor details such as certain kinds of instrumentation, cadences, melodies, harmonies, ornaments, rhythms and so on.
> 2) Now suppose that the scholar is asked whether the Brandenburg Concertos are a work by Bach. Plausibly, the scholar could answer this question without drawing on deep, specific, information, close to the sensory periphery.

In task 1), the prototypical representation, say $\text{BACH}_{\text{prototype}}$, used by the scholar to decide whether the piece he is listening to is typical of Bach might perfectly contain

[9] I am grateful to an anonymous reviewer for providing various potential counterexamples, including this one.

very specific features. The important point is that $\text{BACH}_{\text{prototype}}$ is relatively more abstract than the exemplar representation in *this* task, which is the piece of music, say $\text{BACH}_{\text{exemplar}}$, that she has to classify. In task 2) we deal with a completely different process, again with two representations, say, BACH-WORKS and BANDENBURG-CONCERTO. The question is whether the latter is an exemplar of the former. Indeed, to answer this, one only needs to know that the Brandenburg Concertos are works by Bach (the former is an instance of the latter category). What is needed is that BACH-WORKS is a relatively more abstract representation than BANDENBURG-CONCERTO, and that is sufficient for the latter to be an exemplar of the former. According to the PP model, this is the case if, for instance, BANDENBURG-CONCERTO is represented as a child node of BACH-CONCERTOS. Here the exemplar BANDENBURG-CONCERTO from task 2) is much less concrete than $\text{BACH}_{\text{prototype}}$ from task 1); but that does not undermine the proposed account. What matters is the relative abstractness of the relevant representations *within* each task.

Let us turn to the probabilistic element of PP: the nodes making up the PP model represent whatever they represent in terms of probability distributions. Specifically, a node represents a probability distribution over nodes in the next lower level. For instance,[10] Richard II might be represented as an exemplar of MONARCHS-OF-ENGLAND because the probability distribution over monarchs encoded in MONARCHS-OF-ENGLAND has at a given moment a sharp spike at the child node RICHARD II. Being an exemplar does not imply, however, that *all* lower-level nodes have sharp distributions. For instance, my probability distribution over the hair color feature of Richard II must be very spread-out indeed. As already mentioned, often exemplars are quite schematic (as in the Bach example 2). In the case of a prototype representation, the probability distribution is more broadly spread. A typical feature or exemplar is then one with the largest likelihood. For instance, MONARCHS-OF-ENGLAND might encode a probability distribution over features such that a typical monarch is one who has the most likely features, i.e., those features with the highest probabilities.

Note that in the PP view, there is no explicit "calculation" of similarity formulas, which is central to categorization in exemplar and prototype theories (see, e.g., Machery 2009 for examples of formulas). Rather, similarity is implicit in the fundamental mechanism of the PP model, namely, weighted prediction error minimization. In weighted prediction error minimization, the top-down prediction and the bottom-up input at each level are compared, i.e., their "similarity" is determined. This mechanism can model both the more abstract prototype level (by focusing attention on higher level nodes, i.e., dampening lower-level nodes that represent more details) and the exemplar level (i.e., lower-level nodes are more error sensitive).

5.1.3 Prototype and exemplar effects

As emphasized already, a theory of concepts aims at accounting for a large body of behavioral effects observed during conceptual tasks.

[10] Thanks to an anonymous reviewer for the example, which helped me to make the point clearer.

Prototypes have been motivated by "typicality effects" that could not be explained by the previously prevailing definitional theory of concepts, according to which concepts are definitions or necessary and sufficient properties. A typicality effect arises when we judge certain objects to be more typical members of a category than others. For instance, a sparrow - in normal contexts - is judged to be a more typical bird than an ostrich. In the standard story of prototypes theory, the concept BIRD consists of a set of properties and a typicality rating for each property. A sparrow would in normal circumstances be a more typical bird than an ostrich.

Typicality can be accounted for in terms of representations based on probability distributions through conditional probabilities as they are posited by PP. For instance, if we know that something is a bird, we expect to a higher degree (in a neutral context) that some instantiation is a sparrow rather than an ostrich. So, a sparrow is a more typical bird that an ostrich. In PP jargon: when you are asked to mention a typical bird, your generative model is more likely to "sample" [SPARROW] in the next lower level in the node tree below [BIRD] than [OSTRICH]. This is expressed as the following relation between two conditional probabilities p(OSTRICH | BIRD) < p(SPARROW | BIRD) which are encoded in the PP world model.

The PP model can also provide an account of how *exemplar effects* work. Take, for instance, the *old item advantage effect*: memorized exemplars are more easily categorized than new ones that are equally typical (e.g., Smith and Minda 1998, 2000). Those effects could be modelled within the PP framework as follows. For sensory input like previously encountered and memorized exemplars, the prediction error is better minimized by using the exemplar rather than a prototype. In the case of "deep processing" which is characteristic for exemplar processing and where details matter, the most similar memorized bird exemplar just best "predicts" the target bird you see in front of you because it causes the least prediction error. The fact that details matter is cashed out in terms of the higher error sensitivity of lower-level nodes that represent more specific features. The more specific features, however, are only considered in the prediction if the brain assigns a high precision estimate to the prediction errors on the level of those features, i.e., when it considers details to be relevant and reliable. In the above example, where a person hears a dog barking in a foggy environment, details will be suppressed due to the lack of reliability of the sensory input. Therefore, more abstract prototype representations are used. Barking is a property with high cue validity.

