BEHAVIORAL AND BRAIN SCIENCES (2011) **34**, 169–231
doi:10.1017/S0140525X10003134

Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition

Matt Jones
Department of Psychology and Neuroscience, University of Colorado, Boulder, CO 80309
mcj@colorado.edu **http://matt.colorado.edu**

Bradley C. Love
Department of Psychology, University of Texas, Austin, TX 78712
brad_love@mail.utexas.edu **http://love.psy.utexas.edu**

Abstract: The prominence of Bayesian modeling of cognition has increased recently largely because of mathematical advances in specifying and deriving predictions from complex probabilistic models. Much of this research aims to demonstrate that cognitive behavior can be explained from rational principles alone, without recourse to psychological or neurological processes and representations. We note commonalities between this rational approach and other movements in psychology – namely, Behaviorism and evolutionary psychology – that set aside mechanistic explanations or make use of optimality assumptions. Through these comparisons, we identify a number of challenges that limit the rational program's potential contribution to psychological theory. Specifically, rational Bayesian models are significantly unconstrained, both because they are uninformed by a wide range of process-level data and because their assumptions about the environment are generally not grounded in empirical measurement. The psychological implications of most Bayesian models are also unclear. Bayesian inference itself is conceptually trivial, but strong assumptions are often embedded in the hypothesis sets and the approximation algorithms used to derive model predictions, without a clear delineation between psychological commitments and implementational details. Comparing multiple Bayesian models of the same task is rare, as is the realization that many Bayesian models recapitulate existing (mechanistic level) theories. Despite the expressive power of current Bayesian models, we argue they must be developed in conjunction with mechanistic considerations to offer substantive explanations of cognition. We lay out several means for such an integration, which take into account the representations on which Bayesian inference operates, as well as the algorithms and heuristics that carry it out. We argue this unification will better facilitate lasting contributions to psychological theory, avoiding the pitfalls that have plagued previous theoretical movements.

Keywords: Bayesian modeling; cognitive processing; levels of analysis; rational analysis; representation

1. Introduction

Advances in science are due not only to empirical discoveries and theoretical progress, but also to development of new formal frameworks. Innovations in mathematics or related fields can lead to a new class of models that enables researchers to articulate more sophisticated theories and to address more complex empirical problems than previously possible. This often leads to a rush of new research and a general excitement in the field.

For example in physics, the development of tensor calculus on differential manifolds (Ricci & Levi-Civita 1900) provided the mathematical foundation for formalizing the general theory of relativity (Einstein 1916). This formalism led to quantitative predictions that enabled experimental verification of the theory (e.g., Dyson et al. 1920). More recent mathematical advances have played key roles in the development of string theory (a potential unification of general relativity and quantum mechanics), but in this

MATT JONES is an Assistant Professor in the Department of Psychology and Neuroscience at the University of Colorado, Boulder. He received an M.A. in Statistics (2001) and a Ph.D. in Cognitive Psychology (2003) from the University of Michigan. His research focuses on mathematical modeling of cognition, including learning, knowledge representation, and decision making.

BRADLEY C. LOVE is a full Professor in the Psychology Department at the University of Texas at Austin. In 1999, he received a Ph.D. in Cognitive Psychology from Northwestern University. His research centers on basic issues in cognition, such as learning, memory, attention, and decision making, using methods that are informed by behavior, brain, and computation. Late in 2011, he will relocate to University College London.

 0140-525X/11 $40.00

case the mathematical framework, although elegant, has yet to make new testable predictions (Smolin 2006; Woit 2006). Therefore, it is difficult to evaluate whether string theory represents true theoretical progress.

In the behavioral sciences, we are generally in the more fortunate position of being able to conduct the key experiments. However, there is still a danger of confusing technical advances with theoretical progress, and the allure of the former can lead to the neglect of the latter. As the new framework develops, it is critical to keep the research tied to certain basic questions, such as: What theoretical issues are at stake? What are the core assumptions of the approach? What general predictions does it make? What is being explained and what is the explanation? How do the explanations it provides relate, logically, to those of existing approaches? What is the domain of inquiry, and what questions are outside its scope? This grounding is necessary for disciplined growth of the field. Otherwise, there is a tendency to focus primarily on generating existence proofs of what the computational framework can achieve. This comes at the expense of real theoretical progress, in terms of deciding among competing explanations for empirical phenomena or relating those explanations to existing proposals. By overemphasizing computational power, we run the risk of producing a poorly grounded body of work that is prone to collapse under more careful scrutiny.

This article explores these issues in connection with Bayesian modeling of cognition. Bayesian methods have progressed tremendously in recent years, due largely to mathematical advances in probability and estimation theory (Chater et al. 2006). These advances have enabled theorists to express and derive predictions from far more sophisticated models than previously possible. These models have generated excitement for at least three reasons: First, they offer a new interpretation of the goals of cognitive systems, in terms of inductive probabilistic inference, which has revived attempts at rational explanation of human behavior (Oaksford & Chater 2007). Second, this rational framing can make the assumptions of Bayesian models more transparent than in mechanistically oriented models. Third, Bayesian models may have the potential to explain some of the most complex aspects of human cognition, such as language acquisition or reasoning under uncertainty, where structured information and incomplete knowledge combine in a way that has defied previous approaches (e.g., Kemp & Tenenbaum 2008).

Despite this promise, there is a danger that much of the research within the Bayesian program is getting ahead of itself, by placing too much emphasis on mathematical and computational power at the expense of theoretical development. In particular, the primary goal of much Bayesian cognitive modeling has been to demonstrate that human behavior in some task is rational with respect to a particular choice of Bayesian model. We refer to this school of thought as *Bayesian Fundamentalism*, because it strictly adheres to the tenet that human behavior can be explained through rational analysis – once the correct probabilistic interpretation of the task environment has been identified – without recourse to process, representation, resource limitations, or physiological or developmental data. Although a strong case has been made that probabilistic inference is the appropriate framework for normative accounts of cognition (Oaksford & Chater 2007), the fundamentalist approach primarily aims to reinforce this position, without moving on to more substantive theoretical development or integration with other branches of cognitive science.

We see two significant disadvantages to the fundamentalist approach. First, the restriction to computational-level accounts (cf. Marr 1982) severely limits contact with process-level theory and data. Rational approaches attempt to explain why cognition produces the patterns of behavior that is does, but they offer no insight into how cognition is carried out. Our argument is not merely that rational theories are limited in what they can explain (this applies to all modes of explanation), but that a complete theory of cognition must consider both rational and mechanistic explanations as well as their interdependencies, rather than treating them as competitors. Second, the focus on existence proofs obfuscates the fact that there are generally multiple rational theories of any given task, corresponding to different assumptions about the environment and the learner's goals. Consequently, there is insufficient acknowledgement of these assumptions and their critical roles in determining model predictions. It is extremely rare to find a comparison among alternative Bayesian models of the same task to determine which is most consistent with empirical data (see Fitelson [1999] for a related analysis of the philosophical literature). Likewise, there is little recognition when the critical assumptions of a Bayesian model logically overlap closely with those of other theories, so that the Bayesian model is expressing essentially the same explanation, just couched in a different framework.

The primary aim of this article is to contrast Bayesian Fundamentalism with other Bayesian research that explicitly compares competing rational accounts and that considers seriously the interplay between rational and mechanistic levels of explanation. We call this the *Enlightened Bayesian* approach, because it goes beyond the dogma of pure rational analysis and actively attempts to integrate with other avenues of inquiry in cognitive science. A critical distinction between Bayesian Fundamentalism and Bayesian Enlightenment is that the latter considers the elements of a Bayesian model as claims regarding psychological process and representation, rather than mathematical conveniences made by the modeler for the purpose of deriving computational-level predictions. Bayesian Enlightenment thus treats Bayesian models as making both rational and mechanistic commitments, and it takes as a goal the joint evaluation of both. Our aim is to initiate a discussion of the distinctions and relative merits of Bayesian Fundamentalism and Bayesian Enlightenment, so that future research can focus effort in the directions most likely to lead to real theoretical progress.

Before commencing, we must distinguish a third usage of Bayesian methods in the cognitive and other sciences, which we refer to as *Agnostic Bayes*. Agnostic Bayesian research is concerned with inferential methods for deciding among scientific models based on empirical data (e.g., Pitt et al. 2002; Schwarz 1978). This line of research has developed powerful tools for data analysis, but, as with other such tools (e.g., analysis of variance, factor analysis), they are not intended as models of cognition itself. Because it has no position on whether the Bayesian

framework is useful for describing cognition, Agnostic Bayes is not a topic of the present article. Likewise, research in pure Artificial Intelligence that uses Bayesian methods without regard for potential correspondence with biological systems is beyond the scope of this article. There is no question that the Bayesian framework, as a formal system, is a powerful scientific tool. The question is how well that framework parallels the workings of human cognition, and how best to exploit those parallels to advance cognitive science.

The rest of this article offers what we believe is an overdue assessment of the Bayesian approach to cognitive science, including evaluation of its theoretical content, explanatory status, scope of inquiry, and relationship to other methods. We begin with a discussion of the role that new metaphors play in science, and cognitive science in particular, using connectionism as an historical example to illustrate both the potential and the danger of rapid technical advances within a theoretical framework. An overview of Bayesian modeling of cognition is then presented, that attempts to clarify what is and is not part of a Bayesian psychological theory. Following this, we offer a critical appraisal of the Fundamentalist Bayesian movement. We focus on concerns arising from the limitation to strictly computational-level accounts, by noting commonalities between the Bayesian program and other movements, namely Behaviorism and evolutionary psychology, that have minimized reliance on mechanistic explanations in favor of explaining behavior directly from the environment. Finally, we outline the Enlightened Bayesian perspective, give examples of research in this line, and explain how this approach leads to a more productive use of the Bayesian framework and better integration with other methods in cognitive science. Like many others, we believe that Bayes' mathematical formalism has great potential to aid our understanding of cognition. Our aim is not to undermine that potential, but to focus it by directing attention to the important questions that will allow disciplined, principled growth and integration with existing knowledge. Above all, our hope is that by the time the excitement has faded over their newfound expressive power, Bayesian theories will be seen to have something important to say.

2. Metaphor in science

Throughout the history of science, metaphor and analogy use has helped researchers gain insight into difficult problems and make theoretical progress (Gentner et al. 1997; Nersessian 1986; Thagard 1989). In addition to this evidence gleaned from the personal journals of prominent scientists, direct field observation of modern molecular biologists finds that analogies are commonly used in laboratory discussions (Dunbar 1995). Metaphors and analogies provide powerful means for structuring an abstract or poorly understood domain in terms of a more familiar domain, such as understanding the atom in terms of the solar system (Gentner 1983). Drawing these parallels can lead to insights and be a source of new ideas and hypotheses.

Daugman (2001) reviews the historical usage of metaphor for describing brain function and concludes that current technology has consistently determined the dominant choice of metaphor, from water technology to clocks to engines to computers. Whatever society at large views as its most powerful device tends to become our means for thinking about the brain, even in formal scientific settings. Despite the recurring tendency to take the current metaphor literally, it is important to recognize that any metaphor will eventually be supplanted. Thus, researchers should be aware of what the current metaphor contributes to their theories, as well as what the theories' logical content is once the metaphor is stripped away.

One danger is mistaking metaphors for theories in themselves. In such cases, scientific debate shifts focus from comparisons of theories within established frameworks to comparisons among metaphors. Such debates are certainly useful in guiding future research efforts, but it must be recognized that questions of metaphor are not scientific questions (at best, they are meta-scientific). Metaphors should be viewed as tools or languages, not theories in themselves. Conflating debates over scientific metaphors with scientific debates per se can impede theoretical progress in a number of ways. By shifting focus to the level of competing metaphors, the logical content of specific theories can become neglected. Research that emphasizes existence proofs, demonstrating that a given set of phenomena can be explained within a given framework, tends to ignore critical comparisons among multiple, competing explanations. Likewise, the emphasis on differences in metaphorical frameworks can obscure the fact that theories cast within different frameworks can have substantial logical overlap. In both ways, basic theory loses out, because too much effort is spent debating the best way to analyze or understand the scientific subject, at the expense of actually doing the analysis. Only by identifying competing explanations, and distilling their differences to logical differences in assumptions and empirically testable contrasting predictions, can true theoretical progress be made.

