Predictive coding and the Bayesian brain: Intractability hurdles that are yet to be overcome

Johan Kwisthout (j.kwisthout@donders.ru.nl)
Iris van Rooij (i.vanrooij@donders.ru.nl)
Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour
Montessorilaan 3, 6525 HR Nijmegen, The Netherlands

Abstract

There is a growing body of evidence that the human brain may be organized according to principles of hierarchical predictive coding. A current conjecture in neuroscience is that a brain organized in this way can effectively and efficiently perform genuine Bayesian inferences. Given that many forms of cognition seem to be well characterized as Bayesian inferences, this conjecture has great import for cognitive science. It suggests that hierarchical predictive coding may provide a neurally plausible account of how forms of cognition that are modeled as Bayesian inference may be physically implemented in the brain. Yet, as we show in this paper, the jury is still out on whether or not the conjecture is really true. Specifically, we demonstrate that each key subcomputation invoked in hierarchical predictive coding potentially hides a computationally intractable problem. We identify ways in which computational modelers may or may not overcome these 'intractability hurdles.'

Keywords: theoretical cognitive neuroscience; hierarchical predictive coding; Bayesian modeling; computational complexity theory; NP-hard; intractability; fixed-parameter tractability; approximation

Introduction

The hierarchical predictive coding framework (Rao & Ballard, 1999; Friston, 2002; Friston, 2005) is becoming increasingly popular as an account of perceptual, behavioral, and neural phenomena in cognitive neuroscience. Examples of its explanatory successes include, for instance, binocular rivalry (Hohwy, Roepstorff, & Friston, 2008), neural repetition suppression (Summerfield et al., 2008), motor planning (Brown, Friston, & Bestmann, 2011), and action understanding (Kilner, Friston, & Frith, 2007; Den Ouden, Kok, & De Lange, 2012). Some of the popularity of the framework seems to hinge on its presumed ability to give a neurally plausible and computationally feasible account of how Bayesian inferences can be efficiently implemented in brains. This presupposition is, for instance, reflected in the writings by philosopher Andy Clark on the various virtues of the hierarchical predictive coding framework. For instance, in a recent *BBS* target article he notes:

> It is thus a major virtue of the hierarchical predictive coding account that it effectively implements a computationally tractable version of the so-called Bayesian Brain Hypothesis. (Clark, in press)

In a different place, Clark (in press) phrases the idea as follows:

> the hierarchical (…) predictive processing story, if correct, would rather directly underwrite the claim that the nervous system approximates, using tractable computational strategies, a genuine version of Bayesian inference. (Clark, in press)

If Clark is right, then the hypothesis that brains are organized according to the principles of hierarchical predictive coding is not only of great importance for neuroscience—as a hypothesis about the *modus operandi* of the human brain—but also for contemporary cognitive science. After all, it would directly suggest a candidate explanation of how the probabilistic computations postulated by the many Bayesian models in contemporary cognitive science (Griffiths, Kemp, & Tenenbaum, 2008) may be realistically implemented in the human brain.

The promise that the hierarchical predictive coding framework would yield a tractable way to perform (approximate) Bayesian inference seems all the more important in light of the fact that Bayesian models of cognition have been plagued by complaints about their apparent computational intractability (e.g., Gigerenzer, 2008; Kwisthout, Wareham, & van Rooij, 2011). Having a candidate neural story on offer about how Bayesian inference may yet be tractable for human brains would, of course, further strengthen the already strong case for the Bayesian approach in cognitive science (see, e.g., Tenenbaum, Kemp, Griffiths, & Goodman, 2011, and the references therein). In this paper we set out to analyze to what extent the framework already makes true on this promise, or otherwise could in the future.

The approach that we take is as follows. We formally model each required subcomputation postulated by the hierarchical predictive coding framework for the forward (bottom-up) and backward (top-down) chains of processing. Our models characterize these computations in terms of the basic transformations that they are assumed to perform. We distinguish three key transformations: prediction, error computation, and hypothesis updating. We will show that, unless the causal models underlying these subcomputations are somehow constrained, each of these subcomputations is *in and of itself* intractable to compute (i.e., NP-hard), whether exactly or approximately. This means that just by postulating a predictive coding implementation of Bayesian inference one does not automatically achieve tractability.

