How does Bayesian reverse-engineering work?

Carlos Zednik (czednik@uos.de) and Frank Jäkel (fjaekel@uos.de)
Institute of Cognitive Science, University of Osnabrück
49069 Osnabrück, Germany

Abstract

Bayesian models of cognition and behavior are particularly promising when they are used in reverse-engineering explanations: explanations that descend from the computational level of analysis to the algorithmic and implementation levels. Unfortunately, it remains unclear exactly how Bayesian models constrain and influence these lower levels of analysis. In this paper, we review and reject two widespread views of Bayesian reverse-engineering, and propose an alternative view according to which Bayesian models at the computational level impose pragmatic constraints that facilitate the generation of testable hypotheses at the algorithmic and implementation levels.

Keywords: Bayesian modeling; rational analysis; reverse-engineering; Marr's levels; mechanistic explanation

Introduction

Bayesian models describe cognitive and behavioral phenomena as a form of optimal statistical inference. Using the methodology of *rational analysis* (Anderson, 1990), researchers attempt to specify the statistical inference task to which a particular phenomenon is adapted. This task is defined formally, in terms of a cognitive system's prior knowledge about its environment, recent evidence collected within that environment, hypotheses being compared, and the relative cost or benefit of particular actions. Once the task has been defined in this way, the mathematical framework of *Bayesian decision theory* can be used to derive an optimal solution to the task: how to ideally adjudicate between hypotheses using Bayes' rule to combine prior knowledge with recent evidence, and how to select actions so as to minimize cost or maximize benefit. If the task has been specified correctly, such optimal solutions often provide descriptively adequate and predictively powerful models of the phenomenon being investigated.

Many researchers regard the methodology of Bayesian modeling as a way to *reverse-engineer* the mind. In cognitive science, reverse-engineering is often associated with David Marr (1982), who proposed that cognitive systems ought to be studied at three distinct levels of analysis. At the computational level, researchers seek to understand *what* a system is doing and *why*. At the algorithmic level, they describe *how* the system does what it does. Finally, at the implementation level, they identify *where* in a particular physical system that algorithm is realized. Reverse-engineering explanations involve descending "a triumphant cascade" of these three levels (Dennett, 1987, p. 227). That is, they begin with a computational-level analysis of a particular cognitive or behavioral phenomenon, and invoke that analysis to articulate and test possible algorithms and implementations of that phenomenon.

What role do Bayesian models play in reverse-engineering explanations? It is widely agreed that Bayesian models figure at the computational level of analysis. They help researchers understand what a cognitive system actually does, because they describe and predict its behavior. Moreover, these models allow researchers to understand why a system does what it does, because they show that the system's behavior is an optimal solution to a particular statistical inference task. But how can Bayesian models at the computational level of analysis be used to identify algorithms and implementations at lower levels?

In what follows, we review three different answers to this question. The first two answers—*Bayesian Realism* and *Instrumentalist Bayesianism*—are well-represented in the literature, but are ultimately unsatisfactory. Thus, we propose a third answer—*Pragmatic Bayesianism*—according to which Bayesian models are tools for hypothesis generation: they facilitate the development of novel algorithmic-level and implementation-level analyses.

Bayesian Realism

According to Bayesian Realism, Bayesian models at the computational level of analysis contribute to reverse-engineering explanations because their mathematical structure is reflected in the functional and physical structure of the mechanisms described at the algorithmic and implementation levels. Insofar as a particular cognitive or behavioral phenomenon can be modeled as a (nearly-) optimal solution to a statistical inference task, Bayesian Realism implies that the mechanisms responsible for this phenomenon themselves perform Bayesian inference. That is, they execute algorithms that invoke (or closely approximate) Bayes' rule to combine prior knowledge with new evidence, and are implemented by neural structures that represent prior and posterior probability distributions, as well as likelihood and loss functions.