2.1. The case of connectionism

One illustration of this process within cognitive science comes from the history of connectionism. Connectionism was originally founded on a metaphor with telegraph networks (Daugman 2001) and later on a metaphor between information-processing units and physical neurons (in reaction to the dominant computer metaphor of the 1970s and 1980s). At multiple points in its development, research in connectionism has been marked by technical breakthroughs that significantly advanced the computational and representational power of existing models. These breakthroughs led to excitement that connectionism was the best framework within which to understand the brain. However, the initial rushes of research that followed focused primarily on demonstrations of what could be accomplished within this framework, with little attention to the theoretical commitments behind the models or whether their operation captured something fundamental to human or animal cognition. Consequently, when challenges arose to connectionism's computational power, the field suffered major setbacks, because there was insufficient theoretical or empirical grounding to fall back on. Only after researchers began to take connectionism seriously as a mechanistic model, to address what it could and could not predict, and to consider what

constraints it placed on psychological theory, did the field mature to the point that it was able to make a lasting contribution. This shift in perspective also helped to clarify the models' scope, in terms of what questions they should be expected to answer, and identified shortcomings that in turn spurred further research.

There are of course numerous perspectives on the historical and current contributions of connectionism, and it is not the purpose of the present article to debate these views. Instead, we merely summarize two points in the history of connectionism that illustrate how overemphasis on computational power at the expense of theoretical development can delay scientific progress.

Early work on artificial neurons by McCulloch and Pitts (1943) and synaptic learning rules by Hebb (1949) showed how simple, neuron-like units could automatically learn various prediction tasks. This new framework seemed very promising as a source of explanations for autonomous, intelligent behavior. A rush of research followed, culminated by Rosenblatt's (1962) perceptron model, for which he boldly claimed, "Given an elementary α-perceptron, a stimulus world W, and any classification $C(W)$ for which a solution exists, . . . an error correction procedure will always yield a solution to $C(W)$ in finite time" (p. 111). However, Minsky and Papert (1969) pointed out a fatal flaw: Perceptrons are provably unable to solve problems requiring nonlinear solutions. This straightforward yet unanticipated critique devastated the connectionist movement such that there was little research under that framework for the ensuing 15 years.

Connectionism underwent a revival in the mid-1980s, primarily triggered by the development of *back-propagation*, a learning algorithm that could be used in multilayer networks (Rumelhart et al. 1986). This advance dramatically expanded the representational capacity of connectionist models, to the point where they were capable of approximating any function to arbitrary precision, bolstering hopes that paired with powerful learning rules any task could be learnable (Hornik et al. 1989). This technical advance led to a flood of new work, as researchers sought to show that neural networks could reproduce the gamut of psychological phenomena, from perception to decision making to language processing (e.g., McClelland et al. 1986; Rumelhart et al. 1986). Unfortunately, the bubble was to burst once again, following a series of attacks on connectionism's representational capabilities and lack of grounding. Connectionist models were criticized for being incapable of capturing the compositionality and productivity characteristic of language processing and other cognitive representations (Fodor & Pylyshyn 1988); for being too opaque (e.g., in the distribution and dynamics of their weights) to offer insight into their own operation, much less that of the brain (Smolensky 1988); and for using learning rules that are biologically implausible and amount to little more than a generalized regression (Crick 1989). The theoretical position underlying connectionism was thus reduced to the vague claim that that the brain can learn through feedback to predict its environment, without a psychological explanation being offered of how it does so. As before, once the excitement over computational power was tempered, the shortage of theoretical substance was exposed.

One reason that research in connectionism suffered such setbacks is that, although there were undeniably important theoretical contributions made during this time, overall there was insufficient critical evaluation of the nature and validity of the psychological claims underlying the approach. During the initial explosions of connectionist research, not enough effort was spent asking what it would mean for the brain to be fundamentally governed by distributed representations and tuning of association strengths, or which possible specific assumptions within this framework were most consistent with the data. Consequently, when the limitations of the metaphor were brought to light, the field was not prepared with an adequate answer. On the other hand, pointing out the shortcomings of the approach (e.g., Marcus 1998; Pinker & Prince 1988) was productive in the long run, because it focused research on the hard problems. Over the last two decades, attempts to answer these criticisms have led to numerous innovative approaches to computational problems such as object binding (Hummel & Biederman 1992), structured representation (Pollack 1990), recurrent dynamics (Elman 1990), and executive control (e.g., Miller & Cohen 2001; Rougier et al. 2005). At the same time, integration with knowledge of anatomy and physiology has led to much more biologically realistic networks capable of predicting neurological, pharmacological, and lesion data (e.g., Boucher et al. 2007; Frank et al. 2004). As a result, connectionist modeling of cognition has a much firmer grounding than before.

2.2. Lessons for the Bayesian program?

This brief historical review serves to illustrate the dangers that can arise when a new line of research is driven primarily by technical advances and is not subjected to the same theoretical scrutiny as more mature approaches. We believe such a danger currently exists in regard to Bayesian models of cognition. Principles of probabilistic inference have been prevalent in cognitive science at least since the advent of signal detection theory (Green & Swets 1966). However, Bayesian models have become much more sophisticated in recent years, largely because of mathematical advances in specifying hierarchical and structured probability distributions (e.g., Engelfriet & Rozenberg 1997; Griffiths & Ghahramani 2006) and in efficient algorithms for approximate inference over complex hypothesis spaces (e.g., Doucet et al. 2000; Hastings 1970). Some of the ideas developed by psychologists have been sufficiently sophisticated that they have fed back to significantly impact computer science and machine learning (e.g., Thibaux & Jordan 2007). In psychology, these technical developments have enabled application of the Bayesian approach to a wide range of complex cognitive tasks, including language processing and acquisition (Chater & Manning 2006), word learning (Xu & Tenenbaum 2007b), concept learning (Anderson 1991b), causal inference (Griffiths & Tenenbaum 2009), and deductive reasoning (Chater & Oaksford 1999). There is a growing belief in the field that the Bayesian framework has the potential to solve many of our most important open questions, as evidenced by the rapid increase in the number of articles published on Bayesian models, and by optimistic assessments such as this one made by Chater and Oaksford: "In the [last] decade, probabilistic models have flourished . . . [The current wave of

researchers] have considerably extended both the technical possibilities of probabilistic models and their range of applications in cognitive science" (Chater & Oaksford 2008, p. 25).

One attraction of the Bayesian framework is that it is part of a larger class of models that make inferences in terms of probabilities. Like connectionist models, probabilistic models avoid many of the challenges of symbolic models founded on Boolean logic and classical artificial intelligence (e.g., Newell & Simon 1972). For example, probabilistic models offer a natural account of non-monotonic reasoning, avoiding the technical challenges that arise in the development of non-monotonic logics (see Gabbay et al. 1994). Oaksford and Chater (2007) make a strong case that probabilistic models have greater computational power than propositional models, and that the Bayesian framework is the more appropriate standard for normative analysis of human behavior than is that of classical logic (but see Binmore [2009] for an important counterargument). Unfortunately, most of the literature on Bayesian modeling of cognition has not moved past these general observations. Much current research falls into what we have labeled Bayesian Fundamentalism, which emphasizes promotion of the Bayesian metaphor over tackling genuine theoretical questions. As with early incarnations of connectionism, the Bayesian Fundamentalist movement is primarily driven by the expressive power – both computational and representational – of its mathematical framework. Most applications to date have been existence proofs, in that they demonstrate a Bayesian account is possible without attempting to adjudicate among (or even acknowledge) the multiple Bayesian models that are generally possible, or to translate the models into psychological assumptions that can be compared with existing approaches. Furthermore, amidst the proliferation of Bayesian models for various psychological phenomena, there has been surprisingly little critical examination of the theoretical tenets of the Bayesian program as a whole.

Taken as a psychological theory, the Bayesian framework does not have much to say. Its most unambiguous claim is that much of human behavior can be explained by appeal to what is rational or optimal. This is an old idea that has been debated for centuries (e.g., Kant 1787/1961). More importantly, rational explanations for behavior offer no guidance as to how that behavior is accomplished. As already mentioned, early connectionist learning rules were subject to the same criticism, but connectionism is naturally suited for grounding in physical brain mechanisms. The Bayesian framework is more radical in that, unlike previous brain metaphors grounded in technology and machines, the Bayesian metaphor is tied to a mathematical ideal and thus eschews mechanism altogether. This makes Bayesian models more difficult to evaluate. By locating explanations firmly at the computational level, the Bayesian Fundamentalist program renders irrelevant many major modes of scientific inquiry, including physiology, neuroimaging, reaction time, heuristics and biases, and much of cognitive development (although, as we show in section 6, this is not a necessary consequence of the Bayesian framework itself). All of these considerations suggest it is critical to pin Bayes down, to bring the Bayesian movement past the demonstration phase and get to the real work of using Bayesian models, in integration with other approaches, to understand the detailed workings of the mind and brain.

3. Bayesian inference as a psychological model

Bayesian modeling can seem complex to the outsider. The basic claims of Bayesian modeling can be completely opaque to the non-mathematically inclined. In reality, the presuppositions of Bayesian modeling are fairly simple. In fact, one might wonder what all the excitement is about once the mystery is removed. Here, by way of toy example, we shed light on the basic components at the heart of every Bayesian model. The hope is that this illustration will make clear the basic claims of the Bayesian program.

Constructing a Bayesian model involves two steps. The first step is to specify the set of possibilities for the state of the world, which is referred to as the *hypothesis space*. Each hypothesis can be thought of as a candidate prediction by the subject about what future sensory information will be encountered. However, the term *hypothesis* should not be confused with its more traditional usage in psychology, connoting explicit testing of rules or other symbolically represented propositions. In the context of Bayesian modeling, hypotheses need have nothing to do with explicit reasoning, and indeed the Bayesian framework makes no commitment whatsoever on this issue. For example, in Bayesian models of visual processing, hypotheses can correspond to extremely low-level information, such as the presence of elementary visual features (contours, etc.) at various locations in the visual field (Geisler et al. 2001). There is also no commitment regarding where the hypotheses come from. Hypotheses could represent innate biases or knowledge, or they could have been learned previously by the individual. Thus, the framework has no position on nativist-empiricist debates. Furthermore, hypotheses representing very different types of information (e.g., a contour in a particular location, whether or not the image reminds you of your mother, whether the image is symmetrical, whether it spells a particular word, etc.) are all lumped together in a common hypothesis space and treated equally by the model. Hence, there is no distinction between different types of representations or knowledge systems within the brain. In general, a hypothesis is nothing more than a probability distribution. This distribution, referred to as the *likelihood function*, simply specifies how likely each possible pattern of observations is according to the hypothesis in question.

The second step in constructing a Bayesian model is to specify how strongly the subject believes in each hypothesis before observing data. This initial belief is expressed as a probability distribution over the hypothesis space, and is referred to as the *prior distribution* (or simply, *prior*). The prior can be thought of as an initial bias in favor of some hypotheses over others, in that it contributes extra "votes" (as elaborated in the next two paragraphs) that are independent of any actual data. This decisional bias allows the model's predictions to be shifted in any direction the modeler chooses regardless of the subject's observations. As we discuss in section 5, the prior can be a strong point of the model if it is derived from empirical

statistics of real environments. However, more commonly the prior is chosen ad hoc, providing substantial unconstrained flexibility to models that are advocated as rational and assumption-free.

