Computational Brain & Behavior (2022) 5:234–243
https://doi.org/10.1007/s42113-022-00131-8

ORIGINAL PAPER

Structure Learning in Predictive Processing Needs Revision

Danaja Rutar[1,2] · **Erwin de Wolff**[1] · **Iris van Rooij**[1,3] · **Johan Kwisthout**[1]

Accepted: 31 January 2022 / Published online: 28 April 2022

Abstract
The predictive processing account aspires to explain all of cognition using a single, unifying principle. Among the major challenges is to explain how brains are able to infer the structure of their generative models. Recent attempts to further this goal build on existing ideas and techniques from engineering fields, like Bayesian statistics and machine learning. While apparently promising, these approaches make specious assumptions that effectively confuse structure learning with Bayesian parameter estimation in a fixed state space. We illustrate how this leads to a set of theoretical problems for the predictive processing account. These problems highlight a need for developing new formalisms specifically tailored to the theoretical aims of scientific explanation. We lay the groundwork for a possible way forward.

Keywords Predictive processing · Structure learning · Bayesian inference · Model expansion

Introduction

The predictive processing account is a key theoretical player in present-day cognitive neuroscience. The account postulates that our brains make sense of the world through a cycle of making predictions based on a hierarchical, generative model and updating its internal states to reduce prediction errors (Clark, 2013; Hohwy, 2013; Spratling, 2017; Walsh et al., 2020). While the account has been steadily growing in its explanatory scope, it also has its critics. For instance, some have argued that predictive processing presents old ideas under new labels (Cao, 2020), is not as unifying as proponents like to claim (Litwin & Miłkowski, 2020), or is difficult to falsify (Kogo & Trengove, 2015). Even those sympathetic of the approach, like ourselves, have posited substantive challenges. For instance, it has been shown that structured representations—needed to scale the account to higher-cognition (Griffiths et al., 2010; Kwisthout et al., 2017)—render the postulated (Bayesian) computations intractable and causes uncertainty about how they could be realized by resource-bounded wetware (Blokpoel et al., 2012; Kwisthout & van Rooij, 2020). Despite such critiques and challenges, proponents remain set on making the predictive processing account live up to its aspirations, i.e., to explain *all* of cognitive brain functioning.

The account's explanatory and modelling successes—spanning domains like perception (Den Ouden et al., 2012; Kok et al., 2013), action (Yon et al., 2018), planning (Kaplan & Friston, 2018), communication (Friston & Penny, 2011) and learning (Da Costa et al., 2020; Friston et al., 2017; Smith et al., 2020)—seem to warrant optimism. While we share some of this optimism and see no reason *in principle* why predictive processing cannot ultimately deliver on its promise,[1] it may be premature to take these local successes as directly foreshadowing

Danaja Rutar and Erwin de Wolff contributed equally to this work.

✉ Danaja Rutar
dr571@cam.ac.uk

✉ Erwin de Wolff
erwin.dewolff@donders.ru.nl

Iris van Rooij
iris.vanrooij@donders.ru.nl

Johan Kwisthout
johan.kwisthout@donders.ru.nl

[1] Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, the Netherlands

[2] Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK

[3] Department of Linguistics, Cognitive Science, and Semiotics & Interacting Minds Centre, Aarhus University, Aarhus, Denmark

[1] At least, we do not think its challenges are that more insurmountable than, or even fundamentally different from, those faced by other approaches.

ultimate, global success. All successes to date have been limited to phenomena that can be mathematically cast as Bayesian inference of *parameters* of structurally predefined generative models. Even predictive processing proponents have become increasingly aware of the need to build explanations and mathematical models of how brains are able to learn the *structure* of their generative models (Friston et al., 2017; Smith et al., 2020; Da Costa et al., 2020). Recent attempts, by these same authors, draw inspiration from formalisms developed in engineering applications of Bayesian statistics and machine learning. We believe that when formalisms are adapted from one field to another in this way, it is good practice to evaluate whether or not the conceptual commitments that come along match the commitments the researchers are ready to make (cf. Guest & Martin 2021; van Rooij & Blokpoel 2020). Aspects of formalisms that work well for engineering applications can be inapplicable or misconceived for purposes of scientific explanation (cf. van Rooij & Wareham 2008; van Rooij et al. 2012). In this paper, our focus is on scientific explanations as opposed to engineering applications.

In this theoretical paper, we evaluate the commitments of predictive processing models of (Bayesian) structure learning. We observe that these models make the contentious assumption that the generative model's state space is *fixed* (i.e., pre-defined and immutable).[2] We show that this assumption generates a theoretical problem for this account of structure learning: agents that learn within the confines of a fixed state space inevitably will get stuck in one of two cognitive states that we refer to as *cognitive blindness* and *category conflation*, respectively. The only way out of these states is to allow for true structural changes to occur in generative models that can, in principle, expand the state space. This route will require the predictive processing account to tackle a set of new open questions that we will highlight.