Two arguments speak in favor of Bayesian Realism. The first argument is inspired by the classic "no-miracles" argument for scientific realism in philosophy of science. The no-miracles argument seeks to explain the observation that many well-confirmed scientific theories are exceedingly accurate descriptive and predictive devices. Barring miracles, the best explanation seems to be that these theories are true: their theoretical posits successfully refer, and the structures they describe accurately reflect the structure of the world. In much the same way, Bayesian Realism is motivated by the desire to explain the descriptive and predictive successes of Bayesian models at the computational level of analysis. Barring miracles, the best

explanation seems to be that the mathematical structures and processes used to describe cognition and behavior at this level are reflected in the functional processes and physical structures at the algorithmic and implementation levels of analysis.

This argument is most clearly at work in current neuroscientific research on perception. Following a series of psychophysical studies demonstrating that perceptual cue-combination is performed with near-optimal efficiency (Ernst & Banks, 2002), neuroscientists have sought to identify the neural structures and processes responsible for this efficiency. More often than not, the observed behavioral optimality motivates the *Bayesian Coding Hypothesis* (Knill & Pouget, 2004), which claims that the relevant neural structures and processes represent probability distributions, and combine these distributions by applying Bayes' rule. Consider:

> "Recent psychophysical experiments indicate that humans perform near-optimal Bayesian inference in a wide variety of tasks, ranging from cue integration to decision making to motor control. *This implies that* neurons both represent probability distributions and combine those distributions according to a close approximation to Bayes' rule." (Ma, Beck, Latham, & Pouget, 2006, p. 1432, emphasis added)

How else, if not by representing probability distributions and computing over them with (close approximations to) Bayes' rule, could this kind of behavioral optimality be achieved?

The second argument for Bayesian Realism cites the relative ease by which Bayesian inference could be implemented in the brain. Consider the idea of *probabilistic population coding*. Traditionally, it is thought that a population of neurons represents (in a distributed fashion) exactly one value, such as the direction of perceived motion. It is not hard, however, to interpret the population as representing a full probability distribution over the variable in question. Thus, the neurons that have less probable characteristic stimuli fire less than neurons that represent more probable stimuli. On the assumption that neural populations encode information probabilistically in this way, it is also quite easy to explain how they might be combined using Bayes' rule. For example, if population codes exhibit Poisson-like variability—i.e. the ratio of spike count to spike variance is near 1.0—Bayes' rule can be applied to them by simply adding or subtracting their activation levels (Ma et al., 2006). Notably, it has been observed that sensory neuron populations do in fact exhibit Poisson-like variability (Tolhurst, Movshon, & Dean, 1983).

If Bayesian inference is so easy to implement, it would seem surprising to find that the brain—subject to countless evolutionary and developmental constraints—does not actually do so. Thus, Poisson-like variability and similar measures of brain activity are sometimes referred to as *signatures* of Bayesian inference in the brain: neural processes or properties that, although not yet demonstrably related to any particular cognitive or behavioral phenomenon, are suggestive of Bayesian inference.

These arguments for Bayesian Realism promise a bright future for reverse-engineering explanations in cognitive science. This is because, if true, Bayesian Realism can be used to justify inferences from the mathematical structure of the cognitive, perceptual or behavioral task being solved to the functional and physical structure of the mechanisms solving it. If it can be shown that overt behavior is a form of optimal statistical inference that combines evidence with prior probabilities and likelihood functions, Bayesian Realism implies that the neural mechanisms responsible for this behavior will do so as well. Even before consulting the neuroscience, Bayesian Realists have a pretty good understanding of how the brain works!