Together, the hypotheses and the prior fully determine a Bayesian model. The model's goal is to decide how strongly to believe in each hypothesis after data have been observed. This final belief is again expressed as a probability distribution over the hypothesis space and is referred to as the *posterior distribution* (or *posterior*). The mathematical identity known as *Bayes' Rule* is used to combine the prior with the observed data to compute the posterior. Bayes' Rule can be expressed in many ways, but here we explain how it can be viewed as a simple vote-counting model. Specifically, Bayesian inference is equivalent to tracking evidence for each hypothesis, or votes for how strongly to believe in each hypothesis. The prior provides the initial evidence counts, E_{prior}, which are essentially made-up votes that give some hypotheses a head start over others before any actual data are observed. When data are observed, each observation adds to the existing evidence according to how consistent it is with each hypothesis. The evidence contributed for a hypothesis that predicted the observation will be greater than the evidence for a hypothesis under which the observation was unlikely. The evidence contributed by the ith observation, E_{data_i}, is simply added to the existing evidence to update each hypothesis's count. Therefore, the final evidence, $E_{\text{posterior}}$, is nothing more than a sum of the votes from all of the observations, plus the initial votes from the prior:[1]

$$E_{\text{posterior}}(H) = E_{\text{prior}}(H) + \sum_i E_{data_i}(H) \qquad (1)$$

This sum is computed for every hypothesis, H, in the hypothesis space. The vote totals determine how strongly the model believes in each hypothesis in the end. Thus, any Bayesian model can be viewed as summing evidence for each hypothesis, with initial evidence coming from the prior and additional evidence coming from each new observation. The final evidence counts are then used in whatever decision procedure is appropriate for the task, such as determining the most likely hypothesis, predicting the value of some unobserved variable (by weighting each hypothesis by its posterior probability and averaging their predictions), or choosing an action that maximizes the expected value of some outcome (again by weighted average over hypotheses). At its core, this is all there is to Bayesian modeling.

To illustrate these two steps and how inference proceeds in a Bayesian model, consider the problem of determining whether a fan entering a football stadium is rooting for the University of Southern California (USC) Trojans or the University of Texas (UT) Longhorns, based on three simple questions: (1) Do you live by the ocean? (2) Do you own a cowboy hat? (3) Do you like Mexican food? The first step is to specify the space of possibilities (i.e., hypothesis space). In this case the hypothesis space consists of two possibilities: being a fan of either USC or UT. Both of these hypotheses entail probabilities for the data we could observe, for example, P(ocean | USC) = .8 and P(ocean | UT) = .3. Once these probabilities are given, the two hypotheses are fully specified. The second step is to specify the prior. In many applications, there is no principled way of doing this, but in this example the prior corresponds to the probability that a randomly selected person will be a USC fan or a UT fan; that is, one's best guess as to the overall proportion of USC and UT fans in attendance.

With the model now specified, inference proceeds by starting with the prior and accumulating evidence as new data are observed. For example, if the football game is being played in Los Angeles, one might expect that most people are USC fans, and hence the prior would provide an initial evidence count in favor of USC. If our target person responded that he lives near the ocean, this observation would add further evidence for USC relative to UT. The magnitudes of these evidence values will depend on the specific numbers assumed for the prior and for the likelihood function for each hypothesis; but all that the model does is take the evidence values and add them up. Each new observation adds to the balance of evidence among the hypotheses, strengthening those that predicted it relative to those under which it was unlikely.

There are several ways in which real applications of Bayesian modeling become more complex than the foregoing simple example. However, these all have to do with the complexity of the hypothesis space rather than the Bayesian framework itself. For example, many models have a hierarchical structure in which hypotheses are essentially grouped into higher-level *overhypotheses*. Overhypotheses are generally more abstract and require more observations to discriminate among them; thus, hierarchical models are useful for modeling learning (e.g., Kemp et al. 2007). However, each overhypothesis is just a weighted sum of elementary hypotheses, and inference among overhypotheses comes down to exactly the same vote-counting scheme as described earlier. As a second example, many models assume special mathematical functions for the prior, such as conjugate priors (discussed further in sect. 6), that simplify the computations involved in updating evidence. However, such assumptions are generally made solely for the convenience of the modeler, rather than for any psychological reason related to the likely initial beliefs of a human subject. Finally, for models with especially complex hypothesis spaces, computing exact predictions often becomes computationally intractable. In these cases, sophisticated approximation schemes are used, such as Markov-chain Monte Carlo (MCMC) or particle filtering (i.e., sequential Monte Carlo). These algorithms yield good estimates of the model's true predictions while requiring far less computational effort. However, once again they are used for the convenience of the modeler and are not meant as proposals for how human subjects might solve the same computational problems. As we argue in section 6, all three of these issues are points where Bayesian modeling makes potential contact with psychological theory in terms of how information is represented and processed. Unfortunately, most of the focus to date has been on the Bayesian framework itself, setting aside where the hypotheses and priors come from and how the computations are performed or approximated.

The aim of this section has been to clear up confusion about the nature and theoretical claims of Bayesian models. To summarize: Hypotheses are merely probability distributions and have no necessary connection to explicit

reasoning. The model's predictions depend on the initial biases on the hypotheses (i.e., the prior), but the choice of the prior does not always have a principled basis. The heart of Bayesian inference – combining the prior with observed data to reach a final prediction – is formally equivalent to a simple vote-counting scheme. Learning and one-off decision-making both follow this scheme and are treated identically except for the timescale and specificity of hypotheses. The elaborate mathematics that often arises in Bayesian models comes from the complexity of their hypothesis sets or the tricks used to derive tractable predictions, which generally have little to do with the psychological claims of the researchers. Bayesian inference itself, aside from its assumption of optimality and close relation to vote-counting models, is surprisingly devoid of psychological substance. It involves no representations to be updated; no encoding, storage, retrieval, or search; no attention or control; no reasoning or complex decision processes; and in fact no mechanism at all, except for a simple counting rule.

4. Bayes as the new Behaviorism

Perhaps the most radical aspect of Bayesian Fundamentalism is its rejection of mechanism. The core assumption is that one can predict behavior by calculating what is optimal in any given situation. Thus, the theory is cast entirely at the computational level (in the sense of Marr 1982), without recourse to mechanistic (i.e., algorithmic or implementational) levels of explanation. As a metascientific stance, this is a very strong position. It asserts that a wide range of modes of inquiry and explanation are essentially irrelevant to understanding cognition. In this regard, the Bayesian program has much in common with Behaviorism. This section explores the parallels between these two schools of thought in order to draw out some of the limitations of Bayesian Fundamentalism.

During much of the first half of the 20th century, American psychology was dominated by the Behaviorist belief that one cannot draw conclusions about unobservable mental entities (Skinner 1938; Watson 1913). Under this philosophy, theories and experiments were limited to examination of the schedule of sensory stimuli directly presented to the subject and the patterns of observed responses. This approach conferred an important degree of rigor that the field previously lacked, by abolishing Dualism, advocating rigorous Empiricism, and eliminating poorly controlled and objectively unverifiable methods such as introspection. The strict Empiricist focus also led to discovery of important and insightful phenomena, such as shaping (Skinner 1958) and generalization (Guttman & Kalish 1956).

One consequence of the Behaviorist framework was that researchers limited themselves to a very constrained set of explanatory tools, such as conditioning and reinforcement. These tools have had an important lasting impact, for example, in organizational behavior management (Dickinson 2000) and behavioral therapy for a wide variety of psychiatric disorders (Rachman 1997). However, cognitive constructs, such as representation and information processing (e.g., processes associated with inference and decision-making), were not considered legitimate elements of a psychological theory. Consequently, Behaviorism eventually came under heavy criticism for its inability to account for many aspects of cognition, especially language and other higher-level functions (Chomsky 1959). After the so-called Cognitive Revolution, when researchers began to focus on the mechanisms by which the brain stores and processes information, the depth and extent of psychological theories were dramatically expanded (Miller 2003). Relative to the state of current cognitive psychology, Behaviorist research was extremely limited in the scientific questions that it addressed, the range of explanations it could offer, and the empirical phenomena it could explain.

The comparison of Bayesian modeling to Behaviorism may seem surprising considering that Bayesian models appear to contain unobservable cognitive constructs, such as hypotheses and their subjective probabilities. However, these constructs rarely have the status of actual psychological assumptions. Psychological theories of representation concern more than just what information is tracked by the brain; they include how that information is encoded, processed, and transformed. The Fundamentalist Bayesian view takes no stance on whether or how the brain actually computes and represents probabilities of hypotheses. All that matters is whether behavior is consistent with optimal action with respect to such probabilities (Anderson 1990; 1991b). This means of sidestepping questions of representation can be viewed as a strength of the rational approach, but it also means that Bayesian probabilities are not necessarily psychological beliefs. Instead, they are better thought of as tools used by the researcher to derive behavioral predictions. The hypotheses themselves are not psychological constructs either, but instead reflect characteristics of the environment. The set of hypotheses, together with their prior probabilities, constitute a description of the environment by specifying the likelihood of all possible patterns of empirical observations (e.g., sense data). According to Bayesian Fundamentalism, this description is an accurate one, and by virtue of its accuracy it is determined solely by the environment. There is no room for psychological theorizing about the nature of the hypothesis set, because such theories logically could only take the form of explaining how people's models of the environment are incorrect. According to Bayesian Fundamentalism, by grounding the hypotheses and prior in the environment (Anderson 1990), Bayesian models make predictions directly from the environment to behavior, with no need for psychological assumptions of any sort.

In many Bayesian models, the hypotheses are not expressed as an unstructured set, but instead emerge from a *generative model* of the environment. The generative model (which is a component of the Bayesian model) often takes the form of a causal network, in which the probabilities of observable variables depend on the values of unobservable, latent variables. Hypotheses about observable variables correspond to values of the latent variables. For example, in the topic model of text comprehension, the words in a passage (the observables) are assumed to be generated by a stochastic process parameterized by the weights of various semantic topics within the passage (Griffiths et al. 2007). However, the model makes no claim about the psychological status of the latent variables (i.e., the topic weights). These variables serve only to define the joint distribution over all possible

word sequences, and the model is evaluated only with respect to whether human behavior is consistent with that distribution. Whether people explicitly represent topic weights (or their posterior distributions) or whether they arrive at equivalent inferences based on entirely different representations is outside the scope of the model (Griffiths et al. 2007, p. 212). Therefore, generative models and the latent variables they posit do not constitute psychological constructs, at least according to the fundamentalist viewpoint. Instead, they serve as descriptions of the environment and mathematical tools that allow the modeler to make behavioral predictions. Just as in Behaviorist theories, the path from environmental input to behavioral prediction bypasses any consideration of cognitive processing.

To take a simpler example, Figure 1 shows a causal graphical model corresponding to a simplified version of Anderson's (1991b) rational model of categorization. The subject's task in this example is to classify animals as birds or mammals. The rational model assumes that these two categories are each partitioned into subcategories, which are termed *clusters*. The psychological prediction is that classification behavior corresponds (at a computational level) to Bayesian inference over this generative model. If a subject were told that a particular animal can fly, the optimal probability that it is a bird would equal the sum of the posterior probabilities of all the clusters within the bird category (and likewise for mammal). Critically, however, the clusters do not necessarily correspond to actual psychological representations. All that matters for predicting behavior is the joint probability distribution over the observable variables (i.e., the features and category labels). The clusters help the modeler to determine this distribution, but the brain may perform the computations in a completely different manner. In the discussion of Bayesian Enlightenment below (sect. 6), we return to the possibility of treating latent variables and generative models as psychological assumptions about knowledge representation. However, the important point here is that, according to the Fundamentalist Bayesian view, they are not. Generative models, the hypotheses they specify, and probability distributions over those hypotheses are all merely tools for deriving predictions from a Bayesian model. The model itself exists at a computational level, where its predictions are defined only based on optimal inference and decision-making. The mechanisms by which those decisions are determined are outside the model's scope.