So, according to the PP model, depending on the relevance and reliability of the details, exemplar or prototype modes of processing arise. Note that those are not two strictly dichotomic modes, but a gradation along the abstraction gradient exists. As mentioned, concepts within the PP model serve to modulate the granularity of predictions. Taking up again the example from Sect. 4.2., it is not efficient when a street is crossed to predict the exact, maybe pixel-level, details of the event. Rather the event should be processed on a more aggregated level. For instance, we do not need to predict the exact shape and colour of the car approaching when we try to cross the street. It is sufficient to conceptualize the scene in larger grain, e.g., that some fast-moving car is approaching. Exemplar and prototype formats are manifestation of this context dependent granularity modulation (or choice of abstraction level). Also note that what format, or more precisely, what level of abstraction is used in each

task might vary across individuals. For instance, someone who is especially afraid of sports cars when crossing a street might pay more attention to more detailed features. Maybe someone is especially afraid of a specific car (because in the past Uncle Tim's car has almost hit her, for instance) and, therefore, she mobilizes even more detailed exemplar information for prediction making.

5.2 Theories

Now I argue that a concept can manifest itself in "theory mode" when the surrounding node structure in which the concept is embedded is processed (i.e., processing in a vertically upwards and horizontal direction from the concept's root node). I will first unpack this proposal by explaining how exactly to understand the notion of "theory" and how a theory is realized in the PP model. Then I will walk through an example of how we can account for a classical knowledge effect that motivated the theory format in the first place.

5.2.1 What is a "theory" in the theory-theory of concepts?

It is important to point out that theory-theory is far from being a monolithic position. Discrepancies (or indeterminacies) exist along various dimensions; let me mention two and make explicit what notion of theory I will assume.

Firstly, there are two ways in which the relation between concepts and theories has been spelled out (see, e.g., Weiskopf 2011): concepts are *constituents of theories* or concepts are *miniature theories* that store relevant theoretical (i.e., causal, functional, taxonomic, etc.) knowledge. In the first case, theories are bodies of beliefs or propositional structures with concepts as constituents. In a strong version of this view (e.g., Carey 1985) concepts are individuated as the roles they play in those theories. In the second case, concepts are structures that are themselves little theories (e.g., Keil 1989). However, it is not spelled out in detail what this position exactly amount to in terms of its representational structure. For instance, when Keil says

> most concepts are partial theories themselves in that they embody explanations of the relations between their constituents, of their origins, and of their relations to other clusters of features. (1989:281)

the question arises as to what exactly the embodiment of those items looks like. If those items are articulated as beliefs or propositional structures, how is this then different from the concepts-as-constituents view? Even worse, the view seems then to have the incoherent implication that a concept is both a constituent and a theory of which it is a constituent. So, it is crucial to spell out how the knowledge items are represented. The concept-as-constituents view seems not to have this specific problem because there are two things: some theory and a concept that is a constituent of that theory. In turn, this view does not capture the intuition that a concept indeed seems to be some sort of "information package" including a host of theoretical information. In any case, we have here an unresolved problematic aspect of theory-theory in general because, as Weiskopf points out (2011), "the empirical evidence taken to support the

Theory-Theory does not generally discriminate between them, nor have psychologists always been careful to mark these distinctions."

The advantage of the proposed PP account of concepts is, as I will argue later on, that it spells out a specific representational structure that allows to perfectly make sense of the idea that a concept can be seen to be *both*, a miniature theory *and* a constituent of a theory.

A second aspect where theory-theories vary is the demand regarding the *coherence* of the encoded knowledge. Kwong (2006) usefully distinguishes two different notions of theory, a *literal* and a *liberal* one. A literal theory is analogous to a scientific theory, and cognitive and conceptual development is equivalent to scientific theory formation and change. Here aspects of causal relationships, coherence, and systematic structure are stressed. An example of a literal understanding of a theory notion is Gopnik & Wellman's (2012) account. According to the authors, a theory is a coherent structure of abstract representations, analogous to scientific theories (2012:1086).

On the other hand, in the *liberal* understanding of theory, as endorsed, for instance, by Murphy and Medin (1985), the knowledge structure is more flexible. When they say that "…we use theory to mean any of a host of mental 'explanations,' rather than a complete, organized, scientific account" (1985:426), they allow other, informal types of knowledge structures, i.e., formats, in a theory. Such formats are, for example, empirical generalizations (mere correlations of phenomena) or scripts (procedural knowledge, or a chain of events or acts). Liberal theory theorists put less demand on the coherence of a body of knowledge. A representational knowledge system does not need to exhibit formal consistency and rigor, deductive closure, etc., to count as a theory. Such features might be desirable and are most probably normative; however, they are not plausible as a description of how we cognitive-psychologically store knowledge.

