What can the PGM community contribute to the 'Bayesian Brain' hypothesis?

Johan Kwisthout[†ab]

[a] Department of Artificial Intelligence
[b] Donders Institute for Brain, Cognition and Behaviour
Radboud University Nijmegen
PO Box 9104, 6500HE Nijmegen, The Netherlands

Abstract

Despite the now common view amongst neuroscientists that the brain effectively approximates Bayesian inferences, there are only few researchers in the Probabilistic Graphical Models (PGM) community currently working in this research area. We believe that this is partially due to a misunderstanding of the theoretical challenges that theoretical neuroscience currently faces and the potential contribution that the PGM community can offer in interdisciplinary research. With this paper we hope to remedy such misunderstandings and invite the community to contribute to the mutual benefit of neuroscience and AI alike.

1 Introduction

When discussing recent advances in neuroscience–that postulate that the human brain is at its essence just a Bayesian inferential machine–with scholars in the Probabilistic Graphical Models (PGM) community, our research group occasionally receives lukewarm responses that can best be paraphrased as "I'm just not interested in the brain as an application area of my research". Although there are few things as personal as a research agenda, we still feel that this lack of interest may be at least partially due to a) a misconception of the questions that are currently being addressed in neuroscience and b) lacking some 'insider's insight' in the contribution that the PGM community can offer in interdisciplinary research. With this paper we hope to remedy both. We will give a short overview of the increasingly popular 'Bayesian Brain' hypothesis in neuroscience, in particular its 'predictive processing' manifestation. We will then identify three research areas within this topic where contributions from the PGM community can actually have a huge scientific impact. After identifying potential pitfalls in such interdisciplinary research, including a discussion of the specific (and sometime peculiar) connotations of the neuroscience community with respect to concepts like 'Bayesian', 'uncertainty', and 'prior', we will conclude with an invitation to the community to contribute.

2 The Brain as 'Application Area'

Herman von Helmholtz [39] is traditionally seen as the originator of the view of human perception as (statistical) inference to the best explanation of the causes of the perceptual input. The suggestion that the human brain can be seen as performing some approximate Bayesian inference (integrating prior expectations with newly arriving information) was coined as early as 1957 by Edwin T. Jaynes (first published in [17]). Peter Dayan and colleagues further explored these ideas and proposed the notion of the *Bayesian Brain*, emphasizing on the basis of psychophysical evidence that human perception

† j.kwisthout@donders.ru.nl.

actually is 'Bayes optimal' in combining priors and new signals. The *Bayesian coding* hypothesis postulates that the brain indeed encodes probability distributions in populations of neurons.

In recent years, the *Bayesian Brain* hypothesis has become increasingly popular due to the emergence of Karl Friston's *free energy principle*, providing for a biological and physical foundation; the *predictive processing* view of the brain as a 'prediction machine' that minimizes computational effort by trying to predict its inputs, and the *spiking neural network* research area that shows that probability distributions can be encoded and sampled from using power-efficient networks of spiking neurons. We will elaborate more on these three important recent developments.

2.1 The free energy principle

Friston's *free energy principle* [9, 10] postulates that any biological system that 'resists a tendency to disorder' – be it a single cell or a social network – effectively aims to minimize free energy. In thermodynamics, free energy is the amount of energy that is potentially available, but not put to effective use. In information theory, it is a measure on the discrepancy between our observation of the world and our model of the world, which becomes manifest as the *prediction error* between predicted and observed world state. A biological system that aims to defy disorder seeks to lower entropy (the average of surprise of outcomes). It can do so by minimizing prediction error, that is, aiming to make the predicted world state match the observed world state (adapting one's models of the world), or vice versa (changing one's sensory input by acting upon the world). Because biological systems must remain within certain boundaries to exist, their models of what the world should look like (e.g., have access to a sufficient, but not excess, amount of oxygen to maintain homeostasis) and how they currently perceive the world (e.g., shortage of oxygen) should match, and if not, actions are taken to minimize this prediction error (e.g., breathe faster and deeper). Friston [9, p.295] summarizes this by postulating that *(i) agents resist a natural tendency to disorder by minimizing a free-energy bound on surprise; (ii) this entails acting on the environment to avoid surprises, which (iii) rests on making Bayesian inferences about the world.*