4.1. Consequences of the denial of mechanism

By eschewing mechanism and aiming to explain behavior purely in terms of rational analysis, the Fundamentalist Bayesian program raises the danger of pushing the field of psychology back toward the sort of restrictive state experienced during the strict Behaviorist era. Optimality and probabilistic inference are certainly powerful tools for explaining behavior, but taken alone they are insufficient. A complete science of cognition must draw on the myriad theoretical frameworks and sources of evidence bearing on how cognition is carried out, as opposed to just its end product. These include theories of knowledge representation, decision-making, mental models, dynamic-system approaches, attention, executive control, heuristics and biases, reaction time, embodiment, development, and the entire field of cognitive neuroscience,

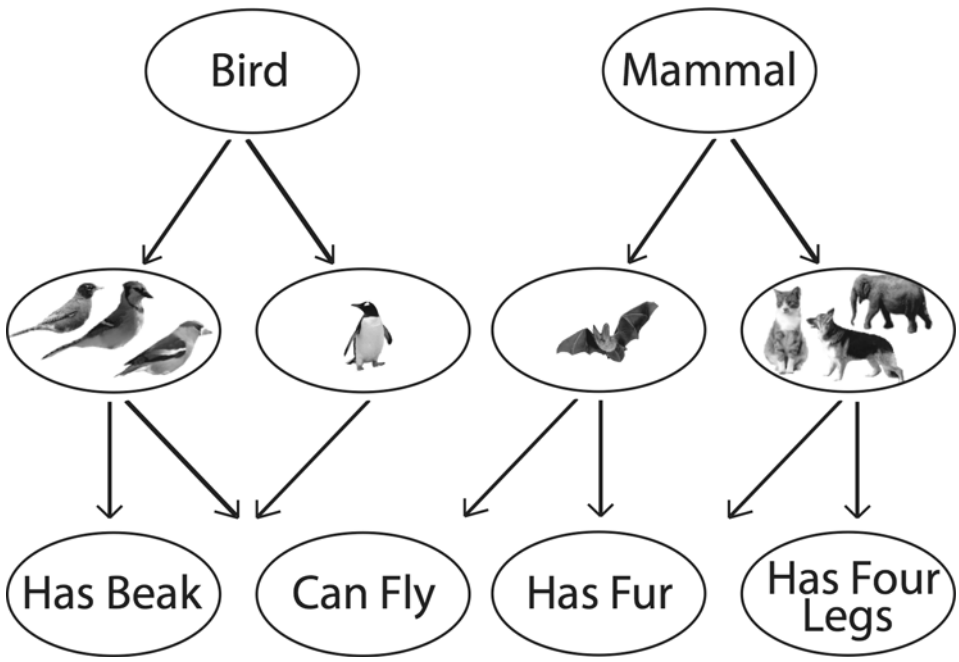

Figure 1. Simplified generative model based on Anderson's (1991b) rational model of categorization. Upper and lower nodes represent observable variables (category labels and features, respectively). Middle nodes represent clusters, which correspond to latent variables. Judgment of category membership based on feature information is assumed to be consistent with Bayesian inference over this probabilistic structure. However, the clusters only serve as mathematical constructs that determine the model's predictions. The question of whether clusters relate to actual psychological representations is outside the scope of the model. The original model treats category labels on par with features (so that clusters are not necessarily nested within categories), but this difference does not alter the point made here.

to name just a few. Many of these lines of research would be considered meaningless within the Behaviorist framework, and, likewise, they are all rendered irrelevant by the strict rational view. Importantly, the limitation is not just on what types of explanations are considered meaningful, but also on what is considered worthy of explanation – that is, what scientific questions are worth pursuing and what types of evidence are viewed as informative.

An important argument in favor of rational over mechanistic modeling is that the proliferation of mechanistic modeling approaches over the past several decades has led to a state of disorganization, wherein the substantive theoretical content of the models cannot be disentangled from the idiosyncrasies of their implementations. Distillation of models down to their computational principles would certainly aid in making certain comparisons across modeling frameworks. For example, both neural network (Burgess & Hitch 1999) and production system (Anderson et al. 1998) models of serial recall have explained primacy effects by using the same assumptions about rehearsal strategies, despite the significant architectural differences in which this common explanation is implemented. The rational approach is useful in this regard in that it eases comparison by emphasizing the computational problems that models aim to solve.

However, 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. As we argue in section 5, mechanistic explanations tend to be better suited for prediction of new phenomena, as opposed to post hoc explanation. Furthermore, algorithmic explanations and neural implementations are an important focus of research in their own right. Much can be learned from consideration of how the brain handles the computational challenge of guiding behavior efficiently and rapidly in a complex world, when optimal decision-making (to the extent that it is even well-defined) is not possible. These mechanistic issues are at the heart of most of the questions of theoretical or practical importance within cognitive science, including questions of representation, timing, capacity, anatomy, and pathology.

For example, connectionist models have proven valuable in reconceptualizing category-specific deficits in semantic memory as arising from damage to distributed representations in the brain (for a review, see Rogers & Plaut 2002), as opposed to being indicative of damage to localized representations (e.g., Caramazza & Shelton 1998). Although these insights rely on statistical analyses of how semantic features are distributed (e.g., Cree & McRae 2003), and, therefore, could in principle be characterized by a Bayesian model, the connectionist models were tremendously useful in motivating this line of inquiry. Additionally, follow-on studies have helped characterize impaired populations and have suggested interventions, including studies involving Alzheimer's patients (Devlin et al. 1998) and related work exploring reading difficulties resulting from developmental disorders and brain injury (Joanisse & Seidenberg 1999; 2003; Plaut et al. 1996).

Even when the goal is only to explain inference or choice behavior (setting aside reaction time), optimal probabilistic inference is not always sufficient. This is because the psychological mechanisms that give rise to behavior often at best only approximate the optimal solution. These mechanisms produce signature deviations from optimality that rational analysis has no way of anticipating. Importantly, considering how representations are updated in these mechanisms can suggest informative experiments.

For example, Sakamoto et al. (2008) investigated learning of simple perceptual categories that differed in the variation among items within each category. To classify new stimuli accurately, subjects had to estimate both the means and variances of the categories (stimuli varied along a single continuous dimension). Sakamoto et al. considered a Bayesian model that updates its estimates optimally, given all past instances of each category, and a mechanistic (cluster) model that learns incrementally in response to prediction error. The incremental model naturally produces *recency effects*, whereby more recent observations have a greater influence on its current state of knowledge (Estes 1957), in line with empirical findings in this type of task (e.g., Jones & Sieck 2003). Simple recency effects are no challenge to Bayesian models, because one can assume non-stationarity in the environment (e.g., Yu & Cohen 2008). However, the incremental model predicts a more complex recency effect whereby, under certain presentation sequences, the recency effect in the estimate of a category's mean induces a bias in the estimate of its variance. This bias arises purely as a by-product of the updating algorithm and has no connection to rational, computational-level analyses of the task. Human subjects exhibited the same estimation bias predicted by the incremental model, illustrating the utility of mechanistic models in directing empirical investigations and explaining behavior.

Departures from strict rational orthodoxy can lead to robust and surprising predictions, such as in work considering the forces that mechanistic elements exert on one another in learning and decision making (Busemeyer & Johnson 2008; Davis & Love 2010; Spencer et al. 2009). Such work often serves to identify relevant variables that would not be deemed theoretically relevant under a Fundamentalist Bayesian view (Clearfield et al. 2009). Even minimal departures from purely environmental considerations, such as manipulating whether information plays the role of cue or outcome within a learning trial, can yield surprising and robust results (Love 2002; Markman & Ross 2003; Ramscar et al. 2010; Yamauchi & Markman 1998). The effects of this manipulation can be seen in a common transfer task, implying that it is the learners' knowledge that differs and not just their present goals.

Focusing solely on computational explanations also eliminates many of the implications of cognitive science for other disciplines. For example, without a theory of the functional elements of cognition, little can be said about cognitive factors involved in psychological disorders. Likewise, without a theory of the physiology of cognition, little can be said about brain disease, trauma, or psychopharmacology. (Here the situation is even more restrictive than in Behaviorism, which would accept neurological data as valid and useful.) Applications of cognitive theory also tend to depend strongly on mechanistic descriptions of the mind. For example, research in human factors relies on models of timing and processing capacity, and applications to real-world decision-making depend on the heuristics underlying human judgment. Understanding these

heuristics can also lead to powerful new computational algorithms that improve the performance of artificially intelligent systems in complex tasks (even systems built on Bayesian architectures). Rational analysis provides essentially no insight into any of these issues.

4.2. Integration and constraints on models

One advantage of Behaviorism is that its limited range of explanatory principles led to strong cohesion among theories of diverse phenomena. For example, Skinner (1957) attempted to explain human verbal behavior by using the same principles previously used in theories of elementary conditioning. It might be expected that the Bayesian program would enjoy similar integration because of its reliance on the common principles of rational analysis and probabilistic inference. Unfortunately, this is not the case in practice, because the process of rational analysis is not sufficiently constrained, especially as applied to higher-level cognition.

Just as mechanistic modeling allows for alternative assumptions about process and representation, rational modeling allows for alternative assumptions about the environment in which the cognitive system is situated (Anderson 1990). In both cases, a principal scientific goal is to decide which assumptions provide the best explanation. With Bayesian models, the natural approach dictated by rational analysis is to make the generative model faithful to empirical measurements of the environment. However, as we observe in section 5, this empirical grounding is rarely carried out in practice. Consequently, the rational program loses much of its principled nature, and models of different tasks become fractionated because there is nothing but the math of Bayesian inference to bind them together.

At the heart of every Bayesian model is a set of assumptions about the task environment, embodied by the hypothesis space and prior distribution, or, equivalently, by the generative model and prior distributions for its latent variables. The prior distribution is the well-known and oft-criticized lack of constraint in most Bayesian models. As explained in section 3, the prior provides the starting points for the vote-counting process of Bayesian inference, thereby allowing the model to be initially biased towards some hypotheses over others. Methods have been developed for using uninformative priors that minimize influence on model predictions, such as Jeffreys priors (Jeffreys 1946) or maximum-entropy priors (Jaynes 1968). However, a much more serious source of indeterminacy comes from the choice of the hypothesis set itself, or equivalently, from the choice of the generative model.

The choice of generative model often embodies a rich set of assumptions about the causal and dynamic structure of the environment. In most interesting cases, there are many alternative assumptions that could be made, but only one is considered. For example, the CrossCat model of how people learn multiple overlapping systems of categories (Shafto et al., in press) assumes that category systems constitute different partitions of a stimulus space, that each category belongs to exactly one system, and that each stimulus feature or dimension is relevant to exactly one category system and is irrelevant to all others. These assumptions are all embodied by the generative model on which CrossCat is based. There are clearly alternatives to these assumptions, for which intuitive arguments can be made (e.g., for clothing, the color dimension is relevant for manufacturing, laundering, and considerations of appearance), but there is no discussion of these alternatives, justification of the particular version of the model that was evaluated, or consideration of the implications for model predictions. Other than the assumption of optimal inference, all there is to a Bayesian model is the choice of generative model (or hypothesis set plus prior), so it is a serious shortcoming when a model is developed or presented without careful consideration of that choice. The neglected multiplicity of models is especially striking considering the rational theorist's goal of determining the – presumably unique – optimal pattern of behavior.

Another consequence of insufficient scrutiny of generative models (or hypothesis sets more generally) is a failure to recognize the psychological commitments they entail. These assumptions often play a central role in the explanation provided by the Bayesian model as a whole, although that role often goes unacknowledged. Furthermore, the psychological assumptions implicitly built into a generative model can be logically equivalent to pre-existing theories of the same phenomena. For example, Kemp et al. (2007) propose a Bayesian model of the shape bias in early word learning, whereby children come to expect a novel noun to be defined by the shape of the objects it denotes, rather than other features such as color or texture. The model learns the shape bias through observation of many other words with shape-based definitions, which shifts evidence to an overhypothesis that most nouns in the language are shape-based. The exposition of the model is a mathematically elegant formalization of abstract induction. However, it is not Bayes' Rule or even the notion of overhypotheses that drives the prediction; rather it is the particular overhypotheses that were built into the model. In other words, the model was endowed with the capability to recognize a particular pattern (viz., regularity across words in which perceptual dimensions are relevant to meaning), so the fact that it indeed recognizes that pattern when presented with it is not surprising or theoretically informative. Furthermore, the inference made by the model is logically the same as the notion of second-order generalization proposed previously by Linda Smith and colleagues (e.g., Smith et al. 2002). Detailed mechanistic modeling has shown how second-order generalization can emerge from the interplay between attentional and associative processes (Colunga & Smith 2005), in contrast to the more tautological explanation offered by the Bayesian model. Therefore, at the level of psychological theory, Kemp et al.'s (2007) model merely recapitulates a previously established idea in a way that is mathematically more elegant but psychologically less informative.