Unfortunately, empirical support for Bayesian Realism is weak: critics have questioned the quality of evidence typically cited in its favor. For example, Bowers & Davis (2012) argue that Poisson-like variability and other neural signatures of Bayesian inference are consistent with several (non-Bayesian) alternatives, and moreover, suggest that Ma et al. over-estimate the prevalence of these signatures in the brain. Similarly, Maloney & Mamassian (2009) demonstrate that many different algorithms can perform optimal Bayesian inference, though not all of them invoke Bayes' rule and represent prior probability distributions. In particular, "any observer that can combine cues linearly and somehow select the correct weights for the linear combination can duplicate the performance of the Bayesian observer"—even a suitably rigged-up lookup table (Maloney & Mamassian, 2009, p. 149).

Without empirical support, the arguments favoring Bayesian Realism are unsound: it is no longer clear whether Bayesian inference really is as easy as Ma et al. contend, and it is unclear whether Bayesian Realism really is the best (as opposed to a merely possible) explanation of the descriptive and predictive success of Bayesian models.

Instrumentalist Bayesianism

Bayesian Instrumentalism is the view that Bayesian models at the computational level are *mere* descriptive and predictive devices, and that they are compatible with a wide variety of algorithms and implementations at lower levels of analysis. As Colombo & Series (2012) have already observed, many proponents of Bayesian modeling seem to adopt such an instrumentalist perspective. In one of the original discussions of rational analysis, John Anderson suggests that this methodology "provides an explanation at a level of abstraction above specific mechanistic proposals" (Anderson, 1991, p. 471). Similarly, Griffiths et al. (2010) argue that "Using probabilistic models to provide a computational-level explanation does not require that hypothesis spaces or probability distributions be explicitly represented by the underlying psychological or neural processes, or that people learn and reason by explicitly

using Bayes' rule" (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010, p. 362).

The most compelling evidence favoring Bayesian Instrumentalism is the *formal independence* of levels. In one oft-cited passage David Marr states:

> "The three levels are coupled, but only loosely. The choice of an algorithm is influenced, for example, by what it has to do and by the hardware in which it must run. But there is a wide choice available at each level, and the explication of each level involves issues that are rather independent of the other two." (Marr, 1982, p. 25)

It is a well-known mathematical fact that every function can be computed by an infinite number of non-equivalent algorithms. Because Bayesian models specify mathematical functions, they are compatible with any number of algorithms. Thus, although the algorithmic level of analysis is minimally constrained insofar as only those algorithms come into question that actually compute the function specified at the computational level, there are still an infinite number of algorithms to choose from. In much the same way, there are innumerable ways in which any particular algorithm might be implemented in physical hardware. Thus, the formal independence of levels implies that developers of Bayesian models at the computational level of analysis ought to be agnostic about the kinds of algorithms and implementations that can be posited at lower levels.

Great care must be taken not to confuse agnosticism about lower levels with a rejection of their explanatory relevance. In an influential recent critique, Jones & Love (2011) outline a position they disparagingly call *Bayesian Fundamentalism*. Like Instrumentalist Bayesianism, this position denies that Bayesian models at the computational level constrain the lower levels of analysis. Rather than be agnostic about these lower levels, however, Bayesian Fundamentalists deny that lower levels of analysis are explanatorily relevant: "human behavior can be explained through rational analysis…without recourse to process representation, resource limitations, or physiological or developmental data" (Jones & Love, 2011, p. 170). This radical position, Jones & Love argue, smacks of behaviorism, and ought to be avoided: "it would be a serious overreaction simply to discard everything below the computational level. As in nearly every other science, understanding *how* the subject of study (i.e., the brain) operates is critical to explaining and predicting its behavior" (Jones & Love, 2011, p. 177, original emphasis).

The most common response to this worry has been to deny that proponents of Bayesian modeling in cognitive science are correctly associated with Bayesian Fundamentalism. In a direct response to Jones & Love's target article, Chater et al. characterize Bayesian Fundamentalism as "purely a construct of Jones & Love's imagination" (Chater et al., 2011, p. 194). Indeed, given their intellectual debt to David Marr—who stresses that a cognitive system must be studied at all three levels "before one can be said to have understood it completely" (Marr, 1982, p. 24)—such an association would be surprising.