I will endorse the liberal view of theories relevant for concepts because the strict view seems psychologically implausible (see also Machery 2009:102). The liberal notion of theory is closely related to the notion of "folk theories." A folk theory, or "intuitive theory" is common sense knowledge about a specific domain, for instance folk biology or folk psychology (e.g., Gerstenberg and Tenenbaum 2017). The building of such folk theories is less systematic and conscious than scientific theory building.

5.2.2 Theories in the PP model

As we have said before, in the proposed PP model, world knowledge is encoded as a huge network of interconnected prediction units (nodes) on many levels of abstraction/complexity. In the upper levels we have prediction units that represent complex situations, contexts, scenes, relations, patterns, patterns of patterns, etc. The lower levels represent for instance concepts of concrete objects or simple features like colour, etc.

The PP framework quite naturally accommodates theory-like structures, as the generative PP model is standardly interpreted as a multilevel *causal model* (e.g., Friston 2010, van Pelt et al. 2016). Nodes that correspond to variables form a proba-

bilistic network. The model is *hierarchical*, i.e., the nodes at one level, roughly, correspond to latent variables that are the causes from which the variable in the next lower level can be derived. However, limiting the relations between the variables to *causal* relations makes the model too narrow (see also Sprevak 2021b). A prediction unit can be more generally interpreted as a prior that constrains the values on lower levels, i.e., nodes and sub-nodes have a more general form of "predictive relation", which can also include part-whole relations or taxonomic relations or object-property relations. The reason is that all of those are "predictive" in the sense that in the same way as causes constrain possible effects, genera constrain possible species, and wholes constrain possible parts.

In theory mode, so I suggest, it is the connectivity of a concept root node with higher-level nodes and nodes on similar levels in the total model hierarchy that is being exploited. In other words, the theory mode of concept processing arises from horizontal and vertical upwards processing *outside* the concept node tree, in addition to vertical downwards processing within the concept node tree below the concept's root node. While exemplar and protype processing remain within the structure of the subordinate nodes of a concept root node, in theory mode, processing expands upwards to more abstract and laterally into neighbouring concepts units.

One might think that theories are represented in terms of high-level, relatively abstract, human-interpretable, lexicalized concepts. For instance, a certain edge form representation in the brain´s visual processing stream is not a concept in the more traditional and common-sense understanding. Perceptual and conceptual representations are normally seen as qualitatively distinct.

However, authors proposing the existence of "folk theories" (e.g., Gerstenberg and Tenenbaum 2017) do not assume representations in symbolic and lexicalized form. A folk theory of physics, which allows for guessing whether certain tower constructions are stable, requires complex "sub-symbolic" sensorimotor representations. Similarly, I have emphasized within the proposed PP view the existence of many ineffable, consciously not accessible, and non-lexicalized nodes on many levels of abstraction (see also Lake et al. 2017 for a discussion of sub-personal "theories" that are not lexicalized). Those sub-symbolic nodes are continuous with the symbolic nodes that correspond to more narrowly understood concepts (e.g., only lexicalized or lexicalizable[11] concepts). All the nodes are "concepts" in virtue of them playing the role of prediction units. They just differ in the degree of abstraction. We could stipulate that only narrowly conceived concepts form theories. But nothing hangs on this rather terminological decision. We can consider theories based on narrow concepts to be "embedded" in the total PP model, which consists of both narrow and inclusively conceived concepts.

[11] A feral child might have the lexicali*zable* concept WOLF, though it is not lexicalized. In contrast, all sorts of ineffable edge-patterns and shapes are used, e.g., in lower levels of the visual pathway there are prediction nodes that are not consciously accessible and lexicalizable in any meaningful way.

5.2.3 Accounting for knowledge effects

The classical knowledge effect I want to focus on here as an example is reported by Rips (1989) in his famous pizza experiment. It provides evidence that sometimes we classify some A to be a B, rather than a C, even if A is more similar to C. Rips asked participants to imagine a circular object of three inches and asked *whether it was more similar* to a quarter or a pizza. The dominant answer was that it was more similar to a coin (because of its small size). Then the participants were asked *whether it is more likely* a pizza or a quarter. The dominant answer was that it was more likely a pizza (because quarters have uniform sizes, while pizza sizes might vary). Here we do not categorize in terms of similarity but rather based on more extended knowledge, e.g., of the manufacturing process of pizzas and quarters from which we can infer their possible variability in size.

Let us now account for the pizza experiment by the PP model. The concept formats involved - prototypes/ exemplars versus theory-like common-sense knowledge - seem to be primed by the task. In the first task, the subjects are explicitly being asked to make a *similarity* judgement while the second task evokes a judgement about the causal chain that brought about each object (pizza versus quarter).