Overview

The remainder of this paper is organized as follows. We first introduce formal concepts and notation from predictive processing and explain a predictive processing view of structure learning called *model expansion*. Next, we present a proof argument that this model predicts that predictive Bayesian brains inevitably get stuck in *category conflation* or *cognitive blindness*. Lastly, we present a way out of this conundrum by considering mathematical possibilities for structural changes in generative models, which we call *structure learning proper*.

Formal Definition of Predictive Processing

Predictive processing proposes that human (and other) brains make predictions about the world based on an internal, hierarchical, generative model. This generative model is taken to represent a person's beliefs about concepts, relations, and transitions in the world (Clark, 2013; Friston et al., 2010). A generative model is defined as a hierarchy of prediction layers (see Fig. 1). Each such layer consists of a set of hypotheses H (or hidden states) with prior probability distribution $P(H)$ and a distribution $P(E \mid H)$ that describes how these hypotheses relate to the sensory input E. Between layers, the predicted $P(E)$ of a higher layer acts as the hypothesis distribution $P(H)$ for the layer below it. By explicitly modelling the causal dynamics of the world, a generative model cannot only make sense of observations, but also make predictions about future observations.

According to predictive processing, predictions form the basis of our interaction with the world. Due to the complexity and random nature of the world we live in, predictions made by a generative model are bound to have some degree of uncertainty. This is called *prediction error* (Den Ouden et al., 2012; Friston & Kiebel, 2009). It is assumed that brains reduce the uncertainty of their predictions as much as possible in order to reduce this prediction error over time (Clark, 2013; Friston et al., 2017). This prediction error reduction is important for a system to persist over time (Friston et al., 2012).

Structure Learning in Predictive Processing

In the predictive processing literature, *learning* typically refers to the process of optimizing the model parameters via the application of Bayes' theorem. Here, we consider another form of learning referred to as *structure* learning. This type of learning is not concerned with parameter optimization but rather with learning the structure of the generative model. More specifically, structure learning pertains to learning the generative models' variables and their functional dependencies (Smith et al., 2020; Da Costa et al., 2020).[3] Recently, the first attempts have been made towards a predictive processing formalization of structure learning. Specifically, *model reduction* (Friston & Penny, 2011) and *model expansion* (Smith et al., 2020) have

[2] One may argue that most cognitive scientists do believe that human generative models are flexible, adapting their structural properties across time and circumstance. Whilst some predictive processing theorists may agree, their generative models do not embody this view, and instead make the contentious assumption that we noted (Smith et al., 2020; Da Costa et al., 2020; Friston et al., 2017). We thank an anonymous reviewer for raising this issue.

[3] Our definition of structure learning differs from the more data-scientific perspective on structure learning, which focuses more on quantifying relations within data through algorithms. See, for example, Madsen et al. (2017) or Pinto et al. (2009).

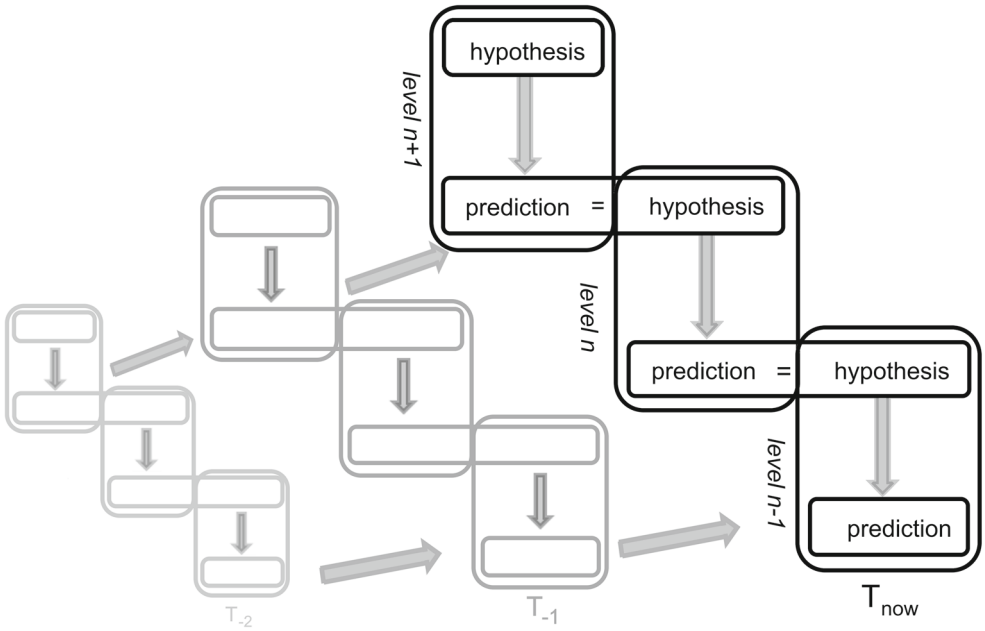

Fig. 1 A hierarchical generative model. Prediction layers at level n of the model act as hypothesis layers level $n-1$. Beliefs of previous timesteps $T-1$ influence the beliefs of future timesteps T_{now}. Adapted with permission from Kwisthout et al. (2017)

been proposed as complementary methods for changing the structure of a generative model. In this paper, we will focus solely on model expansion. We will show how this formalism leads to conceptual problems. We omit discussion of model reduction as this method focusses on reducing the size of the generative model, while the conceptual problems we identify cannot be remedied by reducing the size of a generative model.