2.2 Predictive processing

The Predictive Processing account proposes that the brain continuously predicts its inputs in a hierarchical cascade of (increasingly more concrete) probabilistic predictions [4, 5, 16]. For example, when observing a bowler on a bowling lane, contextual information ("this bowler already hit three strikes in this game") will generate predictions for the result of the throw ("many pins will fall down"). Based on that expectation, more specific predictions will be made for the throwing kinematics, the ball trajectory, where the ball will hit the pins, etc. Violations of predictions will yield prediction errors that need to be 'explained away' by updating ones hypotheses ("even good bowlers will sometimes fail to throw a strike"), taking new contextual information into consideration ("the bowler seems to have injured his wrist whilst throwing") etc. The computations 'under the hood' of this conceptual description can be described and analyzed as various computations on causal Bayesian networks, such as the computation of posterior probability distributions and the tuning of parameters of the network [24]. The rationale behind this account is that processing only the prediction error is less computationally demanding as processing the entire input; however for exact computations, it was shown that this assumption does not hold in general, since processing even a single bit of prediction error is an NP-hard problem [21]; whether *approximate* Bayesian inference is tractable when the prediction error is low is currently an open problem. Despite its popularity as a unifying theory, it is far from clear what the brain's approximation algorithms actually look like; in Clark's [4, p.201] words: *What do the local approximations to Bayesian reasoning look like as we depart further and further from the safe shores of basic perception and motor control? What new forms of representation are then required, and how do they behave in the context of the hierarchical predictive coding regime?*

2.3 Networks of spiking neurons

One of the most promising computational models of neuronal computation in general is the recurrent *network of spiking neurons* model [28]. These biologically inspired networks mimic Boltzmann machines (neural networks that represent a probability distribution that can be sampled from), with a key difference that the neurons are not outputting a zero or one state, but a *spike*; a brief burst of

energy. These networks are energy-efficient and stochastic in nature and they can represent, and reason with, arbitrary probability distributions by means of stochastic sampling in winner-take-all microcircuits [2, 13, 33]. It has been proposed that such sampling methods (like MCMC sampling) are the most promising techniques to describe actual stochastic inferences in the brain [36]. Because of their efficiency – the brain uses a mere 25W of energy – these networks are potentially crucial for future generations of computer hardware by utilizing (rather than trying to filter) the noise that is inherent at the nano-scale [14]. No free lunch is offered, though: As approximate Bayesian inference is an intractable problem [7, 23], there will be problem instances where the convergence time of the network will grow exponentially with the input size, in particular in networks with extreme probability distributions [28].

In terms of Marr's levels of explanation [29], one can see the free energy principle as aiming to answer the 'why' of the Bayesian Brain hypothesis, the predictive processing account describes 'what' is actually being computed, whereas the 'spiking neurons' community studies the 'how' aspect of approximate Bayesian computations in the brain. Where the free energy/predictive processing and the networks of spiking neurons communities were traditionally relatively isolated – as a proxy, one could see them as exponents of the *UK*, respectively *Continental* approach towards theoretical neuroscience – there have been recent mutual research events (for example at the European Institute for Theoretical Neuroscience in Paris) that try to bridge the gap between both communities.

2.4 Organization of this paper

All these developments support the 'Bayesian' view of the brain as it is currently dominant in contemporary neuroscience. We believe that this opens up a significant area of research for the PGM community. Where graphical models are currently used in neuroscience, their role is typically limited to association, clustering, or classification of brain data [1], i.e., as a data-analysis tool rather than as a process-level description of the brain's mechanisms for information processing. Yet, the emergence of the Bayesian brain hypotheses opens up a whole new area of research. In the remainder of this paper we will further elaborate on this. We will show how a formal and computational background can help to bring conceptual clarity and formal rigidity to the field; how neuroscience is in urgent need for new algorithms, implementations, and complexity analyses that computer scientists and AI practitioners can provide, and where new questions in the 'meta'-theory of learning and modifying Bayesian networks emerge.

3 Conceptual Clarity and Rigidity

An important area where researchers with a strong background in computational and formal modeling can make vital contributions is in offering conceptual clarity and formal rigidity, translating verbal theories into complete and consistent computational models, thus exposing ambiguities and gaps in the theory and explicating 'design choices' and their computational consequences [31]. Examples are in the formal explication of the role and nature of the underlying principles of predictive processing [20,34,37], critically assessing the validity of simplifying assumptions [30], and in exposing the consequences of alternative readings of vague or conflicting verbal models [25]. On top of this, the specific background of researchers in the PGM community can contribute significantly to the theory itself, generating new theoretical and empirical questions. The following case study will further exemplify this.