But there are more significant worries than the false specter of fundamentalism. According to Instrumentalist Bayesianism, it is unclear that systematic reverse-engineering is possible: it would seem exceedingly unlikely that a "triumphant cascade" can be descended in a principled way. Although research into the neuroscientific underpinnings of Bayesian inference might be *inspired* by the descriptive and predictive success of Bayesian models of cognition and behavior, such research would not be *justified* by this success. Given the formal independence of levels, there is no reason to believe that the mathematical structure of Bayesian models at the computational level of analysis is reflected at lower levels. Of course, the lower levels should somehow compute and implement the function specified by the Bayesian model, by mapping stimuli onto responses as the model predicts. But there is no reason to believe that e.g. neural populations encode loss functions, posteriors, likelihoods and priors, as opposed to reproducing the modeled stimulus-response behavior in some other way. Thus, even if future neuroscientific research were to eventually confirm the Bayesian Coding Hypothesis, this confirmation would not result from a systematic reverse-engineering effort.

Pragmatic Bayesianism

Bayesian Realism and Instrumentalist Bayesianism are the two most widely-held views on how Bayesian models at the computational level relate to the algorithmic and implementation levels of analysis. Unfortunately, neither view accounts for the possibility of reverse-engineering explanations in cognitive science. Whereas the arguments favoring Bayesian Realism are as of yet inconclusive due to lack of empirical evidence, Instrumentalist Bayesianism makes systematic reverse-engineering impossible.

This section introduces an alternative view. According to *Pragmatic Bayesianism*, Bayesian models at the computational level make reverse-engineering possible by facilitating the generation of novel hypotheses at the algorithmic and implementation levels of analysis. Although levels of analysis may be formally independent, they are *pragmatically dependent*. If a particular cognitive or behavioral phenomenon can be modeled as a form of Bayesian inference, it will be considerably easier to identify possible algorithms to perform this kind of inference, and to identify ways in which these algorithms might be implemented in physical hardware. How so? Because practicing researchers are (a) guided by pragmatic considerations such as their interdisciplinary colleagues' previous research activity, ingenuity and communicative ability, and (b) influenced in their scientific decision-making by the conceptual and theoretical framework of Bayesian statistical inference.

An effective segue into Pragmatic Bayesianism is Colombo & Series' defense of Instrumentalist Bayesianism. Although they do not identify it as such, Colombo & Series

describe one important pragmatic influence on reverse-engineering: "the predictive success of a Bayesian model in a given psychophysical task can *motivate* us to investigate why this is the case" (Colombo & Series, 2012, p. 17, original emphasis). Undeniably, researchers' motivations critically influence the development of algorithms and implementations for a particular kind of Bayesian inference. At the same time, however, Colombo & Series claim that "the discovery that people behave as though they were Bayesian observers does not compel us to make any specific claim at the neural level of implementation" (Colombo & Series, 2012, p. 17). The supposed reason for this is the aforementioned formal independence of levels. However, although there may be no theoretical limit to the number of algorithms that compute a particular mathematical function, pragmatic considerations impose considerable limits on the number of algorithms and implementations that will actually be considered. Importantly, although these algorithms and implementations might reflect the mathematical structure of Bayesian models at the computational level, they need not do so.

Constraints on algorithm-development

Consider recent attempts to develop algorithmic-level analyses to accompany John Anderson's (1991) rational analysis of categorization. One such analysis is developed by Anderson himself, and centers on "a type of iterative algorithm that has appeared in the artificial intelligence literature" (Anderson, 1991, p. 412). By reviewing the categorization literature of the time, Anderson shows that the iterative algorithm accurately predicts qualitative and quantitative human data. Moreover, Anderson suspects (but does not prove) that the iterative algorithm closely approximates the optimal assignment of objects to categories within the constraints of the task environment. Sanborn et al. (2010) later demonstrate that although the iterative algorithm approximates optimal Bayesian inference in the task environments Anderson considers, there is no guarantee that it will do so in general. Thus, Sanborn et al. present two alternative algorithms—*particle filtering* and *Gibbs sampling*—both of which "can approximate the optimal inference to any desired level of precision" (Sanborn et al., 2010, p. 1145). Ultimately, by comparing all three candidate algorithms to experimental data, Sanborn et al. propose particle filtering as the most plausible algorithmic-level analysis of human categorization.