This is not to say that Bayesian inference cannot ever be tractably implemented by a hierarchical predictive coding scheme, but it does mean that this is only possible if the type of Bayesian inference to be performed is already tractable to begin with, e.g., because it is performed on causal structures that have special features that make such inferences tractable (see also Blokpoel, Kwisthout, van Rooij, 2012; Kwisthout, Wareham, & van Rooij, 2011). We will present results on what these special features may be for Bayesian inferences implemented via hierarchical predictive coding.

The remainder of this paper is organized as follows. First, we explain the basic ideas of the hierarchical predictive coding framework in more detail. Next, we propose our formal models of its postulated subprocesses. We then present computational (in)tractability analyses for our formal models and discuss the implications of our findings for the presumed tractability of predictive coding and its implementation of Bayesian inference.

The predictive coding framework

In the predictive coding framework, it is assumed that the brain continuously tries to predict its sensory inputs on the basis of a hierarchy of hypotheses about the world (Friston, 2002). The predictions are formed by the so-called 'backward chain' (top-down processing), in which higher order hypotheses are transformed to lower order hypotheses an eventually predictions about sensory inputs. This form of top-down processing augments the 'forward chain' (bottom-up processing) in which sensory inputs are transformed to increasingly more abstract hypotheses. Fig. 1 presents a schematic illustration. By comparing predicted observations at each level *n* in this hierarchy with the actual observations at the same level *n* in this hierarchy, the system can determine the extent to which its predicted observations in the backward chain match the observations arising from the forward chain, and update its hypotheses about the world accordingly.

An example of the explanatory uses of the predictive coding framework is its explanation of *binocular rivalry* in vision (Hohwy et al., 2008). When an image of a house is presented to the left eye, and an image of a face to the right eye, the subjective experience of the images alternates between a face and a house, rather than some combination of the separate stimuli. As we are not familiar with blended house-faces such a combination would have a low prior probability, hence either a house or a face is predicted to be observed. The actual observation, however, triggers a prediction error: a mismatch between what was predicted (e.g., a face), and what was observed (both a house and a face). This mismatch then leads to an updated hypothesis; taking prior probabilities as well as the prediction error into account, the hypothesis will then shift towards a house, rather than a combination of a face and a house.

The predictive coding framework is often computationally implemented using a hierarchical Bayesian network structure, where posterior probabilities on one level of the hierarchy provide priors for its subordinate level (e.g., Kilner et al., 2007; Lee & Mumford, 2003; Rao & Ballard,

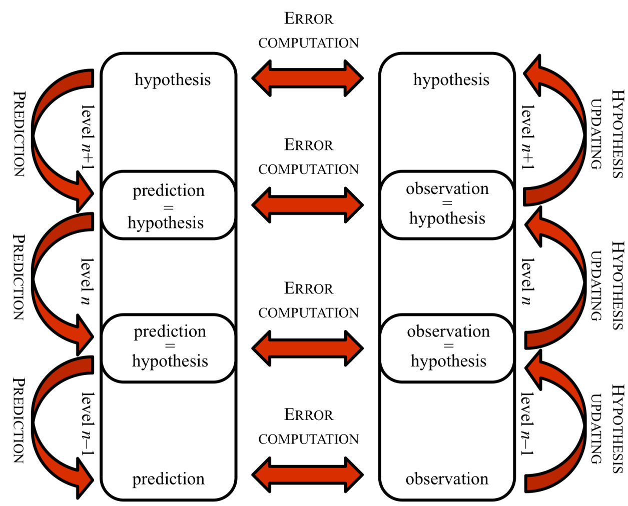

Fig 1. A schematic depiction of the way in which the backward chain (left) and the forward chain (right) in the hierarchical predictive coding framework interact. The red arrows are labeled by the key computational transformations that we analyze in this paper: PREDICTION, ERROR COMPUTATION and HYPOTHESIS UPDATING. See text for more details.