Such causal knowledge is encoded in the PP model as specific experiences but also more abstract generalizations that one might have, which also involve other concepts like PIZZA BAKER, PIZZA OVEN, DOUGH, etc. from experiences with how pizzas are made (see Fig. 1). Hence the concept PIZZA is being processed by carrying out inferences with concept units outside the information package PIZZA itself. A more abstract node in the PP model might be a concept unit representing a complex schema PIZZA-BAKING_SCHEMA which is a sub-domain of common-sense knowledge about baking represented by BAKING-SCHEMA. PIZZA-BAKING_SCHEMA might have sub-nodes that are part of the knowledge about pizza baking, let us say AGENT-FORMS-DOUGH_SCHEMA and HEAT-DOUGH-TO-END-PRODUCT_SCHEMA.[12] AGENT-FORMS-DOUGH_SCHEMA again contains sub-nodes that contain information about how an agent forms the dough, etc. From that knowledge one can infer that it is easy to make, for instance, a pizza that is smaller than usual, simply by applying the same pizza forming process to a reduced quantity of dough. This reduced quantity is possible as the pizza baker is free to choose the quantity she wishes.

Similarly, quarter, might be a node subordinate to a more abstract node corresponding to some frame concept unit, which links quarter in such a way as to encode common-sense knowledge about the role and production of coins. From that knowledge one can infer that it is very unlikely that a coin has the size of the target object. The agents intervening in the coin producing process do not normally have the "freedom" to alter the size of a coin ad hoc.

[12] Here PIZZA-BAKING_SCHEMA could be a concept that encodes a script, i.e., a sequence of actions.

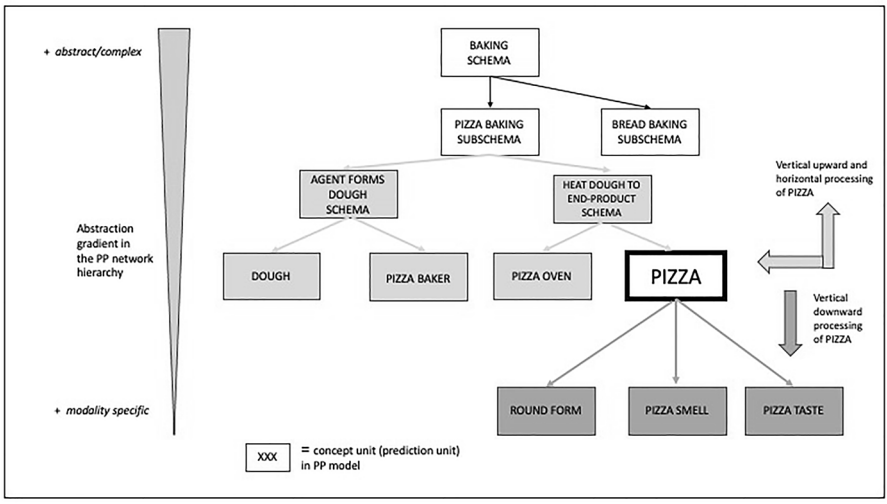

Fig. 1 A schematic toy example of a concept unit network for the concept PIZZA and modes of processing

Taking this way of processing the concept structure, the inference is being made that a pizza can easily have different sizes, while coins do not. Therefore, the target object is more likely to be a pizza[13].

5.2.4 Are concepts then theories or constituents of theories?

With this approach of the theory format in hand, we can now briefly revisit the question discussed in Sect. 5.2.1., namely whether a concept (in its theory format) is a theory or a constituent of a theory. It is easy to see that the dispute now looks merely verbal. A concept can be *both*. A concept, say APPLE, can appear to be a theory when connected nodes are processed that represent theoretically relevant information (i.e., when it is processed in theory mode). But APPLE can also appear to be a "constituent" of some (other) theory, namely when at least the root-node of APPLE is processed as part of the processing in theory mode of some (other) concept, for instance, FRUIT or NUTRITION.

5.3 The functional integration of exemplars/prototypes and theories

One might object that exemplars/prototypes and theories do not seem to have the same status in the concept's information package. There are three properties that prototype and exemplar processing share but that are absent from theory processing. Firstly, prototype and exemplar processing involve nodes of the *sub*-network of the

[13] Given that the PP approach has commitments on the level of neural implementation, at least in principle, there is an avenue for empirical verification/falsification of the model. Admittedly, the current state of the art in brain imaging techniques does not yet provide a sufficient level of temporal and spatial resolution to map out concepts and neural structures in the required way.

concept's root node, while at least some nodes corresponding to theory processing lie outside this sub-network. Secondly, we have also seen that the distinction between exemplars and prototypes is a relative affair, but nothing similar has been said for the theory format. Finally, exemplars and prototypes are closely associated with the notion of similarity, which is not (at least not obviously) the case for theoretical knowledge.