Two things are worth noticing about this series of articles (See also: Griffiths, Vul, & Sanborn, 2012). First, each one of the three proposed algorithms is adapted or straightforwardly coopted from existing research in the discipline of artificial intelligence. Second, although each one of these algorithms approximates Anderson's model of categorization, neither one of them requires explicit representations of the full hypothesis space, prior probability distributions and likelihoods, nor directly invokes Bayes' rule to compute over these representations.

Researchers working in the discipline of artificial intelligence (including machine learning and statistics), have developed many different algorithms for optimally and efficiently computing or approximating Bayesian inference, only a limited number of which directly apply Bayes' rule to full probability distributions. It seems natural to wonder whether algorithms already developed for theoretical reasons or real-world applications might serve double-duty in cognitive science. As the series of articles on categorization demonstrates, describing a particular cognitive or behavioral phenomenon as a form of Bayesian statistical inference at the computational level allows researchers in cognitive science to consider existing artificial intelligence research not just for *motivation* in the way suggested by Colombo & Series, but for *articulating testable hypotheses* at the algorithmic level of analysis. As is exemplified by the particle filtering algorithm advanced by Sanborn et al., these hypotheses need not reflect the mathematical structure of Bayesian models at the computational level of analysis.

There is a clear sense in which any pragmatic consideration that contributes to the generation of testable hypotheses might be thought to facilitate reverse-engineering explanations in cognitive science. At the same time, recall that one of the worries about Instrumentalist Bayesianism was that, although lower-level analyses may be inspired by Bayesian models at the computational level, they are not *justified* by these models. In what sense are psychologists justified in invoking algorithms developed by artificial intelligence researchers who are unconcerned with matters of psychological and biological plausibility?

A useful framework for answering this question is Herbert Simon's influential account of scientific discovery (Simon, Langley, & Bradshaw, 1981). Simon views scientific discovery as a form of problem-solving, in which researchers are tasked with exploring the conceptual space of possible solutions to a particular scientific problem. Because this space is often vast and multidimensional, researchers rely on heuristic strategies that highlight particular areas within the space to the exclusion of others, thereby limiting the number of possible solutions they actually need to consider. Although these heuristic strategies are fallible—they might erroneously highlight an irrelevant area within the space or exclude a relevant one—their use is justified insofar as they allow researchers to efficiently and systematically traverse the space of possible solutions to a particular scientific-discovery problem.

The appeal to existing research in artificial intelligence, statistics, and machine learning that is facilitated by Bayesian models in cognitive science can be understood as a heuristic strategy of this kind. Modeling a particular cognitive or behavioral phenomenon as a form of Bayesian inference is tantamount to defining a particular scientific-discovery problem: the problem of selecting, from among the set of algorithms that *possibly* perform or approximate such inference, the algorithm that *actually* does so in the particular cognitive system being studied. Unfortunately,

because every function can be computed by an infinite number of algorithms, the solution-space is infinite in expanse. Nevertheless, by appealing to the existing literature in artificial intelligence, researchers can concentrate their efforts on particular regions of the space—those regions that have already been explored in theoretical work or real-world applications. Because only a limited number of algorithms have actually been articulated and studied, researchers in cognitive science are able to select from (and if necessary adapt) a handful of well-understood alternatives. Interestingly, this means that the reverse-engineering explanations in cognitive science are constrained in an irreducibly pragmatic way, by the research output of other scientific disciplines.