1998), i.e., each level of the hierarchy can be seen (for the sake of computational analysis) as a separate Bayesian network, consisting of hypothesis nodes *H*, observation nodes *O*, and intermediate nodes *I*. Fig. 2 illustrates such a Bayesian network and how it is embedded in a level *n* of the hierarchy.

A Bayesian network at a given level of a predictive coding hierarchy is sometimes also referred to as a *generative model*, as it can be seen as representing the cognitive agent's causal model of the world. In hierarchical predictive coding, three separate processes are assumed to operate on this generative model; viz., *prediction* (computing the observations that are predicted at level *n*, given the causal model at level *n* and the predictions at level $n + 1$ in the backward chain), *error computation* (computing the deviation between the predicted observations at level *n* in the backward chain and the actual observations at level *n* in the forward chain); and *hypothesis updating* (updating hypotheses at level *n* of the forward chain based on the prediction error between predicted and actual observations at level *n*).[1]

Although several specific computational models, that apply the ideas underlying the predictive coding framework to the analysis of particular perceptual and neural processes, have been developed (e.g., Grush, 2004; Jehee & Ballard, 2009; Rao & Ballard, 1999), such specific models cannot directly be used to address our research question: 'Is Bayesian inference tractable when implemented in a

[1] For completeness, we note that a fourth process can operate in predictive coding, viz., one that updates the parameters of the causal model, based on observations and prediction errors, to allow for better predictions in the future. We will not analyze this (long term) learning aspect in predictive coding here, although likely it is itself computationally no easier than the three subprocesses that we do analyze here (Chickering, Heckerman, & Meek, 2004).

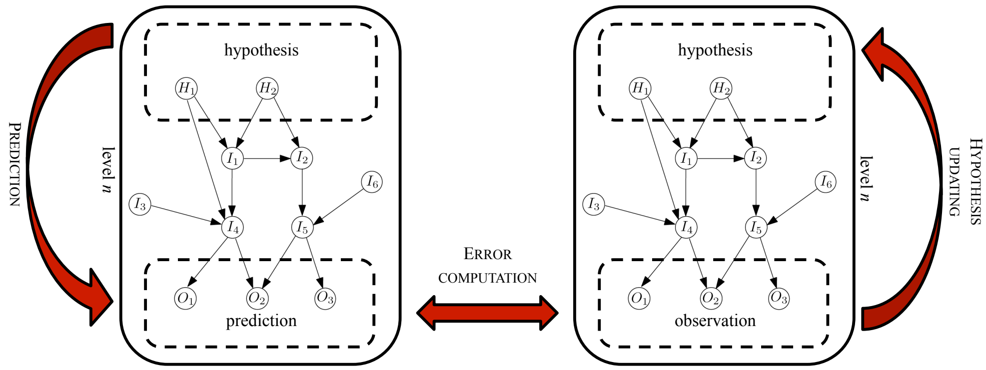

Fig 2. An illustration of a Bayesian network situated at a particular level *n* in the predictive coding hierarchy. The Bayesian network consists of nodes (circles) and arcs (arrows). Nodes represent the variables in the network and arcs represent assumed probabilistic dependencies between variables. The Bayesian network also has an associated conditional probability distribution (not shown here). Nodes partition into hypothesis nodes (*H*), observation nodes (*O*), and intermediate nodes (*I*). In this figure, intermediate variables are situated in between hypothesis and observation nodes within a given level, but this is not necessary for our (in)tractability results to hold; our (in)tractability results hold also if the intermediate nodes are situated outside the level and only connect to the observation nodes.

hierarchal predictive coding architecture?' To address this question, we need instead *generic* computational models, i.e., models that are general enough to be applicable, in principle, to any cognitive domain. We propose such generic models in the next section.

Computational modeling

The hierarchy of coupled subcomputations that we sketched above and in Fig. 1 and 2 can be seen as a candidate algorithmic level explanations of *how* perceptual and cognitive inferences—often well-captured by Bayesian models situated at Marr's (1982) computational level and presumably living higher up in such hierarchies—could be computed by human brains. For such an algorithm to be tractable all of its subcomputations need to be tractable. Here we present models of the three subcomputations prediction, error computation and hypothesis updating also situated at Marr's computational level. That is, the models merely characterize the nature of the transformation the subcomputations achieve while not committing to any further hypotheses of how these subcomputations are again further subdivided in subsubcomputations.