Despite those differences, all three formats should be seen as deeply functionally integrated in the form of a prediction device. To better understand why theoretical information is also integrated with exemplar and prototype information of a given concept, note that - from a neuro-anatomical point of view - the main difference is that processing theoretical information involves nodes on a level higher than (or the same level as) the concept's root-node, while prototype/exemplar information involves nodes at a relatively lower-level. In both cases, however, the concept's root node is involved and connected to those nodes, and the general structure and processing principles are the same in the whole hierarchy. The specific connectivity implements a layered structure of *conditional probabilistic dependencies* among the nodes on different levels. It is this informational dependency dynamics which then integrates the higher and lower-level nodes connected to a given root-node into a functional whole. Let me work this out in further detail.

Remember that a PP model is a generative model with latent variables represented as nodes that "explain" (or "generate", or "sample") features represented by lower-level nodes. While lower-level nodes correspond to concepts that *are "explained"* by some concept in question, higher-level nodes correspond to concepts that *"explain"* that lower-level concept. For instance, while APPLE "explains" RED, FRUIT "explains" APPLE in the sense relevant here. In other words, using the terminology of generative models, RED is a sampled (a "generated") feature from the probability distribution over features represented by APPLE. APPLE, in turn, is sampled with a relatively high probability from FRUIT, which is a probability distribution over fruit types.

Plausibly, the body of knowledge associated with some concept includes both information about what it *is caused/explained by* and what it *is a cause/explanation for*. In this sense, exemplars/prototypes (with more superficial features) *and* theoretical features (representing more abstract causal, taxonomic, mereological, etc. relations) form a functionally integrated information package. The difference is only one of explanatory (or "generative") *direction*.

To bring home my point about the tight functional integration of exemplars/prototypes and theoretical information, it might be useful to refer briefly to Bloch-Mullins' recent work on concepts (e.g., 2018, 2021). There is no space here for a careful discussion of her account and a detailed comparison, but it is worthwhile pointing to some deeper commonalities, which suggest some substantial common ground.

Bloch-Mullins (e.g., 2018: 607) observes, quite correctly in my view, that the problem with the different single-format accounts of concepts is not that they are each on their own unable to cover all of the empirical data from concept research. The problem is that they do not even have sufficient explanatory depth with regards to the restricted scope of the phenomena they were designed to cover. For instance, she argues that the similarity judgements involved in exemplar and prototype applica-

tions cannot be calculated without theoretical (specifically causal) knowledge about how to pick out the relevant dimensions for comparison (pp. 609–614). Theoretical knowledge, in turn, can't be applied in categorization without using similarity judgements to determine the relevant range of values that determine the category of a variable figuring in a causal relation (pp. 615–621). Normally, the values of the variables by which those causal relations (used for categorization) are described are not *identical*, but only *sufficiently similar* to underwrite classification. A second way in which causal knowledge is relevant in categorization is that the dimensions selected for similarity judgements might also include causal relations (Bloch-Mullins 2018, pp. 622 and 624; see also Bloch-Mullins 2021:61–62; Hampton 2006:85–86). I suggest a third way in which similarity intrudes categorization based on causal knowledge: grasping and applying theoretical knowledge is itself recognizing analogies/similarities to abstract (e.g., causal) patterns, i.e., causal knowledge is stored as patterns that demand similarity matching.

I am very sympathetic with Bloch-Mullins' view. In the PP model, the similarity of A and B can be fleshed out as A and B being an instance of (being "sampled from") some concept node. If there is some C that "generates" A and B, then A and B are similar with respect to the features that C encodes. But this idea is transferable to theoretical (i.e., causal, taxonomic, mereological, etc.) features. To see this, let us take one of the examples that motivated the theory format of concepts, namely deep "essences" of living creatures (e.g., Medin & Ortony, 1989; Gelman, 2004). For example, assume that HORSE-A and HORSE-B are representations of horse exemplars in virtue of being sampled by some HORSE-ESSENCE which represents the horse essence that "generates" horses. Our folk-biology might be represented minimally as the knowledge that animals have hidden essences that are responsible for (i.e., cause) the existence of certain animal types. In the PP model, this knowledge is captured by some abstract high-level prediction unit that encodes the very general concept of ANIMAL-ESSENCE as part of some animal folk-theory. There are lower-level child nodes of [ANIMAL-ESSENCE] that correspond to more specific essences like HORSE-ESSENCE, DOG-ESSENCE, etc. Those in turn sample (or "generate") concrete exemplars of the corresponding species, e.g., FIDO (the dog).

The advantage of the PP approach is, as previously pointed out, that similarity calculations are not based on algorithms over an explicit list of features but are the implicit result of holistic prediction error minimization. What is then instantiated as being similar to what depends heavily on the "context" which includes background knowledge, goals, foils under consideration, etc., all of which are represented by other prediction units in the network. PP captures well this highly context dependent dynamics of similarity calculations. Similarity judgements emerge holistically from all of the relevant available information in the PP model.

5.4 In which sense does the PP model refine the coactivation hybrid account?

Let us get back to the end of Sect. 2 where I pointed out two possible improvements to the coactivation account: spelling out more concretely what functional integration amounts to and providing constraints for "admissible" formats. Let us revisit each of them in the light of the proposal just developed.