In summary, Bayesian Fundamentalism is simultaneously more restrictive and less constrained than Behaviorism. In terms of modes of inquiry and explanation, both schools of thought shun psychological constructs, in favor of aiming to predict behavior directly from environmental inputs. However, under Behaviorism this restriction was primarily a technological one. Nothing in the Behaviorist philosophy would invalidate relatively recent tools that enable direct measurements of brain function, such as neuroimaging, EEG, and single-unit recording (at least as targets of explanation, if not as tools through

which to develop theories of internal processes). Indeed, these techniques would presumably have been embraced, as they satisfy the criterion of direct observation. Bayesian Fundamentalism, in contrast, rejects all measures of brain processing out of principle, because only the end product (i.e., behavior) is relevant to rational analysis.[2] At the same time, whereas Behaviorist theories were built from simple mechanisms and minimal assumptions, Bayesian models often depend on complex hypothesis spaces based on elaborate and mathematically complex assumptions about environmental dynamics. As the emphasis is generally on rational inference (i.e., starting with the assumptions of the generative model and deriving optimal behavior from there), the assumptions themselves generally receive little scrutiny. The combination of these two factors leads to a dangerously under-constrained research program, in which the core assumptions of a model (i.e., the choice of hypothesis space) can be made at the modeler's discretion, without comparison to alternatives and without any requirement to fit physiological or other process-level data.

5. Bayes as evolutionary psychology

In addition to the rejection of mechanistic explanation, a central principle of the Fundamentalist Bayesian approach to cognition is that of optimality. The claim that human behavior can be explained as adaptation to the environment is also central to evolutionary psychology. On the surface, these two approaches to understanding behavior seem very different, as their content and methods differ. For example, one core domain of inquiry in evolutionary psychology is mating, which is not often studied by cognitive psychologists, and theories in evolutionary psychology tend not to be computational in nature, whereas rational Bayesian approaches are by definition. Thus, one advantage of rational Bayesian accounts is that they formalize notions of optimality, which can clarify assumptions and allow for quantitative evaluation. Despite these differences, Bayesian Fundamentalism and evolutionary psychology share a number of motivations and assumptions. Indeed, Geisler and Diehl (2003) propose a rational Bayesian account of Darwin's theory of natural selection. In this section, we highlight the commonalities and important differences between these two approaches to understanding human behavior.

We argue that Bayesian Fundamentalism is vulnerable to many of the criticisms that have been leveled at evolutionary psychology. Indeed, we argue that notions of optimality in evolutionary psychology are more complete and properly constrained than those forwarded by Bayesian Fundamentalists, because evolutionary psychology considers other processes than simple adaptation (e.g., Buss et al. 1998). Bayesian Fundamentalism appropriates some concepts from evolutionary psychology (e.g., adaptation, fitness, and optimality), but leaves behind many other key concepts because of its rejection of mechanism. Because it is mechanisms that evolve, not behaviors, Bayesian Fundamentalism's assertions of optimality provide little theoretical grounding and are circular in a number of cases.

Basic evolutionary theory holds that animal behavior is adapted by natural selection, which increases inclusive fitness. High fitness indicates that an animal's behaviors are well suited to its environment, leading to reproductive success. On the assumption that evolutionary pressures tune a species' genetic code such that the observed phenotype gives rise to optimal behaviors, one can predict an animal's behavior by considering the environment in which its ancestors flourished and reproduced. According to evolutionary psychologists, this environment, referred to as the Environment of Evolutionary Adaptedness (EEA), must be understood in order to comprehend the functions of the brain (Bowlby 1969). Thus, evolutionary explanations of behavior tend to focus on the environment – a focus that can occur at the expense of careful consideration of mechanism. However, as discussed extensively further on in section 5.3 and in contrast to Bayesian Fundamentalism, some key concepts in evolutionary psychology do rely on mechanistic considerations, and these concepts are critical for grounding notions of adaptation and optimization. These key concepts are neglected in Bayesian Fundamentalism.

Critically, it is not any function that is optimized by natural selection, but only those functions that are relevant to fitness. To use Oaksford and Chater's (1998a) example, animals may be assumed to use optimal foraging strategies because (presumably) gathering food efficiently is relevant to the global goal of maximizing inclusive fitness (see Hamilton 1964). Thus, in practice, evolutionary arguments, like rational theories of cognition, require specification of the environment and the behaviors that increase fitness. For example, Anderson's (1991b) rational model of category learning is intended to maximize prediction of unknown information in the environment, a behavior that presumably increases fitness.

Like rational approaches to cognition, evolutionary psychology draws inspiration from evolutionary biology and views much of human behavior as resulting from adaptations shaped by natural selection (Buss 1994; Pinker 2002; Tooby & Cosmides 2005). The core idea is that recurring challenges in our ancestral environments (i.e., EEA) shaped our mental capacities and proclivities. This environmental focus is in the same spirit as work in ecological psychology (Gibson 1979; Michaels & Carello 1981). Following from a focus on specific challenges and adaptations, evolutionary theories often propose special-purpose modules. For example, evolutionary psychologists have proposed special-purpose modules for cheater detection (Cosmides & Tooby 1992), language acquisition (Pinker 1995), incest avoidance (Smith 2007), and snake detection (Sperber & Hirschfeld 2003). Much like evolutionary psychology's proliferation of modules, rational models are developed to account for specific behaviors, such as children's ability to give the number of objects requested (Lee & Sarnecka 2010), navigation when disoriented in a maze (Stankiewicz et al. 2006), and understanding a character's actions in an animation (Baker et al. 2009), at the expense of identifying general mechanisms and architectural characteristics (e.g., working memory) that are applicable across a number of tasks (in which the specific behaviors to be optimized differ).

5.1. An illustrative example of rational analysis as evolutionary argument

Perhaps the rational program's focus on environmental adaptation is best exemplified by work in early vision. Early vision is a good candidate for rational investigation,

as the visual environment has likely been stable for millennia and the ability to perceive the environment accurately is clearly related to fitness. The focus on environmental statistics is clear in Geisler et al.'s (2001) work on contour detection. In this work, Geisler and colleagues specify how an ideal classifier detects contours and compare this ideal classifier's performance to human performance. To specify the ideal classifier, the researchers gathered natural image statistics that were intended to be representative of the environment in which our visual system evolved. Implicit in the choice of images are assumptions about what the environment was like. Additionally, the analysis requires assuming which measures or image statistics are relevant to the contour classification problem.

Geisler et al. selected a number of natural images of mountains, forests, coastlines, and so forth, to characterize our ancestral visual environment. From these images, they measured certain statistics they deemed relevant to contour detection. Their chosen measures described relationships among edge segments belonging to the same contour, such as the distance between the segments and their degree of colinearity. To gather these statistics, expert raters determined whether two edge elements belonged to the same contour in the natural images. These measures specify the likelihood and prior in the Bayesian ideal observer. The prior for the model is simply the probability that two randomly selected edge elements belong to the same contour. The likelihood follows from a table of co-occurrences of various distances and angles between pairs of edge elements indexed by whether each pair belongs to the same contour. Geisler et al. compared human performance to the ideal observer in a laboratory task that involved determining whether a contour was present in novel, meaningless images composed of scattered edge elements. Human performance and the rational model closely corresponded, supporting Geisler et al.'s account.

Notice that there is no notion of *mechanism* (i.e., process or representation) in this account of contour detection. The assumptions made by the modeler include what our ancestral environment was like and which information in this environment is relevant. Additionally, it is assumed that the specific behavior modeled (akin to a module in evolutionary psychology) is relevant to fitness. These assumptions, along with demonstrating a correlation with human performance, are the intellectual contribution of the work. Finally, rational theories assume optimal inference as reflected in the Bayesian classification model. Specifying the Bayesian model may be technically challenging, but is not part of the theoretical contribution (i.e., it is a math problem, not a psychology problem). The strength of Geisler et al.'s (2001) work rests in its characterization of the environment and the statistics of relevance.

Unfortunately, the majority of rational analyses do not include any measurements from actual environments, despite the fact that the focus of such theories is on the environment (for a similar critique, see Murphy 1993). Instead, the vast majority of rational analysis in cognition relies on intuitive arguments to justify key assumptions. In some cases, psychological phenomena can be explained from environmental assumptions that are simple and transparent enough not to require verification (e.g., McKenzie & Mikkelsen 2007; Oaksford & Chater 1994). However, more often, Bayesian models incorporate complex and detailed assumptions about the structure of the environment that are far from obvious and are not supported by empirical data (e.g., Anderson 1991b; Brown & Steyvers 2009; Goodman et al. 2008b; Steyvers et al. 2009; Tenenbaum & Griffiths 2001). Cognitive work that does gather environmental measures is exceedingly rare and tends to rely on basic statistics to explain general behavioral tendencies and judgments (e.g., Anderson & Schooler 1991; Griffiths & Tenenbaum 2006). This departure from true environmental grounding can be traced back to John Anderson's (1990; 1991b) seminal contributions in which he popularized the rational analysis of cognition. In those works, he specified a series of steps for conducting such analyses. Step 6 of the rational method (Anderson 1991b) is to revisit assumptions about the environment and relevant statistics when the model fails to account for human data. In practice, this step involves the modeler's ruminating on what the environment is like and what statistics are relevant, rather than actual study of the environment. This is not surprising given that most cognitive scientists are not trained to characterize ancestral environments. For example, at no point in the development of Anderson's (1991b) rational model of category learning is anything in the environment actually measured. Although one purported advantage of rational analysis is the development of zero-parameter, non-arbitrary models, it would seem that the theorist has unbounded freedom to make various assumptions about the environment and the relevant statistics (see Sloman & Fernbach [2008] for a similar critique). As discussed in the next section, similar criticisms have been made of evolutionary psychology.

5.2. Too much flexibility in evolutionary and rational explanations?

When evaluating any theory or model, one must consider its fit to the data and its flexibility to account for other patterns of results (Pitt et al. 2002). Models and theories are favored that fit the data and have low complexity (i.e., are not overly flexible). One concern we raise is whether rational approaches offer unbounded and hidden flexibility to account for any observed data. Labeling a known behavior as "rational" is not theoretically significant if it is always possible for some rational explanation to be constructed. Likewise, evolutionary psychology is frequently derided as simply offering "just so" stories (Buller 2005, but see Machery & Barrett 2006). Adaptationist accounts certainly provide constraint on explanation compared to non-adaptationist alternatives, but taken alone they still allow significant flexibility in terms of assumptions about the environment and the extent to which adaptation is possible. For example, to return to the foraging example, altering one's assumptions about how food rewards were distributed in ancestral environments can determine whether an animal's search process (i.e., the nature and balance of exploitative and exploratory decisions) is optimal. Likewise, the target of optimization can be changed. For example, inefficiencies in an animal's foraging patterns for food-rich environments can be explained after the fact as an adaptation to ensure the animal does not become morbidly obese. On

the other hand, if animals were efficient in abundant environments and became obese, one could argue that foraging behaviors were shaped by adaptation to environments in which food was not abundant. If, no matter the data, there is a rational explanation for a behavior, it is not a contribution to label a behavior as rational. Whereas previous work in the heuristics-and-biases tradition (Tversky & Kahneman 1974) cast the bulk of cognition as irrational using a fairly simplistic notion of rationality, Bayesian Fundamentalism finds rationality to be ubiquitous based on under-constrained notions of rationality.

To provide a recent example from the literature, the persistence of negative traits, such as anxiety and insecurity that lower an individual's fitness, has been explained by appealing to these traits' utility to the encompassing group in signaling dangers and threats facing the group (Ein-Dor et al. 2010). While this ingenious explanation could be correct, it illustrates the incredible flexibility that adaptive accounts can marshal in the face of a challenging data point.

Similar criticisms have been leveled at work in evolutionary biology. For example, Gould and Lewontin (1979) have criticized work that develops hypotheses about the known functions of well-studied organs as "backward-looking." One worry is that this form of theorizing can lead to explanations that largely reaffirm what is currently believed. Work in evolutionary psychology has been criticized for explaining unsurprising behaviors (Horgan 1999), such as, that men are less selective about who they will mate with than are women. Likewise, we see a tendency for rational analyses to largely re-express known findings in the language of Bayesian optimal behavior. The work of Geisler et al. (2001) on contour perception is vulnerable to this criticism as it largely recapitulates Gestalt principles (e.g., Wertheimer 1923/1938) in the language of Bayes. In cognition, the rational rules model (Goodman et al. 2008b) of category learning reflects many of the intuitions of previous models, such as the rule-plus-exception (RULEX) model (Nosofsky et al. 1994), in a more elegant and expressive Bayesian form that does not make processing predictions. In other cases, the intuitions from previous work are re-expressed in more general Bayesian terms in which particular choices for the priors allow the Bayesian model to mimic the behavior of existing models. For example, unsupervised clustering models using simplicity principles based on minimum description length (MDL; Pothos & Chater 2002) are recapitulated by more flexible approaches phrased in the language of Bayes (Austerweil & Griffiths 2008; Griffiths et al. 2008b). A similar path of model development has occurred in natural language processing (Ravi & Knight 2009).