In model expansion, generative models have two distinct types of hypotheses: *latent hypotheses* $\boldsymbol{u} = \{u_1, ..., u_n\}$ (also referred to as "spare slots") and *explicit hypotheses* $\boldsymbol{h} = \{h_1, ..., h_m\}$. The full hypothesis space is the union of the two types of hypotheses, i.e., $H = \boldsymbol{u} \cup \boldsymbol{h}$. Prior to learning all hypotheses are latent, i.e., $\boldsymbol{h} = \emptyset$. Each hypothesis $h \in H$ has a likelihood function $P(E \mid H = h)$. For latent hypotheses, this likelihood function is a noisy uniform distribution. For explicit hypotheses, the likelihood is developed over time based on observations as follows: when making an observation o, the agent determines which hypothesis h best explains that observation based on the posterior probability $h = argmax_{h'} P(H = h' \mid E = o)$. This posterior probability is computed by applying Bayes' rule:

$$P(H \mid E = o) = \frac{P(E = o \mid H) \times P(H)}{P(E = o)} \tag{1}$$

Then, this most explanatory hypothesis is updated in two ways. First, the parameters that define the likelihood function are updated according to Bayes' rule. Secondly, the distribution over the hypotheses is updated such that the prior probability $P(H)$ of the next prediction cycle is set to the posterior probability of the current: $P_{t+1}(H) = P_t(H \mid o)$.[4]

If the best hypothesis h was in $\boldsymbol{u}$ prior to this update, it is "moved" from $\boldsymbol{u}$ to $\boldsymbol{h}$. It is this transition from a latent hypothesis u to an explicit hypothesis h that is referred to as "effectively expanding" the generative model. As such, a latent hypothesis $u \in \boldsymbol{u}$ is made explicit when none of the explicit hypotheses $h \in \boldsymbol{h}$ account better for the observation o than u. Note that the transition from latent to explicit is merely a verbal relabeling. Smith et al. (2020) refer to this kind of state space expansion as "effective" because the dimensions of the state space remain the same throughout. The only thing that changes is the number of explicit hypotheses used by the agent. Crucially, the total number of hypotheses, latent *plus* explicit, cannot be changed through learning. Consequently, the models that make use of model expansion have a fixed state space. That leaves one to wonder whether model expansion, or other models that similarly assume a fixed state space, can truly capture *structure* learning.

Problems Stemming from a Fixed State Space

We now describe two problems that emerge from a fixed-state space model of structure learning (such as model

[4]This dual updating is not a default choice. Sometimes, only the likelihood functions are updated (Smith et al., 2020). Other times, both are updated, and finally there are cases when neither is updated at all.

expansion). We define the problems formally, then provide a real-world example where the problem can occur. We prove that under a plausible assumption (that there are more categories in the world than hypotheses in a fixed state space) either of these two problems will inevitably occur (Fig. 2).

To explain the problems, we need to introduce some definitions. We define the world as a set of categories $\mathcal{O}$. Each element in this set is a specific category $O \in \mathcal{O}$. Furthermore, each such category is a set of instances $o \in O$. Making an observation is defined as evidencing the prediction variables with a particular instance of a category $o \in O$, such that $P(E = o) = 1$.

We define what it means for a category $O \in \mathcal{O}$ to be represented by a concept $h \in H$ as follows:

$$R(O, h) \equiv \forall_{o \in O}(argmax_{h'} P(H = h' \mid o) = h) \tag{2}$$

Thus, we say the hypothesis h *represents* a category O for an agent if for each instance of that category $o \in O$, the hypothesis h is the most probable given the evidence. That is, h is the *maximum a posteriori* (MAP) hypothesis for all $o \in O$. In real life, classification will never be this perfect and will instead be error prone. However, for ease of presentation we will work with this idealized error-free scenario; i.e., we assume that if h represents the category O, then h is the MAP hypothesis for all instances of category O. As will become clear, our conclusions do not rely on this assumption, even if it makes exposition easier.

Category Conflation

The first problem that we consider is category conflation. Category conflation is the cognitive phenomenon where an agent represents two distinct categories $O_1, O_2 \in \mathcal{O}$ in the world by one and the same hypothesis $h \in H$. In addition, there is no other hypothesis $h' \neq h$ that represents either of these categories. As such, the agent will classify instances of both (distinct) categories O_1 and O_2 as instances of h. In other words, for the agent the two categories are indistinguishable.