Before presenting our models, we note that it is difficult to settle on a single candidate computational-level model per subcomputation, because several different—sometimes mutually inconsistent—interpretations of the predictive coding framework can be found in the literature (e.g., Friston, 2002; Friston, 2005; Hohwy, Roepstorff, and Friston, 2008; Kilner, Friston, & Frith, 2007; Lee & Mumford, 2003; see also Knill & Pouget, 2004). For this reason we have developed various variants of the computational-level models of the postulated subprocesses: two for prediction, two for error computation and four for hypothesis updating. We believe that with our eight models we comprehensively cover the most common interpretations in the literature.

Table 1 gives an overview of the 8 models (see the online supplementary materials for more formal details).[2] In the literature, two different conceptualizations of 'prediction' and 'hypothesis updating' can be distinguished, based on whether one views the inference steps as *updating* ones probability distribution over beliefs (Lee & Mumford, 2003, p. 1437), or as *fixating* ones belief to the most probable ones (Kilner et al., 2007, p. 161). These two notions correspond to two different computational problems in Bayesian networks, viz., the problem of computing a posterior probability distribution and the problem of computing the mode of a posterior probability distribution (i.e., the joint value assignment with the highest posterior probability). Following conventions in the machine learning literature, we apply the suffix MAX to the belief fixation model variants, and SUM to the distribution updating variants. For both mathematical and conceptual consistency, we will assume that the subcomputations operational in a given predictive coding hierarchy are either all of the type MAX, or all of the type SUM, rather than a mixture of the two variants (see Fig. 3 for a graphical illustration).

Furthermore, there are two possible conceptualizations of the process of hypothesis updating while using prediction error. One conceptualization (the PRED variant) seeks to minimize prediction error *per se* by replacing the original hypothesis by one that has a higher likelihood given the evidence (Kilner et al., 2007, p. 161), while the other interpretation (the EXPL variant) replaces the original hypothesis with one that has higher posterior probability given the evidence, thus taking both likelihood and prior probability of the hypotheses into account (Friston, 2002, p. 13).

In the PREDICTION problem, a prediction is made based on the causal model and the current hypothesis. This is

[2] Supplementary materials are available online at: http://www.dcc.ru.nl/~johank/material/cogsci13/supplement.pdf

formalized as either (re-)computing the posterior distribution over the observation nodes (SUM) or computing the mode of these nodes (MAX). In ERROR-COMPUTATION, the prediction is compared with the actual observation. For the SUM variant, this is formalized by computing the Kullback-Leibler divergence (Kullback, 1951) between the predicted probability distribution and the observed (at the lowest level) or inferred (at higher levels) probability distribution. For the MAX variant, this is formalized by computing the Hamming distance (Hamming, 1950) between the predicted and observed (or inferred) joint value assignment. Lastly, in HYPOTHESIS-UPDATING, the initial hypothesis—either a distribution over (SUM), or a joint value assignment to (MAX), the hypothesis variables—is updated using the computed error, either to maximize the likelihood (PRED) or the posterior probability (EXPL) of the updated hypothesis, again being either a distribution or a joint value assignment.

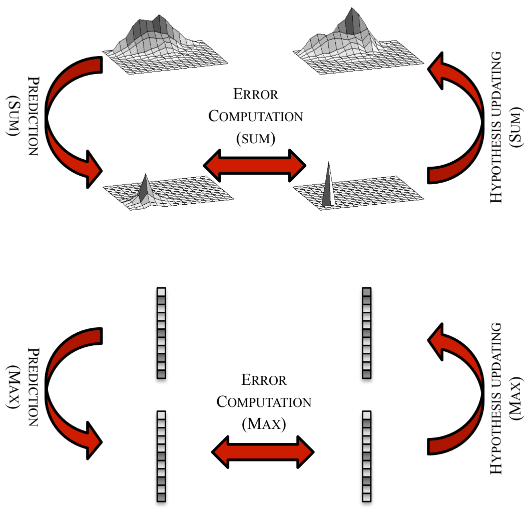

Fig 3. A schematic illustration of the difference between the SUM variants (at the top) and MAX variants (at the bottom) of the computational models of PREDICTION, ERROR COMPUTATION, and HYPOTHESIS UPDATING. See also Table 1.