Constraints on implementation-description

Bayesian models at the computational level of analysis also pragmatically constrain the implementation level of analysis. In order to provide an analysis of implementation, (neuro-)scientists must identify and describe the particular physical structures and processes which realize the algorithm that computes a particular mathematical function. In order to do so, they have several decisions to make: what are the relevant physical structures and processes? Which aspects of these structures and processes should be emphasized? How should they be described? Bayesian models at the computational level often directly influence the outcome of these decisions, but also influence them indirectly, by way of the algorithmic level.

As the previous discussion shows, Bayesian models at the computational level pragmatically constrain the selection of algorithms at the algorithmic level of analysis. In turn, the algorithms considered at this level influence the description of implementing neurobiological mechanisms. Consider once again the particle filtering algorithm proposed by Sanborn et al. (2010). Particle filtering is an example of a general class of algorithms known as *Monte Carlo sampling*. Recently, Fiser et al. (2010) have appealed to this class of algorithms to interpret spontaneous neural activity in the absence of sensory stimulation:

> "Under a sampling-based representational account, spontaneous activity could have a natural interpretation. In a probabilistic framework, if neural activities represent samples from a distribution over external variables, this distribution must be the so-called 'posterior distribution'. The posterior distribution is inferred by combining information from two sources: the sensory input, and the prior distribution describing *a priori* beliefs about the sensory environment. Intuitively, in the absence of sensory stimulation, this distribution will collapse to the prior distribution, and spontaneous activity will represent this prior." (Fiser et al., 2010, pp. 125–127)

The presence of spontaneous neural activity has long been interpreted as stochastic noise (Tolhurst et al., 1983). In contrast, by appealing to the framework of Monte Carlo sampling, Fiser et al. advance an interpretation according to which "a very large component of high spontaneous activity is probably not noise but might have a functional role in cortical computation" (Fiser et al., 2010, p. 125). Thus, because they adopt a theoretical perspective that is "colored" by a particular class of algorithms, Fiser et al. arrive at a very different way of describing particular neural structures and processes. Indeed, on their interpretation, spontaneous neural activity is not merely a neural signature of Bayesian inference, but of Bayesian inference by way of Monte Carlo sampling. Insofar as Bayesian models at the computational level suggest Monte Carlo sampling (or more specifically according to Sanborn et al., particle filtering) as a possible algorithmic-level account of behavior and cognition, these models also *indirectly* suggest particular ways of interpreting, individuating and describing certain neurobiological structures and processes.

Bayesian models at the computational level may also influence the implementation level quite directly. In recent philosophical research on mechanistic explanation in neuroscience, Carl Craver (2013) identifies three ways in which neuroscientists' decision-making is influenced by available characterizations of a mechanism's function. First, mechanisms are *defined* in functional terms: they are always mechanisms for something. Thus, neurotransmitters are "used to send signals from one cell to another" (Craver, 2013, p. 135), much like soda machines are used to dispense cans of soda in exchange for money. Second, mechanisms are typically *delineated* by appealing to functional characterizations which serve to distinguish a mechanism from its background or environment. In Craver's words:

> "it takes considerable scientific effort, abstraction, and idealization to distinguish components from contraband, activities from incidental interactions, and causes from background conditions. And this filtering process requires (essentially) fixing on some behavior, process, or function for which a mechanistic explanation will be sought" (Craver, 2013, p. 140).

Third and finally, the way mechanisms are *decomposed* also typically relies on characterizations of function. Following Craver, such characterizations determine the particular physical structures and processes that are actually relevant to the production of the phenomenon being investigated.