Given the definition in Eq. 2, we can formally define *category conflation* given hypotheses H and categories $O_1, O_2 \in \mathcal{O}$ as follows:

$$Conf(O_1, O_2, H) \equiv O_1 \neq O_2 \wedge \exists_{h \in H}(R(O_1, h) \wedge R(O_2, h)) \tag{3}$$

Let us illustrate the idea with a concrete example. Imagine an agent observing ducks and geese. Imagine further that due to the structure of its generative model, the agent believes that ducks and geese are different instances of the same category (we will call this combined category *guck* as opposed to the excellent alternative *deese*). It is not difficult to see how this situation may arise. Ducks and geese are quite similar to one another, especially visually. Furthermore, they can often be found in the same places, such as public parks, within the same general "group" of animals. Lastly, ducks and geese are strikingly different from, say, the plants, trees and decorations that tend to be near them. Whatever the exact reason for the confusion, the agent now has a problem. Whenever a duck or goose is seen, the agent infers that this is a *guck*. It has conflated two categories (ducks and geese) into a single hypothesis (gucks).

Cognitive Blindness

The second problem that we consider is cognitive blindness. Cognitive blindness is the cognitive phenomenon where a specific category $O \in \mathcal{O}$ in the world is not represented by any hypothesis $h \in H$ in the generative model of the agent. Therefore, the agent is blind to the category O. Using the definition of representation in Eq. 2, we formally define *cognitive blindness* with hypothesis variables H and a category $O \in \mathcal{O}$ as follows:

$$Blind(O, H) \equiv \neg\exists_{h \in H} R(O, h) \tag{4}$$

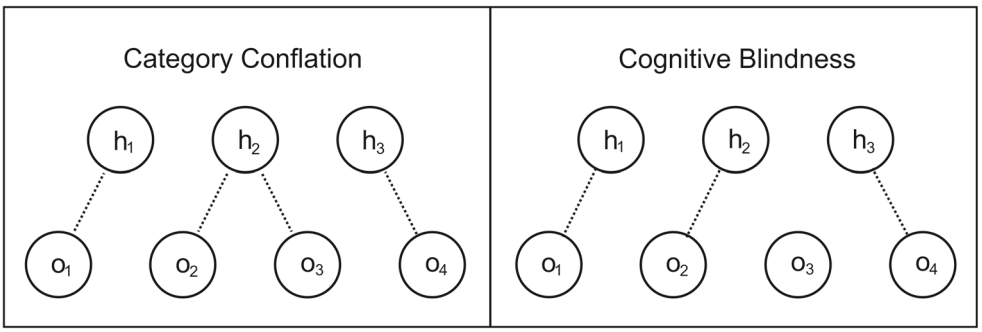

Fig. 2 A visual representation of the two problems discussed in this paper. Hypotheses are indicated with a letter h, categories with a letter O, and dotted lines denote that a category is represented by a hypothesis ($R(O, h)$). In the situation on the left, categories O_2 and O_3 are both represented by h_2. This is an example of *category conflation*. In the situation on the left, category O_3 is not represented by any hypothesis. This is an example of *cognitive blindness*

Let us again illustrate the idea with a concrete example. Imagine that an agent has three hypotheses, representing flamingos, geese and swans respectively. At some point, this agent encounters an ostrich. Now, the agent is confused. Because the ostrich shares some similarities with each of the agent's hypotheses (such as colour, "kind"—they are all birds, and long shaped neck), it can be interpreted as belonging to any of these hypotheses. Which bird the ostrich is classified as will depend on the circumstances wherein the agent sees the ostrich (How far away is the agent? How clearly can be the colours be seen? Is only the neck visible?). Because these circumstances will inevitably vary, ostriches will sometimes be interpreted as a flamingo, sometimes as a goose and sometimes as a swan. Any one of the hypotheses that share enough features with the category ostrich can therefore be deemed the most likely hypothesis of an ostrich. As such, the agent is cognitively blind to the concept ostrich.

Proof

We will now prove that a generative model with a fixed state space will necessarily either conflate categories or be blind to categories if there are more categories in the world than the agent can represent in such a model. The assumption may seem bold, but we should consider the number of categories that we need to learn during our lifetime. All the animals, plants, tools, words, songs, types of food, faces, furniture and so much more need to be represented by a "spare slot." Furthermore, even if evolution conveniently provided us with the right number of such slots for a particular time, we would be missing a means of representing all the new categories that are constantly being developed and discovered around us (e.g., the notions of a "neutrino," "blockchain," "tweet," and "bitcoin" did not exist prior to 1930, 1998, 2006, and 2009 respectively).