In the predictive processing theory, stochastic predictions are compared with actual observations and only the residual (non-predicted) signal is processed by prediction error minimization. This prediction error, however, is dependent on the state space of the prediction and its granularity; for example, when we predict and observe a (non-specified) tree, or when we predict to see an oak and see a chestnut tree. This observation – made from an information-theoretic point of view – led to a further refinement of the predictive processing account with the notion of *levels of detail* of models and predictions (Figure 1), and spawned various research projects. One particular empirical result that is based on these insights is the development of a predictive processing account of how psychedelics effect the brain's information processing [35]. Here it was proposed that psychedelics such as psilocybin hyper-activate $5HT_{2A}$ receptors in layer-5 pyramid cells, effectively leading to over-detailed, diffuse predictions that cause many of the sympoms associated with psilocybin administration.

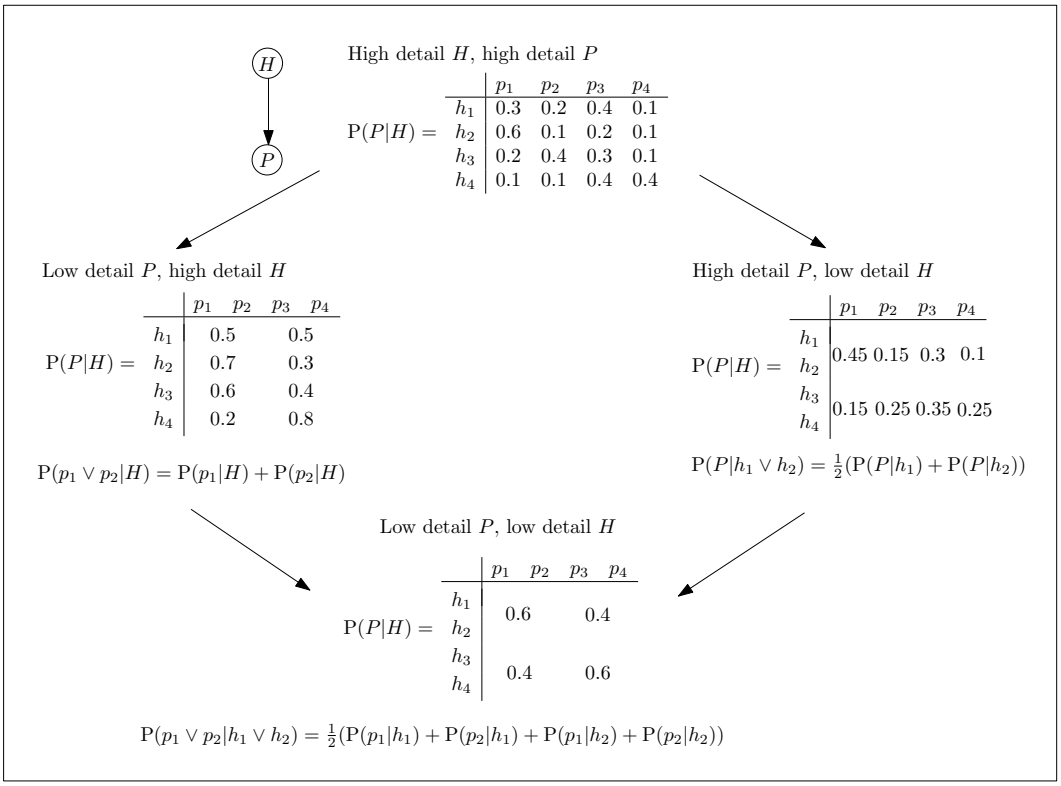

Figure 1: A formalization of the relationship between different levels of detail of hypotheses and predictions. In this example we have a singleton hypothesis and a singleton prediction, each with vour different values. Observe that the actual hypotheses, as well as the predictions, can be *clustered*, redefining the conditional probability distributions in a straightforward way. We can thus lower the detail of the predictions (leftmost CPT), lower the detail of the hypotheses (rightmost CPT), or both (bottom CPT).