First, there is a more specific notion of *functional integration* that emerges from the PP model. The whole coactivation package of a concept serves as a context-sensitive prediction device for the category represented by the concept. A coactivation package, we have seen, consists of a root-node and the depending sub-network of lower-level nodes. The root-node is the result of abstraction and convolution of lower-level nodes, therefore in a sense it is closely connected to (i.e., it "contains" information of) all sub-nodes. Those subordinate nodes correspond to exemplar and prototypical information. Furthermore, as this package is integrated into the whole overall model, it has external connections to other lateral and higher-level nodes. Those nodes correspond to more theoretical and abstract knowledge associated with the concept, namely causal, taxonomic, mereological, etc., information that "explains" the concept.

Processing in the PP model is holistic, so all of the nodes are interlocked and have an influence on the overall state of the information package associated with the concept, i.e., on which other nodes are selected, and which are not.

With the PP model, an account of the context sensitive modulation of the subparts of a coactivation package comes for free because it is a core feature of the general PP framework. It can be put to work to select the processing depth and direction that determine the appearance of the concept formats.

Secondly, the PP model provides constraints for possible formats, namely those imposed by the PP architecture. One needs to be able to derive the format from the representational resources provided by PP. We have seen that we can derive the three generally accepted, classical formats: exemplars, prototypes, and theories. An interesting next step - that needs to be carried out elsewhere, however - would be to explore whether other candidate formats like definitions, scripts or ideals could be derived from, or are consistent with, the proposed PP model.

6 Conclusions

This paper has attempted to put forward a cognitive-computational model of hybrid concepts within the Predictive Processing framework. In the view proposed here, formats are - contrary to most other hybrid accounts - not to be understood as components of a concept. Rather, formats correspond to different directions and depths of processing of the same concept structure.

The model aims to further develop and improve Vicente & Martínez Manrique's hybrid account with regard to two aspects. Firstly, it spells out what "functional integration" of the formats more specifically amounts to. Functional integration is necessary for a genuine hybrid account. Formats are functionally integrated in the PP model because they arise as optimal (i.e., prediction error minimizing) ways of processing a unified representational structure. Critical for the functional integration is the context-sensitive selection of subparts of the structure (which then appear as different formats). Such a format selection mechanism comes for free in the PP model. Secondly, the proposed model provides constraints for possible formats because it supplies more detail about how concepts are represented and processed in the mind,

providing more specific computational, algorithmic and implementational level commitments.

Acknowledgements I would like to thank Mark Sprevak and three anonymous reviewers for their very useful comments.

Funding and Competing interests The author has no relevant financial or non-financial interests to disclose.

References

Anderson, J. R., and J. Betz. 2001. A Hybrid Model of Categorization. *Psychonomic Bulletin and Review* 8: 629–647.

Barsalou, L. W. 1990. On the indistinguishability of exemplar memory and abstraction in category representation. *Advances in social cognition* 3: 61–88.

Barsalou, L. W. 2016. On Staying Grounded and Avoiding Quixotic Dead Ends. *Psychonomic Bulletin & Review* 23 (4): 1122–1142.

Bastos, A. M., W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries, and K. J. Friston. 2012. Canonical Microcircuits for Predictive Coding. *Neuron* 76 (4): 695–711.

Bloch-Mullins, C. L. 2018. Bridging the Gap between Similarity and Causality: An Integrated Approach to Concepts. *The British Journal for the Philosophy of Science* 69 (3): 605–632.

Bloch-Mullins, C. L. 2021. Similarity Reimagined (with Implications for a Theory of Concepts). *Theoria* 87 (1): 31–68.

Carey, S. (1985). *Conceptual change in childhood*. MIT press.

Clark, A. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36 (3): 181–204.

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

Danks, D. 2014. *Unifying the mind: Cognitive representations as graphical models*. MIT Press.

Eliasmith, C. 2013. *How to build a brain: A neural architecture for biological cognition*. Oxford University Press.

Erickson, M. A., and J. K. Kruschke. 1998. 'Rules and Exemplars in Category Learning'. *Journal of Experimental Psychology: General* 127: 107–140.

Friston, K. 2010. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Gelman, S.A. (2004). Psychological essentialism in children. *Trends in cognitive sciences* 8.9: 404–409.

Gerstenberg, T., and J. B. Tenenbaum. 2017. Intuitive theories. *Oxford Handbook of Causal Reasoning*, 515–548.

Gopnik, A., and H. M. Wellman. 2012. Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological bulletin* 138 (6): 1085.

Hampton, J. A. 2003. Abstraction and context in concept representation. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* 358 (1435): 1251–1259.

Hampton, J. A. 2006. Concepts as prototypes. In *The Psychology of Learning and Motivation: Advances in Research and Theory*, ed. B. H. Ross, vol. 46, 79–113. Amsterdam: Elsevier.