1. Let $| \mathcal{O} | > | H |$
2. By tautology, we have that: $(\forall_{O \in \mathcal{O}} \exists_{h \in H} R(O, h)) \vee \neg(\forall_{O \in \mathcal{O}} \exists_{h \in H} R(O, h))$
3. For step 2, suppose that $\neg(\forall_{O \in \mathcal{O}} \exists_{h \in H} R(O, h))$
4. From 3, it follows that $\exists_{O \in \mathcal{O}} \neg \exists_{h \in H} R(O, h)$
5. Step 4 is equivalent to $\exists_{O \in \mathcal{O}} Blind(O, H)$. Therefore, supposing 3 under assumption 1 leads to cognitive blindness.
6. For step 2, suppose that $\forall_{O \in \mathcal{O}} \exists_{h \in H} R(O, h)$
7. From 1 and 6, it follows by pigeonhole principle that $\exists_{O_1, O_2 \in \mathcal{O}} (O_1 \neq O_2 \wedge \exists_{h \in H} (R(O_1, h) \wedge R(O_2, h)))$
8. Step 7 is equivalent to $\exists_{O_1, O_2 \in \mathcal{O}} (Conf(O_1, O_2, H))$ Therefore, supposing 6 under assumption 1 leads to category conflation.
9. Combining steps 1, 2, 3, 5, 6, and 8, we conclude that every model must either suffer from category conflation or cognitive blindness, given that there are more categories in the world than there are concepts in the model.

Human Generative Models are Not Fixed

Our proof shows that structure learning on a fixed state space, as proposed by, e.g., *model expansion* (Smith et al., 2020), inevitably will end up conflating categories or be blind to categories in a world with ever-expanding number of categories.

Arguably, humans may also conflate categories from time to time, and they can at times be cognitively blind to certain features in the world. Be that as it may, unlike the fixed models adopted in the model expansion account of structure learning, humans certainly do not appear to be stuck indefinitely when these states arise. When humans find themselves with conflated categories, or when they come across something unfamiliar to them, they will often readily introduce new concepts to try and explain (i.e., resolve) the problem. For instance, while a child may indeed believe initially that ducks and geese are one and the same species guck, they will learn at some point that the animals are different species. Similarly, an explorer that first sees an ostrich might consider all other bird hypotheses to be insufficiently explanatory, and will introduce a new hypothesis "ostrich" instead. New hypotheses added in these situations are often wrong, perhaps most of the time so, but regardless of the verity of the new concept, the cognitive problem is actively being solved. In fact, humans continually hone their skills in structure learning through tools acquired through formal education, scientific training and explicit feedback from other people. For examples of successful structure learning, see the section below. So, if the predictive processing account aims to explain how the brain *actually* performs structure learning, we need to move beyond fixed models.

A Way Forward: Structure Learning Proper

In the previous section, we have argued that current attempts to model structure learning in predictive processing run into problems due to their reliance on a fixed state space. Therefore, predictive processing models will need to incorporate operations that allow for changing the *structure* of the generative models, if they want to avoid these problems. We refer to these types of operations as *structure learning proper*. We define structure learning proper as the process of going from a generative model G to a generative model G' in a way that can not be explained by Bayesian (parameter) updating alone. Table 1 presents an overview of a number of changes that fall under structure learning

Table 1 Options for structural changes to generative models formalized as Bayesian networks. Note that the possible preconditions mentioned here are mere suggestions. They should not be interpreted as a commitment to a particular stance

Structural changes	Real-life examples	Formalized transformations	Possible preconditions
Add a new variable	A child learns about the existence of bacteria.	$V_n \leftarrow V_{n-1} \cup \{v\}$, where $v \notin V_{n-1}$	An observation o which is novel to all variables $u \in V_{n-1}$.
Remove a variable	A child stops believing in Santa Claus.	$V_n \leftarrow V_{n-1} - \{v\}$, where $v \in V_{n-1}$	The prior probability of v is below some existence-threshold λ.
Merge two variables	Physicists discover that "temperature" and "energy" are the same thing.	$V_n \leftarrow V_{n-1} - \{v_1, v_2\} + \{\hat{v}\}$, where $\hat{v} = v_1 + v_2$	v_1 and v_2 share the exact same relations with the other variables in V_{n-1}.
Split a variable	Psychologists realize that "sex" and "gender" are distinct.	$V_n \leftarrow V_{n-1} - \{\hat{v}\} + \{v_1, v_2\}$, where $\hat{v} = v_1 + v_2$	Distinct groups of values of v relate sufficiently differently to other variables in V_{n-1}.
Add a new causal connection	Students learn that thunderstrokes are the result of unstable air mass.	$A_n \leftarrow A_{n-1} \cup \{a\}$, where $a \notin A_{n-1}$	A stable, sizeable improvement of the likelihood of observations given $a \in A_n$ compared to $a \notin A_n$.
Remove a causal connection	A person concludes that the amount of water has no effect on the colour of aloe vera.	$A_n \leftarrow A_{n-1} - \{a\}$ where $a \in A_{n-1}$	The difference in the likelihood of observations given $a \in A_n$ is not worth the complexity of the model compared to $a \notin A_n$.
Add a new value for a variable	A headmaster learns that the "fist bump" is a new greeting.	$val(X) \leftarrow \{x_1, ..., x_n\} \cup \{x_{n+1}\}$ with $x_{n+1} \notin val(X)$	An observation o that is classified as an instance of H is better explained by a new value for h.
Remove a value of a variable	An incorrectly learned word is removed after a period of non-use.	$val(X) \leftarrow \{x_1, ..., x_n\} - \{x_k\}$ with $x_k \in val(X)$	Some decay of the weights of $val(X)$ reduces the weight of x_k to zero.
Merge two values of a variable	Two different sounds are concluded to be the same phoneme 'r'.	$P(\hat{h}) \leftarrow P(h_1) + P(h_2)$ $P(E \mid \hat{h}) \leftarrow \frac{P(E\mid h_1)\times P(h_1)}{P(h_1)+P(h_2)} + \frac{P(E\mid h_2)\times P(h_2)}{P(h_1)+P(h_2)}$	Minimal difference between the two likelihood functions of the values h_1 and h_2.
Split a value of a variable	A culture is introduced to the blue/green distinction.	Replace $\hat{h}$ by some $\{h_1, h_2\}$ such that the merging of h_1 with h_2 gives back $\hat{h}$	$\hat{h}$ has a variance that exceeds some maximum variance.