4 Theory, Algorithms, and Analysis

The Bayesian Brain hypothesis stipulates that the brain approximates probabilistic inference by means of variational Bayes methods [12] or sampling approaches [36]. As exact inference is PP-complete [26] there cannot exist efficient approximation algorithms in general, unless BPP equals PP. That means that there must be constraints on the inputs in order to render this approximate computation tractable; the field of *parameterized complexity* studies such constraints. Recent developments in this area allow for the analysis of stochastic computations where the probability of answering incorrectly is parameterized, rather than the computation time [22, 23]. This allows for the study of so-called *fixed error randomized tractable* approximations, relative to 'ecologically valid' parameters, viz. parameters that can plausibly be assumed to be small in the computations as performed by the brain. In particular the parameterized complexity of approximate inference, as parameterized by the prediction error of the generative model, is an important open problem that would significantly contribute to the Bayesian Brain hypothesis in general and predictive processing in particular: It is claimed [4] that the brain can be efficient *because* it tries to minimized prediction error and thus that inference can be tractable when prediction error is low.

Apart from process-level considerations (under what constraints can the approximations postulated by predictive processing be tractable), one can study the properties and plausibility of neuronal implementations of such approximations using networks of spiking neurons. Crucial properties here are the power efficiency of such networks [28], the nature of the *noise* in the brain and its consequences for efficient sampling [13], and the general question how many resources are needed for effective computations [27]. Computational complexity theory offers an indication of the resources needed for a particular computational problem to be solved, as a function of the input size of a problem. These resources most

notably, time and memory are typically fairly coarse and built on a theoretical abstract model of computation: Turing machines. Here, the 'time' resource refers to the number of state transitions in the machine, and the 'memory' resource refers to the number of memory cells on the tape that are used. It has been proposed by a working group at the Dagstuhl seminar on Resource-Bounded Problem Solving (seminar 14341) to have a more refined, brain-focused model of computation in the brain, based on networks of spiking neurons, and have complexity measures based on brain resources, such as spiking rates, network size, and connectivity [15]. The development of such a model of computation would allow for seminal contributions to the Bayesian Brain hypothesis by analyzing the fundamental limits of brain computations.

5 Meta-theory of Bayesian Networks

When learning a Bayesian network from data one might reconstruct the structure of the network, the probability distributions, and even the distributions over hidden variables. Crucially, though, one needs to settle beforehand on the variables and their state space. This is to be contrasted with how generative models in the Bayesian brain hypothesis are actually constructed: Here, one somehow needs to 'learn' new variables and the values they can take, both for potential causes and their observable manifestations. The question then arises *when* a Bayesian learner realizes that the current model is insufficient and new hypotheses should be formed, as well as *what* these hypotheses should look like [3]. This problem comes on top of 'normal' model revision by Jeffrey updating [19], where just the probability distributions are updated in the light of new evidence; this aspect of model revision can be elegantly related to predictive processing concepts such as precision-weighted prediction error [32]. The problem of adding new variables and values of variables to a network in the light of unresolvable prediction error is a major open problem.

When a prediction error is to be accounted for, one can either update ones current beliefs about the actual hypotheses or try to reduce uncertainty by observing hidden variables. These predictive processing sub-processes (belief revision and adding observations) correspond to aspects of parameter tuning and sensitivity analysis [6] and selecting evidence [38]. Algorithmic and analytical aspects of these problems are of direct relevance to the Bayesian Brain hypothesis.

A vital open problem in the predictive processing account relates to the trade-off between making predictions that are very *detailed* and predictions that are likely to be *correct*. For example, when predicting the outcome of a throw at a bowling lane, a prediction over a distribution containing values like 'pin four will be hit by the ball from the left side and will topple over pins seven and eight' is very detailed, but probably always gives a huge prediction error. On the other hand, a prediction like 'the ball will hit the pins and some will fall' is likely to be correct, but as a prediction not very informative. There are reasons to believe that particular neurotransmitters (in particular serotonin) control this *level of detail* [35], but from a more meta-perspective it is completely open how causal Bayesian models can be 'flexible' in their granularity and how algorithms on such models may trade-off information gain and prediction error.

6 Potential Pitfalls

In the previous sections we highlighted several research areas and tentative research questions where the PGM community can substantially contribute to the 'Bayesian Brain' with a potential for considerable impact. Notwithstanding this potential, there are also pitfalls to avoid that are inherent risks of interdisciplinary work, in particular when the research fields have different cultures and tradition and use specific terminology that may be misunderstood. Here we enumerate a few potential pitfalls.