Intractability results

For our (in)tractability analyses we build on concepts and techniques from computational complexity theory (Arora, & Barak, 2009; Garey & Johnson, 1979). In line with the literature (Gigerenzer, 2008; Kwisthout et al., 2011), we will adopt NP-hardness as a formalization (of at least one relevant interpretation) of the notion of 'intractability'.

A computation that is NP-hard cannot be computed in so-called polynomial time, i.e., a time taking on the order of n^a steps, where n is a measure of the input size (e.g., n may the number of the nodes in the Bayesian network) and a is a (small) constant.[3] In other words, all algorithms performing an NP-hard computation require super-polynomial (e.g., exponential) time, at least for some of the inputs. To illustrate why super-polynomial time is considered intractable for all but very small inputs, consider that, for instance, an exponential running time on the order of 2^n already consumes more than a million basic steps for $n = 20$, and for $n = 35$ the number of steps exceeds the number of seconds in a millennium.

Using proof techniques from computational complexity theory, we have derived the following intractability results, which hold regardless of whether the computation is to be performed exactly or approximately (for full proofs and technical details, refer to the online supplementary materials; see footnote 1):

- PREDICTION (MAX/SUM) is NP-hard, even if there is only one hypothesis and one observation variable.
- ERROR COMPUTATION (SUM) is NP-hard, even if there is only one hypothesis and one observation variable.
- ERROR COMPUTATION (MAX) is polynomial-time computable.
- HYPOTHESIS UPDATING (MAX/SUM, PRED) is NP-hard, even if there is only one hypothesis and one observation variable and the prediction error is arbitrarily small.
- HYPOTHESIS UPDATING (MAX/SUM, EXPL) is NP-hard, even if there is only one hypothesis and one observation variable and there is no prediction error at all.

Some important observations and implications can be drawn from our results: Without constraints on the network, both PREDICTION and HYPOTHESIS UPDATING are intractable (NP-hard) for either the SUM or the MAX interpretation, and for either the PRED and EXPL interpretation. These results hold regardless the tractability of the MAX interpretation of ERROR COMPUTATION. These findings make clear that a predictive coding implementation of Bayesian inference does not yet make the latter tractable; i.e., the subcomputations PREDICTION and HYPOTHESIS-UPDATING are intractable. This holds even under stringent additional assumptions, such as that each level in the hierarchy contains a causal model with at most one hypothesis variable and at most one observation variable, and even when the prediction error is arbitrarily small (for the PRED version of HYPOTHESIS UPDATING) or completely absent (for the EXPL version of HYPOTHESIS UPDATING). The latter brings the—possibly counterintuitive—point home that a brain in possession of a hierarchy of causal models that accurately predict sensory inputs is not yet a brain that can easily update its beliefs about states of the world. Perhaps similarly counterintuitive is our finding that even ERROR COMPUTATION itself can be intractable, at least under

[3] This statement assumes that the famous P ≠ NP conjecture is true (Garey & Johnson, 1979). This conjecture is widely believed to be true, both on theoretical and empirical grounds (for an accessible explanation see also Scott Aaronson's blog 'Reason to believe': http://scottaaronson.com/blog/?p=122).