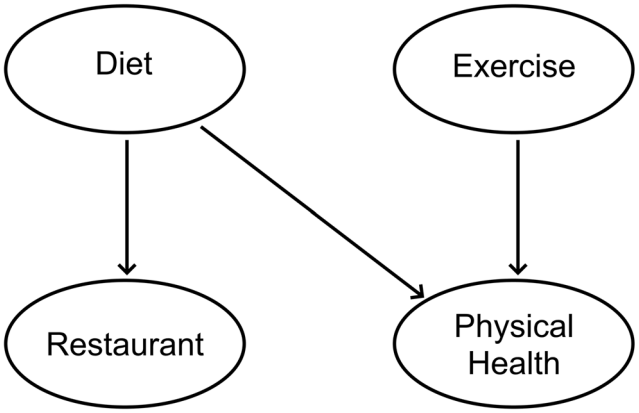

Fig. 3 Example of a Bayesian network. The bubbles represent variables, and the arrows denote (causal) relations between those variables. Here, the variable "Diet" influences "Restaurant": what kind of diet you are on limits the choice of restaurants you can go to. Both "Diet" and "Exercise" influence "Physical Health": A healthy diet and more exercise lead to better physical health

proper. These changes are presented both verbally and formally. The verbal description gives an intuitive name for the type of change made to the generative model. In addition, Table 1 also gives intuitive real-world examples for each listed simple change. With these examples, we want to illustrate that our proposed changes are not just theoretical constructs, but also reflect aspects of human learning and how humans make changes to their generative models. In real life, some of these changes may occur simultaneously and some may happen more often than others. For ease of presentation, we illustrate the possible structural changes in isolation and in a minimal form.

The formal description takes the verbal description and translates it into a mathematical formalisation of the same change. To allow for these formal translations, we define the generative models postulated by the predictive processing account as *Bayesian networks* (see Fig. 3).[5] In the Bayesian networks that we consider we designate two types of variables: hypothesis variables H and prediction variables E. Hypothesis variables are equivalent to hidden states or concepts in predictive processing literature. Likewise, prediction variables are equivalent to outcomes. The values of a Bayesian variable X, which can either be a hypothesis or a prediction variable, are denoted with $val(X) = \{x_1,...,x_n\}$, and a single value as x. The structure of the generative model is defined as a directed graph $G = (V, A)$, where V is the set of vertices in G, and A is the set of arcs in G. Here, vertices are variables, and arcs indicate dependencies between those variables.

[5]The Hidden Markov Models (HMM) commonly used in the predictive processing literature are a special case of Bayesian networks. For arguments for adopting the more general formalism see Kwisthout et al. (2017).

Lastly, the table lists a conceivable precondition for each simple change to occur (or for it to be justified to occur). These preconditions serve as an inspiration and illustration, and should not be interpreted as any definitive answer. The reason why we only provide an example is that for most of these changes these preconditions have not been both formally defined and empirically investigated. What the real preconditions are is therefore an open question. This paper does not aim to give answers to these questions, because each question would demand a research project to address, if not a body of research. This then is the challenge: In order for predictive processing to account for structure learning proper, it must investigate the formal and empirical underpinnings of the preconditions under which our proposed changes occur in human structure learning. We stress that the ultimate goal here is not to find statistical or engineering solutions, but to formulate an explanatory scientific account of structure learning proper at Marr's (1982) computational level.

A Note on Bayesian Non-parametrics

Before we close, we reflect on the relationship between the challenge that we pose and an existing approach in computational cognitive science that takes inspiration from Bayesian non-parametrics. Readers familiar with that approach may believe that that method already addresses the challenge that we pose. Here, we briefly explain how and why it does not.