One motivation for rational analysis was to prevent models with radically different assumptions from making similar predictions (Anderson 1991b). In reality, the modeler has tremendous flexibility in characterizing the environment (see Buller [2005] for similar arguments). For example, the studies by Dennis and Humphreys (1998) and Shiffrin and Steyvers (1998) both offer rational accounts of memory (applicable to word-list tasks) that radically differ, but both do a good job with the data and are thought-provoking. According to the rational program, analysis of the environment and the task should provide sufficient grounding to constrain theory development. Cognitive scientists (especially those trained in psychology) are not expert in characterizing the environment in which humans evolved, and it is not always clear what this environment was like. As in experimental sciences, our understanding of past environments is constantly revised, rather than providing a bedrock from which to build rational accounts of behavior. Adding further complexity, humans can change the environment to suit their needs rather than adapt to it (Kurz & Tweney 1998).

One factor that provides a number of degrees of freedom to the rational modeler is that it is not clear which environment (in terms of when and where) is evolutionarily relevant (i.e., for which our behavior was optimized). The relevant environment for rational action could be the local environment present in the laboratory task, similar situations (however defined) that the person has experienced, all experiences over the person's life, all experiences of our species, all experiences of all ancestral organisms traced back to single cell organisms, and so on. Furthermore, once the relevant environment is specified and characterized, the rational theorist has considerable flexibility in characterizing which relevant measures or statistics from the environment should enter into the optimality calculations. When considered in this light, the argument that rational approaches are parameter-free and follow in a straightforward manner from the environment is tenuous at best.

5.3. Optimization occurs over biological mechanisms, not behaviors

It is non-controversial that many aspects of our behavior are shaped by evolutionary processes. However, evolutionary processes do not directly affect behavior, but instead affect the mechanisms that give rise to behavior when coupled with environmental input (McNamara & Houston 2009). Assuming one could properly characterize the environment, focusing solely on how behavior should be optimized with respect to the environment is insufficient, as the physical reality of the brain and body is neglected. Furthermore, certain aspects of behavior, such as the time to execute some operation (e.g., the decision time to determine whether a person is a friend or foe), are closely linked to mechanistic considerations.

Completely sidestepping mechanistic considerations when considering optimality leads to absurd conclusions. To illustrate, it may not be optimal or evolutionarily advantageous to ever age, become infertile, and die; but these outcomes are universal and follow from biological constraints. It would be absurd to seriously propose an optimal biological entity that is not bounded by these biological and physical realities, but this is exactly the reasoning Bayesian Fundamentalists follow when formulating theories of cognition. Certainly, susceptibility to disease and injury impact inclusive fitness more than many aspects of cognition do. Therefore, it would seem strange to assume that human cognition is fully optimized while these basic challenges, which all living creatures past and present face, are not. Our biological reality, which is ignored by Bayesian Fundamentalists, renders optimal solutions – defined solely in terms of choice behavior – unrealistic and fanciful for many challenges.

Unlike evolutionary approaches, rational approaches to cognition, particularly those in the Bayesian Fundamentalist tradition, do not address the importance of mechanism in the adaptationist story. Certain physical limitations and realities lead to the prevalence of certain designs. Which design prevails is determined in part by these physical realities and the contemporaneous competing designs in the gene pool. As Marcus (2008) reminds us, evolution is the survival of the best current design, not survival of the globally optimal design. Rather than the globally optimal design winning out, often a locally optimal solution (i.e., a design better than similar designs) prevails (Dawkins 1987; Mayr 1982). Therefore, it is important to consider the trajectory of change of the mechanism (i.e., current and past favored designs), rather than to focus exclusively on which design is globally optimal.

As Marcus (2008) notes, many people are plagued with back pain because the human spine is adapted from animals that walk on four paws, not two feet. This is clearly not the globally optimal design, indicating that the optimization process occurs over constraints not embodied in rational analyses. The search process for the best design is hampered by the set of current designs available. These current designs can be adapted by descent-with-modification, but there is no purpose or forethought to this process (i.e., there is no intelligent designer). It simply might not be possible for our genome to code for shock absorbers like those in automobiles, given that the current solution is locally optimal and distant from the globally optimal solution. In the case of the human spine, the current solution is clearly not globally optimal, but is good enough to get the job done. The best solution is not easily reachable and might never be reached. If evolution settles on such a bad design for our spine, it seems unlikely that aspects of cognition are fully optimized. Many structures in our brains share homologs with other species. Structures more prominent in humans, such as the frontal lobes, were not anticipated, but like the spine, resulted from descent-with-modification (Wood & Grafman 2003).

The spine example makes clear that the history of the mechanism plays a role in determining the present solution. Aspects of the mechanism itself are often what is being optimized rather than the resulting behavior. For example, selection pressures will include factors such as how much energy certain designs require. The human brain consumes 25% of a person's energy, yet accounts for only 2% of a person's mass (Clark & Sokoloff 1999). Such non-behavioral factors are enormously important to the optimization process, but are not reflected in rational analyses, as these factors are tied to a notion of mechanism, which is absent in rational analyses. Any discussion of evolution optimizing behavior is incomplete without consideration of the mechanism that generates the behavior. To provide an example from the study of cognition, in contrast to Anderson's (1991b) rational analysis of concepts solely in terms of environmental prediction, concepts might also serve other functions, such as increasing "cognitive economy" in limited-capacity memory systems that would otherwise be swamped with details (Murphy 1993; Rosch 1978).

The notion of incremental improvement of mechanisms is also important because it is not clear that globally optimal solutions are always well defined. The optimality of Bayesian inference is well supported in "small worlds" in which an observer can sensibly assign subjective probabilities to all possible contingencies (Savage 1954). However, Binmore (2009) argues that proponents of Bayesian rationality overextend this reasoning when moving from laboratory tasks to the natural world. Normative support for the Bayesian framework breaks down in the latter case because, in an unconstrained environment, there is no clear rational basis for generating prior probabilities. Evolutionary theory does not face this problem because it relies on incremental adjustment rather than global optimization. Furthermore, shifting focus to the level of mechanism allows one to study the relative performance of those mechanisms without having to explicitly work out the optimal pattern of behavior in a complex environment (Gigerenzer & Todd 1999).

The preceding discussion assumes that we are optimized in at least a local sense. This assumption is likely invalid for many aspects of the mechanisms that give rise to behavior. Optimization by natural selection is a slow process that requires consistent selective pressure in a relatively stable environment. Many of the behaviors that are considered uniquely human are not as evolutionarily old as basic aspects of our visual system. It is also not clear how stable the relevant environment has been. To provide one example, recent simulations support the notion that many syntactic properties of language cannot be encoded in a language module, and that the genetic basis of language use and acquisition could not coevolve with human language (Chater et al. 2009).

Finally, while rational theorists focus on adaptation in pursuit of optimality, evolutionary theorists take a broader view of the products of evolution. Namely, evolution yields three products: (1) adaptations, (2) by-products, and (3) noise (Buss et al. 1998). An *adaptation* results from natural selection to solve some problem, whereas a *by-product* is the consequence of some adaptation. To use Bjorklund and Pelligrini's (2000) example, the umbilical cord is an adaptation, whereas the belly button is a by-product. Noise includes random effects resulting from mutations, drift, and so on. Contrary to the rational program, one should not take all behaviors and characteristics of people to be adaptations that increase (i.e., optimize) fitness.

5.4. Developmental psychology and notions of capacity limitation: What changes over time?

Although rational Bayesian modeling has a large footprint in developmental psychology (Kemp et al. 2007; Sobel et al. 2004; Xu & Tenenbaum 2007b), development presents basic challenges to the rational approach. One key question for any developmental model is what develops. In rational models, the answer is that nothing develops. Rational models are mechanism-free, leaving only information sampled to change over time. Although some aspects of development are driven by acquisition of more observations, other aspects of development clearly reflect maturational changes in the mechanism (see Xu & Tenenbaum 2007b, p. 169). For example, some aspects of children's performance are indexed by prefrontal development (Thompson-Schill et al. 2009) rather than the degree of experience within a domain. Likewise, teenage boys' interest in certain stimuli is likely

attributable more to hormonal changes than to collecting examples of certain stimuli and settling on certain hypotheses.

These observations put rational theories of development in a difficult position. People's mental machinery clearly changes over development, but no such change occurs in a rational model. One response has been to posit rational theories that are collections of discrepant causal models (i.e., hypothesis spaces). Each discrepant model is intended to correspond to a different stage of development (Goodman et al. 2006; Lucas et al. 2009). In effect, development is viewed as consisting of discrete stages, and a new model is proposed for each qualitative developmental change. Model selection is used to determine which discrepant model best accounts for an individual's current behavior. Although this approach may be useful in characterizing an individual's performance and current point in development, it does not offer any explanation for the necessity of the stages or why developmental transitions occur. Indeed, rather than accounts of developmental processes, these techniques are best viewed as methods to assess a person's conceptual model, akin to user modeling in tutoring systems (Conati et al. 1997). To the extent that the story of development is the story of mechanism development, rational theories have little to say (e.g., Xu & Tenenbaum 2007b).

Epigenetic approaches ease some of these tensions by addressing how experience influences gene expression over development, allowing for bidirectional influences between experience and genetic activity (Gottlieb 1992; Johnson 1998). One complication for rational theories is the idea that different selection pressures are exerted on organisms at different points in development (Oppenheim 1981). For adults, rigorous play wastes energy and is an undue risk, but for children, rigorous play may serve a number of adaptive functions (Baldwin & Baldwin 1977). For example, play fighting may prepare boys for adult hunting and fighting (Smith 1982). It would seem that different rational accounts are needed for different periods of development.

Various mental capacities vary across development and individuals. In adult cognition, Herbert Simon introduced the notion of *bounded rationality* to take into account, among other things, limitations in memory and processing capacities (see Simon 1957a). One of the proposals that grew out of bounded rationality was optimization under constraints, which posits that people may not perform optimally in any general sense, but, if their capacities could be well characterized, people might be found to perform optimally, given those limitations (e.g., Sargent 1993; Stigler 1961). For instance, objects in the environment may be tracked optimally, given sensory and memory limitations (Vul et al. 2009).

Although the general research strategy based on bounded rationality can be fruitful, it severely limits the meaning of labeling a behavior as rational or optimal. Characterizing capacity limitations is essentially an exercise in characterizing the mechanism, which represents a departure from rational principles. Once all capacity limitations are detailed, notions of rationality lose force. To provide a perverse example, each person can be viewed as an optimal version of himself given his own limitations, flawed beliefs, motivational limitations, and so on. At such a point, it is not clear what work the rational analysis is doing. Murphy (1993) makes a similar argument about the circularity of rational explanations: Animals are regarded as optimal with respect to their ecological niche, but an animal's niche is defined by its behaviors and abilities. For example, if one assumes that a bat's niche involves flying at night, then poor eyesight is not a counterexample of optimality.

Although these comments may appear negative, we do believe that considering capacity limitations is a sound approach that can facilitate the unification of rational and mechanistic approaches. However, we have doubts as to the efficacy of current approaches to exploring capacity limitations. For example, introducing capacity limitations by altering sampling processes through techniques like the particle filter (Brown & Steyvers 2009) appears to be motivated more by modeling convenience than by examination of actual cognitive mechanisms. It would be a curious coincidence if existing mathematical estimation techniques just happened to align with human capacity limitations. In section 6, we consider the possibility of using (mechanistic) psychological characterizations of one or more aspects of the cognitive system to derive bounded-optimality characterizations of decision processes. Critically, the potential of such approaches lies in the mutual constraint of mechanistic and rational considerations, as opposed to rational analysis alone.