Notably, each one of these three constraints is pragmatic in character: it concerns influences on a researcher's decision-making, focus of research, and descriptive emphasis. Although the neural structures and processes that compose a mechanism are real things in the world, the particular way in which they are described is invariably tied to previously available characterizations of function. Now, recall that Bayesian models figure at Marr's computational level not just because they allow researchers to describe what a cognitive system actually does, but also because they help them understand *why* the system behaves as it does. Specifically, Bayesian models show that the system behaves

as it does because this particular behavior is an optimal solution to the task environment within which the system is situated. Thus, Bayesian models seem ideally suited for imposing the kinds of pragmatic constraints on implementation identified by Craver.

Consider again the work in theoretical neuroscience discussed in the context of Bayesian Realism above. Much of this research is inspired by the descriptive and predictive success of Bayesian models in cognitive psychology and psychophysics. Notably, this success not only motivates neuroscientists to look for possible neural implementations of Bayesian inference, but also regularly suggests the particular form these implementations might take: the functional and physical structure of mechanisms at the implementation level is assumed to reflect the mathematical structure of Bayesian models at the computational level. Thus for example, in a passage already quoted above, Ma et al. (2006) claim that the descriptive success of Bayesian models "implies that" neurons represent probability distributions and implement Bayes' rule.

Although this kind of research has yet to provide conclusive evidence in favor of the Bayesian Coding Hypothesis, it confirms Craver's philosophical analysis. Specifically, it shows that characterizations of function—in this case, Bayesian models—influence neuroscientists' decisions about how to define, delineate, and decompose mechanisms. Thus, Bayesian models at the computational level of analysis directly influence the implementation level by suggesting possible ways of interpreting the activity of certain neural mechanisms, but also by suggesting which particular neural structures and processes to include in descriptions of these mechanisms. Because Bayesian models at the computational level pragmatically constrain algorithmic and implementation level analysis, they are a viable starting point for reverse-engineering explanations in cognitive science.

Conclusion

Although Bayesian Realism makes reverse-engineering explanations easy, empirical support for this position is weak. Many practicing researchers have therefore endorsed Instrumentalist Bayesianism. Unfortunately, this position makes systematic reverse-engineering impossible. Unlike these more established alternatives, Pragmatic Bayesianism both provides a satisfying account of scientific practice and allows for systematic reverse-engineering in cognitive science.

References

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The Adaptive Nature of Human Categorization. *Psychological Review*, *98*(3), 409–429.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*(3), 389–414.

Chater, N., Goodman, N., Griffiths, T. L., Kemp, C., Oaksford, M., & Tenenbaum, J. B. (2011). The imaginary fundamentalists: The unshocking truth about Bayesian cognitive science. *Behavioral and Brain Sciences*, *34*, 194–196.

Colombo, M., & Series, P. (2012). Bayes in the Brain--On Bayesian Modelling in Neuroscience. *The British Journal for the Philosophy of Science*, *63*(3), 697–723. doi:10.1093/bjps/axr043

Craver, C. F. (2013). Functions and Mechanisms: A Perspectivalist View. In P. Huneman (Ed.), *Functions: Selection and Mechanisms* (pp. 133–158). Dordrecht: Springer.

Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(January), 429–433.

Fiser, J., Berkes, P., Orban, G., & Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, *14*(3), 119–130. doi:10.1016/j.tics.2010.01.003

Griffiths, T. L., 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. doi:10.1016/j.tics.2010.05.004

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging Levels of Analysis for Probabilistic Models of Cognition. *Current Directions in Psychological Science*, *21*(4), 263–268. doi:10.1177/0963721412447619

Jones, M., & Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*, 169–231.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–9.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438.

Maloney, L. T., & Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Visual Neuroscience*, *26*, 147–155.

Marr, D. (1982). *Vision*. New York, NY: Henry Holt & Co.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational Approximations to Category Learning. *Psychological Review*, *117*(4), 1144–1167.

Simon, H. A., Langley, P. W., & Bradshaw, G. L. (1981). Scientific Discovery as Problem Solving. *Synthese*, *47*(1), 1–27.

Tolhurst, D. J., Movshon, J. A., & Dean, A. F. (1983). The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Research*, *23*(8), 775–785.