Bayesian non-parametrics is a method firmly rooted in Bayesian statistics that has found appreciation in modelling cognition (Griffiths et al., 2006; Perfors et al., 2011; Chater et al., 2010; Austerweil & Griffiths, 2013). A central tenet of Bayesian non-parametrics is that generative models can be defined without having to specify the number of values within a variable, the amount of layers in a belief model and other such parameters; hence the name "non-parametric" (see Gershman & Blei 2012; Griffiths & Ghahramani 2011, for reviews). In other words, this method allows modellers to define a generative model that is *not* fixed, but can flexibly expand and shrink. This property notwithstanding, the approach has some limitations relevant to the challenge that we pose.

The first limitation of Bayesian non-parametrics is that it only captures a proper subset of the possible structural changes listed in Table 1. Namely, to the best of our knowledge Bayesian non-parametrics is not (yet) capable of discovering new variables or establishing or removing causal connections between variables. Of course, one might argue that is it not necessary that our brains perform *all* the changes listed in Table 1. We can agree with this argument to some degree. However, the very structural

changes missing in Bayesian non-parametrics are so central to our cognition that any account of human learning that cannot account for them is incomplete at best, and flawed at worst. Given that Bayesian non-parametrics is missing the capacity to enact some of the cognitively relevant structural changes, there will *necessarily* be situations where Bayesian non-parameterics predicts a structural change that does not match the structural change that *actually* occurs in human learning. Without a more complete account of structure learning we have no way of predicting which situations these could be.

The second, and arguably more important limitation of Bayesian non-parametrics is that it commits to a very particular stance of when and why values are changed in a generative model. This view, if considered at Marr's computational level (Marr, 1982), postulates that generative models change in order to optimize the likelihood of future observations. This is indeed one possible precondition for a structural change to a generative model. It is, in fact, one formalisation of the precondition for adding a new value that we gave in Table 1, row 7. However, we can conceive of many other formalisations of "better explained" (column 4, row 7), that will lead to (very) different preconditions. Consider, for example, Kwisthout (2013), who proposes that in judging the quality of an explanation cognizers may weigh both informativeness and probability. This idea conflicts with the metric of likelihood alone used in Bayesian non-parametrics. Without knowing which preconditions most accurately describe structural changes as they apply to human learning, we should not treat Bayesian non-parametrics as the only possible answer.

To summarize, while we appreciate the perspective that Bayesian non-parametrics offers on a relevant subset of the changes that make up structure learning proper, we conclude that there are still important changes that are unaccounted for. Furthermore, we argue that there are other plausible preconditions possible that are not captured by Bayesian non-parametrics. Our challenge to predictive processing theorists, thus, remains to investigate which of these preconditions best explains human structure learning.

Conclusion

Learning in predictive processing has mostly been conceptualized as *parameter* learning. Recently, several developments have been made that aim to tackle the problem of *structure* learning. Structure learning is important as it can help explain why generative models have the structure that they do, as well as how this structure is learned over time. However, despite the name *structure learning*, these new developments in predictive processing are essentially a particular type of Bayesian parameter learning in a fixed state space.

We showed that generative models with a fixed state space will inevitably lead to at least one of the two conceptual problems. These problems are *category conflation*, where different categories in the world are interpreted as the same concept, and *cognitive blindness*, where a category in the world is not represented by any concept. Furthermore, we argued that when humans run into these problems, they have strategies that allow them to solve them. These strategies are best understood as structural changes to a generative model.

We presented an exhaustive set of minimal changes that can be made to a Bayesian generative model, defining structure learning *proper* (see Table 1). The idea that such changes are possible is not new: structure learning proper has been studied in the past, with or without a cognitive perspective (Tsamardinos et al., 2006; Piantadosi et al., 2016; Piantadosi, 2021; Perfors, 2012; Chickering, 1996; Chickering et al. 1994). Either this research has not provided a formal account of all possible and cognitive relevant structural changes or they have not explored the conditions under which these might occur in human learning. Given that we provide a formal characterisation of possible structural changes in this paper, our proposed way forward is to research cognitively plausible preconditions of when each change occurs.

We take the position that current theories of structure learning should integrate the proposed proper structural changes into their explanatory toolbox. The importance of structural changes, like the ones we suggest, cannot be understated as a means to explain the richness of human learning. If predictive processing wishes to live up to its ambition to explain all of learning, it needs to make the transition from its current formalisation of "effective" structure learning to structure learning proper.

Funding DR was supported by a Donders Centre for Cognition grant awarded to JK, EdW was supported by a Donders Centre for Cognition grant awarded to JK and IvR. IvR acknowledges the support of a Distinguished Lorentz fellowship by the Netherlands Institute for Advanced Studies in the Humanities and Social Sciences (NIAS-KNAW) and the Lorentz Center.