To return to development, one interesting consideration is that reduced capacity at certain points in development is actually seen as a benefit by many researchers. For example, one proposal is that children's diminished working-memory capacity may facilitate language acquisition by encouraging children to focus on basic regularities (Elman 1993; Newport 1990). "Less is more" theories have also been proposed in the domain of metacognition. For example, children who overestimate their own abilities may be more likely to explore new tasks and be less self-critical in the face of failure (Bjorklund & Pellegrini 2000). Such findings seem to speak to the need to consider the nature of human learners, rather than the nature of the environment. Human learners do not seem to "turn off" a harmful capacity to narrow the hypothesis space when it might be prove beneficial to do so.

6. The role of Bayesian modeling in cognitive science

The observations in the preceding sections suggest that, although Bayesian modeling has great potential to advance our understanding of cognition, there are several conceptual problems with the Fundamentalist Bayesian program that limit its potential theoretical contributions. One possible reason is that most current work lacks a coherent underlying philosophy regarding just what that contribution should be. In this section, we lay out three roles for Bayesian modeling in cognitive science that potentially avoid the problems of the fundamentalist approach and that better integrate with other modes of inquiry. We make no strong commitment that any of the approaches proposed in this section will succeed, but we believe these are the viable options if one wants to use Bayes' Rule or probabilistic inference as a component of psychological theory.

First, Bayesian inference has proven to be exceedingly valuable as an analysis tool for deciding among scientific hypotheses or models based on empirical data. We refer to such approaches as Bayesian Agnosticism, because they take no stance on whether Bayesian inference is itself a useful psychological model. Instead, the focus is on using Bayesian inference to develop model-selection techniques that are sensitive to true model complexity and that avoid many of the logical inconsistencies of frequentist hypothesis testing (e.g., Pitt et al. 2002; Schwarz 1978).

Second, Bayesian models can offer computational-level theories of human behavior that bypass questions of cognitive process and representation. In this light, Bayesian analysis can serve as a useful starting point when investigating a new domain, much like how ideal-observer analysis can be a useful starting point in understanding a task, and thus assist in characterizing human proficiency in the task. This approach is in line with the Fundamentalist Bayesian philosophy, but, as the observations of the previous sections make clear, several changes to current common practice would greatly improve the theoretical impact of computational-level Bayesian modeling. Foremost, rational analysis should be grounded in empirical measurement of the environment. Otherwise, the endeavor is almost totally unconstrained. Environmental grounding has yielded useful results in low-level vision (Geisler et al. 2001) and basic aspects of memory (Anderson & Schooler 1991), but the feasibility of this approach with more complex cognitive tasks remains an open question. Furthermore, researchers are faced with the questions of what is the relevant environment (that behavior is supposedly optimized with respect to) and what are the relevant statistics of that environment (that behavior is optimized over). There is also the question of the objective function that is being optimized, and how that objective might vary according to developmental trajectory or individual differences (e.g., sex or social roles). It may be impossible in cases to specify what is optimal in any general sense without considering the nature of the mechanism. All of these questions can have multiple possible answers, and finding which answers lead to the best explanation of the data is part of the scientific challenge. Just as with mechanistic models, competing alternatives need to be explicitly recognized and compared. Finally, an unavoidable limitation of the pure rational approach is that behavior is not always optimal, regardless of the choice of assumptions about the environment and objective function. Evolution works locally rather than globally, and many aspects of behavior may be by-products rather than adaptations in themselves. More importantly, evolution is constrained by the physical system (i.e., the body and brain) that is being optimized. By excluding the brain from psychological theory, Bayesian Fundamentalism is logically unable to account for mechanistic constraints on behavior and unable to take advantage of or inform us about the wealth of data from areas such as neurophysiology, development, or timing.[3]

Third, rather than putting all the onus on rational analysis by attempting to explain behavior directly from the environment, one could treat various elements of Bayesian models as psychological assumptions subject to empirical test. This approach, which we refer to as Bayesian Enlightenment, seems the most promising, because it allows Bayesian models to make contact with the majority of psychological research and theory, which deals with mechanistic levels of analysis. The remainder of this section explores several avenues within Bayesian Enlightenment. We emphasize up front that all of these directions represent significant departures from the Fundamentalist Bayesian tenet that behavior can be explained and understood without recourse to process or representation.

6.1. Bayesian Enlightenment: Taking Bayesian models seriously as psychological theories

The most obvious candidate within the Bayesian framework for status as a psychological construct or assumption is the choice of hypothesis space or generative model. According to the Fundamentalist Bayesian view, the hypotheses and their prior distribution correspond to the true environmental probabilities within the domain of study. However, as far as predicting behavior is concerned, all that should matter is what the subject *believes* (either implicitly or explicitly) are the true probabilities. Decoupling information encoded in the brain from ground truth in the environment (which cannot always be determined) allows for separation of two different tenets of the rationalist program. That is, the question of whether people have veridical mental models of their environments can be separated from the question of whether people reason and act optimally with respect to whatever models they have. A similar perspective has been proposed in game theory, whereby distinguishing between an agent's model of the opponent(s) and rational behavior with respect to that model can resolve paradoxes of rationality in that domain (Jones & Zhang 2003). Likewise, Baker et al. (2009) present a model of how people reason about the intentions of others in which the psychological assumption is made that people view others as rational agents (given their current knowledge).

Separating Bayesian inference from the mental models it operates over opens up those models as a fruitful topic of psychological study (e.g., Sanborn et al. 2010b). Unfortunately, this view of Bayesian modeling is at odds with most applications, which focus on the inferential side and take the generative model for granted, leaving that critical aspect of the theory to be hand-coded by the researcher. Thus, the emphasis on rationality marginalizes most of the interesting psychological issues. The choice of the generative model or hypothesis space reflects an assumption about how the subject imputes structure to the environment and how that structure is represented. There are often multiple options here (i.e., there is not a unique Bayesian model of most tasks), and these correspond to different psychological theories. Furthermore, even those cases that ground the hypothesis space in empirical data from natural environments tend not to address how it is learned by individual subjects. One strong potential claim of the Bayesian framework is that the most substantial part of learning lies in constructing a generative model of one's environment, and that using that model to make inferences and guide behavior is a relatively trivial (albeit computationally intensive) exercise in conditional probability. Therefore, treating the generative model as a psychological construct enables a shift of emphasis to this more interesting learning problem. Research focusing on how people develop models of their environment (e.g., Griffiths & Tenenbaum 2006;

Mozer et al. 2008; Steyvers et al. 2003) can greatly increase the theoretical utility of Bayesian modeling by bringing it into closer contact with the hard psychological questions of constructive learning, structured representations, and induction.

Consideration of generative models as psychological constructs also highlights a fundamental difference between a process-level interpretation of Bayesian learning and other learning architectures such as neural networks or production systems. The Bayesian approach suggests that learning involves working backward from sense data to compute posterior probabilities over latent variables in the environment, and then determining optimal action with respect to those probabilities. This can be contrasted with the more purely feed-forward nature of most extant models, which learn mappings from stimuli to behavior and use feedback from the environment to directly alter the internal parameters that determine those mappings (e.g., connection weights or production utilities). A similar contrast has been proposed in the literature on reinforcement learning, between model-based (planning) and model-free (habit) learning, with behavioral and neurological evidence that these exist as separate systems in the brain (Daw et al. 2005). Model-based reinforcement learning and Bayesian inference have important computational differences, but this parallel does suggest a starting point for addressing the important question of how Bayesian learning might fit into a more complete cognitive architecture.

Prior distributions offer another opportunity for psychological inquiry within the Bayesian framework. In addition to the obvious connections to biases in beliefs and expectations, the nature of the prior has potential ties to questions of representation. This connection arises from the principle of conjugate priors (Raiffa & Schlaifer 1961). A *conjugate prior* for a Bayesian model is a parametric family of probability distributions that is closed under the evidence-updating operation of Bayesian inference, meaning that the posterior is guaranteed also to lie in the conjugate family after any number of new observations have been made. Conjugate priors can dramatically simplify computational and memory demands, because the learner needs to store and update only the parameters of the conjugate family, rather than the full evidence distribution. Conjugate priors are a common assumption made by Bayesian modelers, but this assumption is generally made solely for the mathematical convenience of the modeler rather than for any psychological reason. However, considering a conjugate prior as part of the psychological theory leads to the intriguing possibility that the parameters of the conjugate family constitute the information that is explicitly represented and updated in the brain. If probabilistic distributions over hypotheses are indeed part of the brain's computational currency, then they must be encoded in some way, and it stands to reason that the encoding generally converges on one that minimizes the computational effort of updating knowledge states (i.e., of inferring the posterior after each new observation). Therefore, an interesting mechanistic-level test of Bayesian theory would be to investigate whether the variables that parameterize the relevant conjugate priors are consistent with what is known based on more established methods about knowledge representation in various psychological domains. Of course, it is unlikely that any extant formalism (currently adopted for mathematical convenience) will align perfectly with human performance, but empirically exploring and evaluating such possibilities might prove a fruitful starting point.

A final element of Bayesian models that is traditionally considered as outside the psychological theory but that may have valuable process-level implications involves the algorithms that are often used for approximating exact Bayesian inference. Except in models that admit a simple conjugate prior, deriving the exact posterior from a Bayesian model is in most practical cases exceedingly computationally intensive. Consequently, even the articles that propose these models often resort to approximation methods such as Markov-Chain Monte Carlo (MCMC; Hastings 1970) or specializations such as Gibbs sampling (Geman & Geman 1984) to derive approximate predictions. To the extent that Bayesian models capture any truth about the workings of the brain, the brain is faced with the same estimation problems that confront Bayesian modelers, so it too likely must use approximate methods for inference and decision-making. Many of the algorithms used in current Bayesian models correspond to important recent advances in computer science and machine learning, but until their psychological predictions and plausibility are addressed, they cannot be considered part of cognitive theory. Therefore, instead of being relegated to footnotes or appendices, these approximation algorithms should be a focus of the research because this is where a significant portion of the psychology lies. Recent work investigating estimation algorithms as candidate psychological models (e.g., Daw & Courville 2007; Sanborn et al. 2010a) represents a promising step in this direction. An alternative line of work suggests that inference is carried out by a set of simple heuristics that are adapted to statistically different types of environments (Brighton & Gigerenzer 2008; Gigerenzer & Todd 1999). Deciding between these adaptive heuristics and the more complex estimation algorithms mentioned above is an important empirical question for the mechanistic grounding of Bayesian psychological models.

A significant aspect of the appeal of Bayesian models is that their assumptions are explicitly laid out in a clean and interpretable mathematical language that, in principle, affords the researcher a transparent view of their operation. This is in contrast to other computational approaches (e.g., connectionism), in which it can be difficult to separate theoretically important assumptions from implementational details. Unfortunately, as we have argued here, this is not generally the case in practice. Instead, unexamined yet potentially critical assumptions are routinely built into the hypothesis sets, priors, and estimation procedures. Treating these components of Bayesian models as elements of the psychological theory rather than as ancillary assumptions is an important prerequisite for realizing the transparency of the Bayesian framework. In this sense, the shift from Bayesian Fundamentalism to Enlightenment is partly a shift of perspective, but it is one we believe could have a significant impact on theoretical progress.

6.2. Integrating Bayesian analysis with mechanistic-level models

Viewing Bayesian models as genuine psychological theories in the ways outlined here also allows for potential

integration between rational and mechanistic approaches. The most accurate characterization of cognitive functioning is not likely to come from isolated considerations of what is rational or what is a likely mechanism. More promising is to look for synergy between the two, in the form of powerful rational principles that are well approximated by efficient and robust mechanisms. Such an approach would aid understanding not just of the principles behind the mechanisms (which is the sole focus of Bayesian Fundamentalism), but also of how the mechanisms achieve and approximate those principles and how constraints at both levels combine to shape behavior (see Oaksford & Chater [2010] for one thorough example). We stress that we are not advocating that every model include a complete theory at all levels of explanation. The claim is merely that there must be contact between levels. We have argued this point here for rational models, that they should be informed by considerations of process and representation; but the same holds for mechanistic models as well, that they should be informed by consideration of the computational principles they carry out (Chater et al. 2003).