- '**Terminology**' — An informal quiz at the interdisciplinary Lorentz Center workshop 'Perspectives on Human Probabilistic Inference'[2] on the association that participants had with the word 'Bayesian' was illuminative to us. For some participants *Bayesian* was a synonym of *probabilistic*, for others it concerned the semantics of probability distributions (*subjective*, as contrasted with *frequentist*), yet others associated *Bayesian* with *Bayes' rule* for updating distributions. Despite the traditional interpretation of 'Bayesian' as 'subjective degrees of belief' [18], it is not

[2] http://www.lorentzcenter.nl/lc/web/2014/627/info.php3?wsid=627&venue=Oort

uncommon for proponents of the Bayesian Brain hypothesis to have a strong frequentist view on probabilities as describing the objective state of the world [8]. Similarly diverse (and sometimes counterintuitive) associations could be elicited for terms like 'prior', 'uncertainty', 'information', and 'structure'. The bottom line is to be aware of potential misunderstandings and to be explicit of one's intended meaning of such terms in communication with neuroscientists.

- '**Culture and tradition**' — In computer science and artificial intelligence, acceptance of a paper to a prestigious conference such as AAAI, UAI, FOCS or STOC is distinctive. Many scholars focus their publication strategy on such conferences, rather than journal papers. In neuroscience, a conference publication is close to irrelevant when it comes to evaluating research output; much more emphasis is put on the impact factor of the journals one is publishing in. Culture and tradition put emphasis on different 'golden standards' of excellence in research, validity of research methodology, and importance of research topics. Awareness of such issues and an open mind may help avoid or solve misunderstandings.

- '**Interdisciplinary**' — Members of interdisciplinary teams have different backgrounds and distinct areas of expertise; that is exactly the main benefit of having interdisciplinary collaborations at all. There is a fine line between 'nitpicking on details' versus 'allowing crucial misconceptions to exist' in interdisciplinary collaborations, and it requires some expertise to see what is important and what not. For example, it is rarely important to insist on the distinction between NP-hardness and NP-completeness of a problem, but the difference between an observation and an intervention in (causal) Bayesian networks may well be important to clarify. Don't assume your neuroscience collaborators share your background, and don't be afraid to ask for clarification about what seems obvious to them.

- '**Selling your work**' — An elegant intractability proof or a new formalization of a verbal theory is typically not sufficient for publication in neuroscience outlets. In order to get published one should aim to understand the problems that neuroscientists care about, make clear why your contribution is instrumental in solving these problems, and write in a way that connects to their background and expectations. It might be difficult to convince one's departmental chair or (grant) reviewers of the relevance of this work. Our approach is to seek for niches that both allow for a significant PGM contribution *and* solve crucial problems with respect to the Bayesian Brain.

7 Conclusion

Despite the potential pitfalls we identified in the previous section, we strongly believe computer scientists and AI practitioners working in the PGM area can make a vital interdisciplinary contribution to contemporary theoretical neuroscience. With this paper we hope to have given an overview of crucial open problems in the Bayesian Brain hypothesis and a sketch of the contributions that the PGM community can offer. We conclude this paper with this quote from Karl Friston [11] that (probably inadvertently) illustrates the importance of research on probabilistic graphical models for theoretical neuroscience: *Life (. . .) is an inevitable and emergent property of any (ergodic) random dynamical system that possesses a Markov blanket.* We would like to invite the community to bring their toolbox of computational and formal modeling and help to advance this fascinating research area — who knows what else may emerge!

References

[1] C. Bielza and P. Larrañaga. Bayesian networks in neuroscience: A survey. *Frontiers in Computational Neuroscience*, 8:Article 131, 2014.

[2] L. Buesing, J. Bill, B. Nessler, and W. Maass. Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology*, 7(11):e1002211, 2011.

[3] C. D. Carroll and C. Kemp. Hypothesis space checking in intuitive reasoning. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, 2013.

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

[5] A. Clark. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press, 2015.

[6] V. M. H. Coupé, F. V. Jensen, U. B. Kjærulff, and L. C. van der Gaag. A computational architecture for n-way sensitivity analysis of Bayesian networks. Technical report, Aalborg University, 2000.

[7] P. Dagum and M. Luby. Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence*, 60(1):141–153, 1993.

[8] C.D. Fiorillo. Beyond Bayes: On the need for a unified and Jaynesian definition of probability and information within neuroscience. *Information 2012, 3(2)*, 3(2):175–203, 2012.

[9] K.J. Friston. The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, 13(7):293–301, 2009.

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

[11] K.J. Friston. Life as we know it. *Journal of the Royal Society Interface*, 10(86):20130475., 2013.

[12] K.J. Friston and K. E. Stephan. Free-energy and the brain. *Synthese*, 159:417–458, 2007.

[13] S. Habenschuss, Z. Jonke, and W. Maass. Stochastic computations in cortical microcircuit models. *PLoS Computational Biology*, 9(11):e1003037, 2013.