Table 1
Definitions of the model variants

Problem	Variant	Input	Output
PREDICTION	SUM	Bayesian network B_L, prior distribution Pr_H over the hypothesis nodes H	The posterior distribution Pr_{Op} over the prediction nodes O
	MAX	Bayesian network B_L, joint value assignment h to the hypothesis nodes H	A joint value assignment o_p to the prediction nodes O with a maximum posterior probability $Pr(o_p \mid h)$
ERROR COMPUTATION	SUM	Distribution Pr_{Op} over the prediction nodes O, observation Pr_{Oa}	$D_{KL}(Pr_{Op}, Pr_{Oa})$, i.e., the Kullback-Leibler divergence between Pr_{Op} and Pr_{Oa}
	MAX	Joint value assignment o_p to the prediction nodes O, observation o_a	$D_H(o_p, o_a)$, i.e., the Hamming distance between o_p and o_a
HYPOTHESIS UPDATING	SUM, PRED	Bayesian network B_L, candidate prior distribution Pr_{Hc} over the hypothesis nodes H, prediction error $D_{KL}(Pr_{Op}, Pr_{Oa})$	Updated distribution Pr_{Hu} such that $D_{KL}(\text{PREDICTION}(Pr_{Hu}), Pr_{Oa}) < D_{KL}(Pr_{Op}, Pr_{Oa})$
	SUM, EXPL	Bayesian network B_L, candidate prior distribution Pr_{Hc} over the hypothesis nodes H, prediction error $D_{KL}(Pr_{Op}, Pr_{Oa})$	The posterior distribution Pr_{Hu} over the hypothesis nodes H
	MAX, PRED	Bayesian network B_L, candidate joint value assignment h_c to the hypothesis nodes H, prediction error $D_H(o_p, o_a)$	A joint value assignment h_u such that $D_H(\text{PREDICTION}(h_u), o_a) < D_H(o_p, o_a)$
	MAX, EXPL	Bayesian network B_L, candidate joint value assignment h_c to the hypothesis nodes H, prediction error $D_H(o_p, o_a)$	A joint value assignment h_u to the hypothesis nodes H with a maximum posterior probability $Pr(h_u \mid o_a)$

the SUM interpretation (an interpretation some believe holds for human brains; e.g., Lee & Mumford, 2003).

Tractability results

The previous section reports intractability (NP-hardness) results for unconstrained causal models in hierarchical predictive coding. It is known, however, that a computation that is NP-hard on unconstrained inputs can sometimes be computed in a time that is non-polynomial only in some (possibly small) parameter of the input and polynomial in the size of the input. More formally, given an NP-hard computation Q, with input parameters $k_1, k_2, \ldots, k_m$, there may exist a procedure for computing Q that runs in time $f(k_1, k_2, \ldots, k_m)n^a$, where f is a function and a is a constant independent of the input size n. A fixed-parameter (fp-) tractable procedure can perform an otherwise intractable computation quite efficiently, even for large inputs, provided only that the parameters $k_1, k_2, \ldots, k_m$ are small (e.g., for $k = 3$ and $n = 35$ an fp-tractable running time on the order of $2^k n = 280$ compares favorable with an exponential running time $2^n > 10^{10}$). Here, we investigate whether or not such fp-tractable procedures also exist for the subcomputations in the hierarchical predictive coding framework.

In our analyses (full formal details appear in the supplementary materials; see footnote 2), we consider the following parameters:

- the maximum number of values each variable in the Bayesian network can take (c);
- the treewidth of the network (t), a graph-theoretical property that can loosely be described as a measure on the 'localness' of connections in the network (Bodlaender, 1993);
- the size of the observation space ($|O|$) or hypothesis space ($|H|$); and
- the probability of the most probable prediction or most probable hypothesis ($1 - p$).

Our main findings are that each of the subcomputations PREDICTION, ERROR COMPUTATION, and HYPOTHESIS UPDATING can be performed tractably when the topological structure of the Bayesian network is constrained (small t), *and* each variable can take a small number of distinct values (small c), *and* the search space of possible predictions and hypotheses is small (small $|O|$ and small $|H|$). Specifically, both SUM and MAX variants are computable in fp-tractable time $O(c^{|O|} \cdot c^t \cdot n)$ for PREDICTION and ERROR COMPUTATION, and $O(c^{|H|} \cdot c^t \cdot n)$ for HYPOTHESIS UPDATING.[4] In addition, the MAX-variants are also tractable when the prediction or hypothesis space may be large, but instead the most probable prediction (hypothesis) has a high probability (i.e., $1 - p$ is small). Specifically, PREDICTION and HYPOTHESIS UPDATING are computable in fp-tractable time $O(c^{(\log(p) / \log(1-p))} \cdot c^t \cdot n)$ and, as we had shown in the previous section, ERROR COMPUTATION can be computed in polynomial time.