References

Austerweil, J. L., & Griffiths, T. (2013). A nonparametric Bayesian framework for constructing exible feature representations. *Psychological Review*, *120*(4), 817.

Blokpoel, M., Kwisthout, J., & van Rooij, I. (2012). When can predictive brains be truly Bayesian? *Frontiers in Psychology*, *3*, 406.

Chickering, D. M. (1996). Learning Bayesian networks is NP-complete. In *Learning from data* (pp. 121–130). New York: Springer.

Chickering, D. M., Geiger, D., Heckerman, D., et al. (1994). Learning Bayesian networks is NP-hard. Technical report, Technical Report MSR-TR-94-17 Microsoft Research.

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., & Friston, K. (2020). Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology*, *99*, 102447.

Den Ouden, H. E., Kok, P., & De Lange, F. P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in Psychology*, *3*, 548.

Cao, R. (2020). New labels for old ideas: Predictive processing and the inter- pretation of neural signals. *Review of Philosophy and Psychology*, *11*(3), 517–546.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 811–823.

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

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221.

Friston, K., & Penny, W. (2011). Post hoc Bayesian model selection. *Neuroimage*, *56*(4), 2089–2099.

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

Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, *3*, 130.

Friston, K., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., & Ondobaka, S. (2017). Active inference, curiosity and insight. *Neural Computation*, *29*(10), 2633–2683.

Gershman, S. J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*(1), 1–12.

Griffiths, T., & Ghahramani, Z. (2011). The Indian buffet process: an introduction and review. *Journal of Machine Learning Research*, *12*(4), 1185–1224.

Griffiths, T. L., Navarro, D. J., & Sanborn, A. N. (2006). A more rational model of categorization. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 28).

Griffiths, T., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802.

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

Kaplan, R., & Friston, K. (2018). Planning and navigation as active inference. *Biological Cybernetics*, *112*(4), 323–343.

Kogo, N., & Trengove, C. (2015). Is predictive coding theory articulated enough to be testable? *Frontiers in Computational Neuroscience*, *9*, 111.

Kok, P., Brouwer, G. J., van Gerven, M. A., & de Lange, F. P. (2013). Prior expectations bias sensory representations in visual cortex. *Journal of Neuroscience*, *33*(41), 16275–16284.

Kwisthout, J. (2013). Most inforbable explanations: Finding explanations in Bayesian networks that are both probable and informative. In *European conference on symbolic and quantitative approaches to reasoning and uncertainty* (pp. 328–339). Springer.

Kwisthout, J., & van Rooij, I. (2020). Computational resource demands of a predictive Bayesian brain. *Computational Brain & Behavior*, *3*(2), 174–188.

Kwisthout, J., Bekkering, H., & van Rooij, I. (2017). To be precise, the details don't matter: On predictive processing, precision, and level of detail of predictions. *Brain and Cognition*, *112*, 84–91.

Litwin, P., & Miłkowski, M. (2020). Unification by fiat: Arrested development of predictive processing. *Cognitive Science*, *44*(7), 12867.

Madsen, A. L., Jensen, F., Salmerón, A., Langseth, H., & Nielsen, T. D. (2017). A parallel algorithm for Bayesian network structure learning from large data sets. *Knowledge-Based Systems*, *117*, 46–55.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information* (pp. 69–73). New York: The MIT Press.

Perfors, A. (2012). Bayesian models of cognition: What's built in after all? *Philosophy Compass*, *7*(2), 127–138.

Perfors, A., Tenenbaum, J. B., Griffiths, T., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*(3), 302–321.

Piantadosi, S. T. (2021). The computational origin of representation. *Minds and Machines*, *31*(1), 1–58.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, *123*(4), 392.

Pinto, P. C., Nagele, A., Dejori, M., Runkler, T. A., & Sousa, J. M. (2009). Using a local discovery ant algorithm for Bayesian network structure learning. *IEEE Transactions on Evolutionary Computation*, *13*(4), 767–779.

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition*, *112*, 92–97.

Smith, R., Schwartenbeck, P., Parr, T., & Friston, K. J. (2020). An active inference approach to modeling structure learning: Concept learning as an example case. *Frontiers in Computational Neuroscience*, 14.

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, *65*(1), 31–78.

van Rooij, I., & Blokpoel, M. (2020). Formalizing verbal theories: A tutorial by dialogue. *Social Psychology*, *51*(5), 285–298.

van Rooij, I., & Wareham, T. (2008). Parameterized complexity in cognitive modeling: Foundations, applications and opportunities. *The Computer Journal*, *51*(3), 385–404.

van Rooij, I., Wright, C. D., & Wareham, T. (2012). Intractability and the use of heuristics in psychological explanations. *Synthese*, *187*(2), 471–487.

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences*, *1464*(1), 242.

Yon, D., Gilbert, S. J., de Lange, F. P., & Press, C. (2018). Action sharpens sensory representations of expected outcomes. *Nature Communications*, *9*(1), 1–8.