[14] T.J. Hamilton, S. Afshar, A. van Schaik, and J. Tapson. Stochastic electronics: A neuro-inspired design paradigm for integrated circuits. *Proceedings of the IEEE*, 5:843–859, 2014.

[15] Y. Haxhimusa, I. van Rooij, S. Varma, and H. T. Wareham. Resource-bounded problem solving (dagstuhl seminar 14341). *Dagstuhl Reports*, 4(8), 2014.

[16] J. Hohwy. *The Predictive Mind*. Oxford University Press, 2013.

[17] E.T. Jaynes. How does the brain do plausible reasoning? In G. J. Erickson and C. R. Smith, editors, *Maximum-Entropy and Bayesian Methods in Science and Engineering*, 1988.

[18] E.T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[19] R.C. Jeffrey. *The Logic of Decision*. University of Chicago Press, 1965.

[20] J. W. Kay and W. A. Phillips. Coherent infomax as a computational goal for neural systems. *Bulletin of Mathematical Biology*, 73(2):344–372, 2011.

[21] J. Kwisthout. Minimizing relative entropy in hierarchical predictive coding. In L.C. van der Gaag and A.J. Feelders, editors, *Proceedings of PGM'14*, LNCS 8754, pages 254–270, 2014.

[22] J. Kwisthout. Tree-width and the computational complexity of map approximations in bayesian networks. *Journal of Artificial Intelligence Research*, 53:699–720, 2015.

[23] J. Kwisthout. The parameterized complexity of approximate inference in Bayesian networks. In *Proceedings of Machine Learning Research*, volume 52, 2016.

[24] J. Kwisthout, H. Bekkering, and I. van Rooij. To be precise, the details don't matter: On predictive processing, precision, and level of detail of predictions. *Brain and Cognition*, in press.

[25] J. Kwisthout and I. van Rooij. Predictive processing and the Bayesian brain: Intractability hurdles that are yet to overcome. *Journal of Mathematical Psychology*, under review.

[26] M. L. Littman, J. Goldsmith, and M. Mundhenk. The computational complexity of probabilistic planning. *Journal of Artificial Intelligence Research*, 9:1–36, 1998.

[27] W. Maass. Neural computation: a research topic for theoretical computer science? Some thoughts and pointers. In *Bulletin of the European Association for Theoretical Computer Science (EATCS)*, volume 72. 2000.

[28] W. Maass. Noise as a resource for computation and learning in networks of spiking neurons. *Proceedings of the IEEE*, 102(5):860–880, 2014.

[29] D. Marr. *Vision: A computational investigation into the human representation and processing of visual information*. New York: Freeman, 1982.

[30] M. Otworowska, J. Kwisthout, and I. van Rooij. Counter-factual mathematics of counterfactual predictive models. *Frontiers in Consciousness Research*, 5:801, 2014.

[31] M. Otworowska, J. Riemens, C. Kamphuis, P. Wolfert, L. Vuurpijl, and J. Kwisthout. The robohavioral methodology: Developing neuroscience theories with FOES. In *Proceedings of the 27th Benelux Conference on AI (BNAIC'15)*, 2015.

[32] M. Otworowska, L. Zaadnoordijk, E. de Wolff, J. Kwisthout, and I. van Rooij. Causal learning in the crib: A predictive processing formalisation and babybot simulation. In *Proceedings of the Sixth Joint IEEE International Conference on Developmental Learning and Epigenetic Robotics*, 2016.

[33] D. Pecevski, L. Bueling, and W. Maass. Probabilistic inference in general graphical models through sampling in stochastic networks of spiking neurons. *PLoS Computational Biology*, 7(12):1–25, 2011.

[34] W.A. Phillips. Cognitive functions of intracellular mechanisms for contextual amplification. *Brain and Cognition*, in press.

[35] S. Pink-Hashkes and J. Kwisthout. A predictive processing account of psychedelia. In *Interdisciplinary Conference on Psychedelics Research*, 2016.

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

[37] C. Thornton. Predictive processing simplified: The infotropic machine. *Brain and Cognition*, in press.

[38] L. C. van der Gaag and H. L Bodlaender. On stopping evidence gathering for diagnostic Bayesian networks. In W. Liu, editor, *Proceedings of the Eleventh European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, volume 6717 of *LNCS*, pages 170–181, 2011.

[39] H. von Helmholtz. *Handbuch der Physiologischen Optik*. Leipzig: Leopold Voss, 1867.