Hampton, J. A. 2015. Categories, prototypes and exemplars. In *The Routledge Handbook of Semantics*, 141–157. Routledge.

Harpaintner, M., N. M. Trumpp, and M. Kiefer. 2018. The Semantic Content of Abstract Concepts: A Property Listing Study of 296 Abstract Words. *Frontiers in Psychology* 9: 1748.

Harpaintner, M. 2020. Neurocognitive architecture of the semantics of abstract concepts. Dissertation University of Ulm.

Harpaintner, M., E.-J. Sim, N. M. Trumpp, M. Ulrich, and M. Kiefer. 2020. The grounding of abstract concepts in the motor and visual system: An fMRI study. *Cortex; A Journal Devoted To The Study Of The Nervous System And Behavior* 124: 1–22.

Hilgetag, C. C., and A. Goulas. 2020. 'Hierarchy' in the organization of brain networks. *Philosophical Transactions of the Royal Society B: Biological Sciences* 375 (1796): 20190319.

Hoenig, K., E.-J. Sim, V. Bochev, B. Herrnberger, and M. Kiefer. 2008. Conceptual flexibility in the human brain: Dynamic recruitment of semantic maps from visual, motor, and motion-related areas. *Journal of Cognitive Neuroscience* 20 (10): 1799–1814.

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

Hohwy, J. 2020. New directions in predictive processing. *Mind & Language* 35 (2): 209–223.

Hubel, D. H., and T. N. Wiesel. 1959. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology* 148 (3): 574–591.

Kanai, R., Y. Komura, S. Shipp, and K. Friston. 2015. Cerebral hierarchies: Predictive processing, precision and the pulvinar. *Philosophical Transactions of the Royal Society B: Biological Sciences* 370 (1668): 20140169–20140169.

Keil, F. C. 1989. Conceptual development and category structure. In: Neisser, U. (Ed.). *Concepts and conceptual development: Ecological and intellectual factors in categorization* (1). CUP Archive.

Keil, F. 2010. Hybrid vigor and conceptual structure. *Behavioral and Brain Sciences*, *33*(2–3), 215.Concepts, Kinds, and Cognitive Development, Cambridge, MA: MIT Press.

Keller, G. B., and T. D. Mrsic-Flogel. 2018. Predictive Processing: A Canonical Cortical Computation. *Neuron* 100 (2): 424–435.

Kemmerer, D. 2015. Are the motor features of verb meanings represented in the precentral motor cortices? Yes, but within the context of a flexible, multilevel architecture for conceptual knowledge. *Psychonomic Bulletin & Review* 22 (4): 1068–1075.

Kiefer, M., and F. Pulvermüller. 2012. Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex; A Journal Devoted To The Study Of The Nervous System And Behavior* 48 (7): 805–825.

Kruschke, J. K. 2005. Category learning. In *The handbook of cognition*, eds. K. Lamberts, and R. L. Goldstone, 183–201. Sage.

Kuhnke, P., M. Kiefer, and G. Hartwigsen. 2021. Task-Dependent Functional and Effective Connectivity during Conceptual Processing. *Cerebral Cortex* 31 (7): 3475–3493.

Kwong, J. M. 2006. Why concepts can't be theories. *Philosophical Explorations* 9 (3): 309–325.

Lake, B. M., T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. 2017. 'Building machines that learn and think like people'. *Behavioral and Brain Sciences* 40.

Lenci, A., G. E. Lebani, and L. C. Passaro. 2018. The Emotions of Abstract Words: A Distributional Semantic Analysis. *Topics in Cognitive Science* 10 (3): 550–572.

Löhr, G. 2020. Concepts and categorization: Do philosophers and psychologists theorize about different things? *Synthese* 197 (5): 2171–2191.

Machery, E. 2009. *Doing Without Concepts*. Oxford University Press.

Margolis, E., and S. Laurence. 1999. *Concepts: Core Readings*. Mit Press.

Margolis, E., and S. Laurence. 2010. Concepts and Theoretical Unification. *Behavioral and Brain Sciences* 33: 219–220.

Marr, D. 1982. *Vision*. Cambridge, MA: MIT Press.

Medin, D. L., and M. M. Schaffer. 1978. Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning (pp. 179–195). Cambridge University Press.

Michel, C. 2020a. Concept contextualism through the lens of Predictive Processing. *Philosophical Psychology* 33 (4): 624–647.

Michel, C. 2020b. Overcoming the modal/amodal dichotomy of concepts. *Phenomenology and the Cognitive Sciences*. https://doi-org.ezproxy.is.ed.ac.uk/10.1007/s11097-020-09678-y.