Conclusion

To assess the purported tractability of a hierarchical predictive coding implementation of Bayesian inference we formulated explicit computational-level characterizations of

[4] Here the expression O(.) denotes the standard *big-Oh* notation for order of magnitude (see e.g. Garey & Johnson, 1979).

the basic subcomputations that the predictive coding framework postulates. Computational complexity analyses of these subcomputations reveal that they themselves are intractable, unless the causal models that are coded by the hierarchy are somehow constrained.

To be clear, we do not wish to suggest that the brain cannot tractably implement Bayesian inferences, nor do we suggest that the brain cannot do so using a predictive coding hierarchy. In fact, we think that the hypothesis that the brain implements tractable Bayesian inferences, possibly using predictive coding, is not implausible. What our results show, however, is that *if* the brain indeed tractably implements Bayesian inference, then that is not *because* the inferences are implemented using predictive coding *per se*. Furthermore, our results suggest that the key to tractable inference in the brain may rather lie in the properties of the causal models that it represents at the different levels of such hierarchy. Some of these properties are topological in nature (i.e., the structure of the causal models), whereas others pertain to the number of competing hypotheses that the brain considers per level of the hierarchy. In light of our findings, we propose that an important topic of empirical investigation in cognitive neuroscience may be whether or not the brain's causal models have the properties that are necessary for tractable Bayesian inference.

Acknowledgements

We thank Mark Blokpoel for discussions that inspired this research and for comments on earlier versions of this paper.

References

Arora, S., & Barak, B. (2009). *Computational complexity: A modern approach.* Cambridge: Cambridge University Press.

Blokpoel, M., Kwisthout, J., & van Rooij, I. (2012). When can predictive brains be truly Bayesian? *Frontiers in Theoretical and Philosophical Psychology*, 3 (406), 1–3.

Bodlaender, H. (1993). A tourist guide through treewidth. *Acta Cybernetica, 11*, 1–21.

Brown, H., Friston, K., & Bestmann, S. (2011). Active inference, attention, and motor preparation. *Frontiers in Psychology, 2*(218), 1–10.

Chickering, D., Heckerman, D., & Meek, C. (2004). Large-sample learning of Bayesian networks is NP-Hard. *Journal of Machine Learning Research, 5*, 1287–1330.

Clark, A. (in press). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*.

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

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

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B, 350*, 815–836.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences, 13*, 293–301.

Friston, K. (2012). *Policies and priors*. In Gutkin, B., & Ahmed, S.H. (Eds.), Computational neuroscience of drug addiction (pp. 237–283). New York: Springer.

Garey, M., & Johnson, D. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. San Francisco, CA: W.H. Freeman and Co.

Griffiths, T., Kemp, C., & Tenenbaum, J. (2008). Bayesian models of cognition. In Ron Sun (Ed.), *Cambridge handbook of computational cognitive modeling.* Cambridge: Cambridge University Press.

Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences, 27*, 377–442.

Hamming, R. (1950). Error detecting and error correcting codes. *Bell System Technical Journal, 29(2)*, 147–160.

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition, 108*(3), 687–701.

Jehee J., & Ballard D. (2009). Predictive feedback can account for biphasic responses in the lateral geniculate nucleus. *PLoS Computational Biology, 5*, 1–10.

Kilner J., Friston K., & Frith C. (2007). Predictive coding: an account of the mirror neuron system. *Cognitive Processes*, *8*, 159–166.

Knill, D. & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neuroscience, 27*(12), 712–719 .

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*, 79–86.

Kwisthout, J. (2011). Most probable explanations in Bayesian networks: Complexity and tractability. *International Journal of Approximate Reasoning*, *52,* 1452–1469.

Kwisthout, J., & van Rooij, I. (in press). Bridging the gap between theory and practice of approximate Bayesian inference. *Cognitive Systems Research.*

Kwisthout, J., Wareham, T., & van Rooij, I. (2011). Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science, 35*, 779–784.

Lee, T.S. & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America, 20*(7), 1434–1448.

Rao, R., & Ballard, D. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*, 79–87.

Summerfield, C., Trittschuw, E., Monti, J., Mesulam, M.-M., & Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. *Nature Neuroscience, 11*(9), 1004–1006.

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