Murphy, G. L., and D. L. Medin. 1985. The role of theories in conceptual coherence. *Psychological Review* 92 (3): 289.
Nosofsky, R. M. 1986. Attention, similarity, and the identification categorization relationship. *Journal of Experimental Psychology: General* 115: 39–57.
Nosofsky, R. M., T. J. Palmeri, and S. McKinley. 1994. *Rule-plus-exception model of classification learning Psychological Review* 101: 53–79.
Osherson, D. N., and E. E. Smith. 1981. On the adequacy of prototype theory as a theory of concepts. *Cognition* 9 (1): 35–58.
Pecher, D. 2018. Curb Your Embodiment. *Topics in Cognitive Science* 10 (3): 501–517.
Piccinini, G., and S. Scott. 2006. Splitting concepts. *Philosophy of Science* 73 (4): 390–409.
Posner, M. I., and S. W. Keele. 1968. On the genesis of abstract ideas. *Journal of Experimental Psychology* 77 (3p1), 353–363.
Prinz, J. J. 2002. *Furnishing the mind: Concepts and their perceptual basis*. MIT Press.
Raut, R. V., A. Z. Snyder, and M. E. Raichle. 2020. Hierarchical dynamics as a macroscopic organizing principle of the human brain. *Proceedings of the National Academy of Sciences*, 117(34), 20890–20897.
Rice, C. 2016. Concepts as Pluralistic Hybrids. *Philosophy and Phenomenological Research* 92 (3): 597–619.
Rips, L. J. 1989. Similarity, typicality, and categorization. In *Similarity and analogical reasoning*, eds. S. Vosniadou, and A. Ortony, 21–59. Cambridge University Press.
Rosch, E. 1978. Principles of categorization. In *Cognition and categorization*, eds. E. Rosch, and B. B. Lloyd, 27–48. Hillsdale, NJ: Lawrence Erlbaum.
Smith, E., and D. Medin. 1999. The exemplar view. In *Concepts: Core Readings*, eds. E. Margolis, and S. Laurence, 207–222. MIT Press.
Smith, J. D., and J. P. Minda. 1998. Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning Memory and Cognition* 24: 1411–1436.
Smith, J. D., and J. P. Minda. 2000. Thirty categorization results in search of a model. *J Exp Psyc : Learning Memory and Cognition* 26: 3–27.
Sprevak, M. 2021a. Predictive coding I: Introduction. PhilSci-Archive URL: http://philsci-archive.pitt.edu/id/eprint/19365.
Sprevak, M. 2021b. Predictive coding III: Algorithm. PhilSci-Archive URL: http://philsci-archive.pitt.edu/id/eprint/19488.
Van Dam, W. O., M. Van Dijk, H. Bekkering, and S.-A. Rueschemeyer. 2012. Flexibility in embodied lexical-semantic representations. *Human Brain Mapping* 33 (10): 2322–2333.
van Pelt, S., L. Heil, J. Kwisthout, S. Ondobaka, I. van Rooij, and H. Bekkering. 2016. Beta-and gamma-band activity reflect predictive coding in the processing of causal events. *Social cognitive and affective neuroscience* 11 (6): 973–980.
Vanpaemel, W., G. Storms, and B. Ons. 2005. A varying abstraction model for categorization. In *Proceedings of the Annual Conference of the Cognitive Science Society* (Vol. 27, pp. 2277–2282). Lawrence Erlbaum Associates; Mahwah, NJ.
Verbeemen, T., W. Vanpaemel, S. Pattyn, G. Storms, and T. Verguts. 2007. Beyond exemplars and prototypes as memory representations of natural concepts: A clustering approach. *Journal of Memory and Language* 56 (4): 537–554.
Vicente, A., and F. Martínez Manrique. 2016. The Big Concepts Paper: A Defence of Hybridism. *British Journal for the Philosophy of Science* 67 (1): 59–88.
Vigliocco, G., S.-T. Kousta, P. A. Della Rosa, D. P. Vinson, M. Tettamanti, J. T. Devlin, and S. F. Cappa. 2014. The Neural Representation of Abstract Words: The Role of Emotion. *Cerebral Cortex* 24 (7): 1767–1777.
Voorspoels, W., G. Storms, and W. Vanpaemel. 2011. Representation at different levels in a conceptual hierarchy. *Acta psychologica* 138 (1): 11–18.
Walsh, K. S., D. P. McGovern, A. Clark, and R. G. O'Connell. 2020. Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences* 1464 (1): 242–268.
Weilnhammer, V. A., H. Stuke, P. Sterzer, and K. Schmack. 2018. The Neural Correlates of Hierarchical Predictions for Perceptual Decisions. *The Journal of Neuroscience* 38 (21): 5008–5021.
Weiskopf, D. A. 2009. The plurality of concepts. *Synthese* 169 (1): 145–173.

Weiskopf, D. A. 2011. The theory-theory of concepts. In James Fieser & Bradley Dowden (eds.), *Internet Encyclopedia of Philosophy.*https://iep.utm.edu/theory-theory-of-concepts/ (Last access: 16 April 2022).

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

Williams, D. 2018. Predictive Processing and the Representation Wars. *Minds and Machines* 28 (1): 141–172.

Yee, E., and S. L. Thompson-Schill. 2016. Putting concepts into context. *Psychonomic Bulletin & Review* 23 (4): 1015–1027.