With reference to the problem of model fractionation discussed earlier, one way to unite Bayesian models of different phenomena is to consider their rational characterizations in conjunction with mechanistic implementations of belief updating and knowledge representation, with the parsimony-derived goal of explaining multiple computational principles with a common set of processing mechanisms. In this way the two levels of analysis serve to constrain each other and to facilitate broader and more integrated theories. From the perspective of theories as metaphors, the rationality metaphor is unique in that is has no physical target, which makes it compatible with essentially any mechanistic metaphor and suggests that synthesis between the two levels of explanation will often be natural and straightforward (as compared to the challenge of integrating two distinct mechanistic architectures). Daw et al. (2008) offer an excellent example of this approach in the context of conditioning, by mapping out the relationships between learning algorithms and the rational principles they approximate, and by showing how one can distinguish behavioral phenomena reflecting rational principles from mechanistic signatures of the approximation schemes.

Examples of work that integrates across levels of explanation can also be found in computational neuroscience. Although the focus is not on explaining behavior, models in computational neuroscience relate abstract probabilistic calculations to operations in mechanistic neural network models (Denève 2008; Denève et al. 1999). Other work directly relates and evaluates aspects of Bayesian models to brain areas proposed to perform the computation (Doll et al. 2009; Soltani & Wang 2010). For example, Köver and Bao (2010) relate the prior in a Bayesian model to the number of cells devoted to representing possible hypotheses. This work makes contact with all three of Marr's (1982) levels of analysis by making representational commitments and relating these aspects of the Bayesian model to brain regions.

An alternative to the view of mechanisms as approximations comes from the research of Gigerenzer and colleagues on adaptive heuristics (e.g., Gigerenzer & Todd 1999). Numerous studies have found that simple heuristics can actually outperform more complex inference algorithms in naturalistic prediction tasks. For example, with certain datasets, linear regression can be outperformed in cross-validation (i.e., transfer to new observations) by a simple tallying heuristic that gives all predictors equal weight (Czerlinski et al. 1999; Dawes & Corrigan 1974). Brighton and Gigerenzer (2008) explain how the advantage of simple heuristics is rooted in the bias-variance dilemma from statistical estimation theory – specifically, that more constrained inference algorithms can perform better on small datasets because they are less prone to overfitting (e.g., Geman et al. 1992). Although this conclusion has been used to argue against computational-level theories of rationality in favor of ecological rationality based on mechanisms adapted to specific environments (Gigerenzer & Brighton 2009), we believe the two approaches are highly compatible. The connection lies in the fact that any inference algorithm implicitly embodies a prior expectation about the environment, corresponding to the limitations in what patterns of data it can fit and hence the classes of environments in which it will tend to succeed (cf. Wolpert 1996). For example, the tallying heuristic is most successful in environments with little variation in true cue validities and in cases where the validities cannot be precisely estimated (Hogarth & Karelaia 2005). This suggests that tallying should be matched or even outperformed by Bayesian regression with a prior giving more probability to more homogeneous regression weights. The point here is that the ecological success of alternative algorithms (tallying vs. traditional regression) can inform a rational analysis of the task and hence lead to more accurate normative theories. This sort of approach could alleviate the insufficient environmental grounding and excessive flexibility of Bayesian models discussed in section 5. Formalizing the relationship between algorithms and implicit priors – or between statistical regularities in particular environments and algorithms that embody those regularities – is therefore a potentially powerful route to integrating mechanistic and rational approaches to cognition.

Another perspective on the relationship between Bayesian and mechanistic accounts of cognition comes from the recognition that, at its core, Bayes' Rule is a model of the decision process. This is consistent with (and partly justifies) the observation that most work in the Bayesian Fundamentalist line avoids commitments regarding representation. However, the thesis that inference and decision-making are optimal is meaningful only in the context of the knowledge (i.e., beliefs about the environment) with respect to which optimality is being defined. In other words, a complete psychological theory must address both how knowledge is acquired and represented and how it is acted upon. As argued in section 4.1, questions of the structure of people's models of their environments, and of how those models are learned, are better addressed by traditional, mechanistic psychological methods than by rational analysis. Taken together, these observations suggest a natural synthesis in which psychological mechanisms are used to model the learner's state of knowledge, and rational analysis is used to predict how that knowledge is used to determine behavior.

The line between knowledge and decision-making, or representation and process, is of course not so well defined as this simple proposal suggests, but the general idea is that rational analysis can be performed not in the environment but instead within a mechanistic model,

thus taking into account whatever biases and assumptions the mechanisms introduce. This approach allows the modeler to postulate decision rules that are optimal with respect to the representations and dynamics of the rest of the model. The result is a way of enforcing "good design" while still making use of what is known about mental representations. It can improve a mechanistic model by replacing what might otherwise be an arbitrary decision rule with something principled, and it also offers an improvement over rational analysis that starts and ends with the environment and is not informed by how information is actually represented. This approach has been used successfully to explain, for example, aspects of memory as optimal retrieval, given the nature of the encoding (Shiffrin & Steyvers 1998); patterns of short-term priming as optimal inference with unknown sources of feature activation (Huber et al. 2001); and sequential effects in speeded detection tasks as optimal prediction with respect to a particular psychological representation of binary sequences (Wilder et al. 2009). A similar approach has also been applied at the neural level, for example, to model activity of lateral intraparietal (LIP) neurons as computing a Bayesian posterior from activity of middle temporal (MT) cells (Beck et al. 2008). One advantage of bringing rational analysis inside cognitive or neural models is that it facilitates empirical comparison among multiple Bayesian models that make different assumptions about knowledge representation (e.g., Wilder et al. 2009). These lines of research illustrate that the traditional identification of rational analysis with computational-level theories is an artificial one, and that rational analysis is in fact applicable at all levels of explanation (Danks 2008).

A complementary benefit of moving rational analysis inside psychological models is that the assumption of optimal inference can allow the researcher to decide among multiple candidate representations, through comparison to empirical data. The assumption of optimal inference allows for more unambiguous testing of representation, because representation becomes the only unknown in the model. This approach has been used successfully in the domain of category induction by Tenenbaum et al. (2006). However, such conclusions depend on a strong assumption of rational inference. The question of rational versus biased or heuristic inference has been a primary focus of much of the judgment and decision-making literature for several decades, and there is a large body of work arguing for the latter position (e.g., Tversky & Kahneman 1974). On the other hand, some of these classic findings have been given rational reinterpretations under new assumptions about the learner's knowledge and goals (e.g., Oaksford & Chater 1994). This debate illustrates how the integration of rational and mechanistic approaches brings probabilistic inference under the purview of psychological models where it can be more readily empirically tested.

Ultimately, transcending the distinction between rational and mechanistic explanations should enable significant advances of both and for cognitive science as a whole. Much of how the brain operates reflects characteristics of the environment to which it is adapted, and therefore an organism and its environment can be thought of as a joint system, with behavior depending on aspects of both subsystems. There is of course a fairly clear line between organism and environment, but that line has no more epistemological significance than the distinctions between different sources of explanation within either category. In other words, the gap between an explanation rooted in some aspect of the environment and one rooted in a mechanism of neural or cognitive processing should not be qualitatively wider than the gap between explanations rooted in different brain regions, different processing stages or modules, or uncertainty in one latent variable versus another. The joint system of organism and environment is a complex one, with a large number of constituent processes; and a given empirical phenomenon (of behavior, brain activity, etc.) can potentially be ascribed to any of them. Just as in other fields, the scientific challenge is to determine which explanation is best in each case, and for most interesting phenomena the answer will most likely involve an interaction of multiple, disparate causes.

7. Conclusions

The recent advances in Bayesian modeling of cognition clearly warrant excitement. Nevertheless, many aspects of current research practice act to severely limit the contributions to psychological theory. This article traces these concerns to a particular philosophy that we have labeled Bayesian Fundamentalism, which is characterized by the goal of explaining human behavior solely in terms of optimal probabilistic inference, without recourse to mechanism. This philosophy is motivated by the thesis that, once a given task is correctly characterized in terms of environmental statistics and goals of the learner, human behavior in that task will be found to be rational. As the numerous citations throughout this article demonstrate, Bayesian Fundamentalism constitutes a significant portion (arguably the majority) of current research on Bayesian modeling of cognition.

Establishing the utility of the Bayesian framework, and the rational metaphor more generally, is an important first step, and convincing arguments have been made for this position (e.g., Oaksford & Chater 2007). However, excessive focus on this meta-scientific issue severely limits the scope and impact of the research. Focusing on existence proofs distracts from the more critical work of deciding among competing explanations and identifying the critical assumptions behind models. In the context of rational Bayesian modeling, existence proofs hide the fact that there are generally many Bayesian models of any task, corresponding to different assumptions about the learner's goals and model of the environment. Comparison among alternative models would potentially reveal a great deal about what people's goals and mental models actually are. Such an approach would also facilitate comparison to models within other frameworks, by separating the critical assumptions of any Bayesian model (e.g., those that specify the learner's generative model) from the contribution of Bayes' Rule itself. This separation should ease recognition of the logical relationships between assumptions of Bayesian models and of models cast within other frameworks, so that theoretical development is not duplicated and so that the core differences

between competing theories can be identified and tested.

The total focus on rational inference that characterizes Bayesian Fundamentalism is especially unfortunate from a psychological standpoint because the updating of beliefs entailed by Bayes' Rule is psychologically trivial, amounting to nothing more than vote counting. Much more interesting are other aspects of Bayesian models, including the algorithms and approximations by which inference is carried out, the representations on which those algorithms operate (e.g., the parameters of conjugate priors), and the structured beliefs (i.e., generative models) that drive them. The Enlightened Bayesian view takes these seriously as psychological constructs and evaluates them according to theoretical merit rather than mathematical convenience. This important shift away from Bayesian Fundamentalism opens up a rich base for psychological theorizing, as well as contact with process-level modes of inquiry.

It is interesting to note that economics, the field of study with the richest history of rational modeling of behavior and the domain in which rational theories might be expected to be most accurate, has increasingly questioned the value of rational models of human decision-making (Krugman 2009). Economics is thus moving away from purely rational models toward theories that take into account psychological mechanisms and biases (Thaler & Sunstein 2008). Therefore, it is surprising to observe a segment of the psychological community moving in the opposite direction. Bayesian modeling certainly has much to contribute, but its potential impact will be much greater if developed in a way that does not eliminate the psychology from psychological models. We believe this will be best achieved by treating Bayesian methods as a complement to mechanistic approaches, rather than as an alternative.

ACKNOWLEDGMENTS

This research was supported in part by the Air Force Office of Scientific Research (AFOSR) Grant no. FA9550-07-1-0178 and Army Research Laboratory (ARL) Grant no. W911NF-09-2-0038 to Bradley C. Love. We thank John Anderson, Colin Bannard, Lera Boroditsky, David Buss, Matt Keller, Mike Mozer, Mike Oaksford, Randy O'Reilly, Michael Ramscar, Vladimir Sloutsky, and Alan Yuille for helpful comments on an earlier draft of this article.

NOTES

1. Formally, $E_{\text{posterior}}$ equals the logarithm of the posterior distribution, E_{prior} is the logarithm of the prior, and $E_{data}(H)$ is the logarithm of the likelihood of the data under hypothesis H. The model's prediction for the probability that hypothesis H is correct, after data have been observed, is proportional to $\exp[E_{\text{posterior}}(H)]$ (cf. Luce 1963).

2. Bayesian analysis has been used to interpret neural spike recordings (e.g., Gold & Shadlen 2001), but this falls outside Bayesian Fundamentalism, which is concerned only with behavioral explanations of cognitive phenomena.

3. Note that we refer here to Bayesian models that address behavior, not those that solely aim to explain brain data without linking to behavior, such as Mortimer et al.'s (2009) model of axon wiring.

Open Peer Commentary

Evolutionary psychology and Bayesian modeling

doi:10.1017/S0140525X11000173

Laith Al-Shawaf and David Buss

Department of Psychology, University of Texas, Austin, TX 78712.

dbuss@psy.utexas.edu **www.davidbuss.com**

Abstract: The target article provides important theoretical contributions to psychology and Bayesian modeling. Despite the article's excellent points, we suggest that it succumbs to a few misconceptions about evolutionary psychology (EP). These include a mischaracterization of evolutionary psychology's approach to optimality; failure to appreciate the centrality of mechanism in EP; and an incorrect depiction of hypothesis testing. An accurate characterization of EP offers more promise for successful integration with Bayesian modeling.

Jones & Love (J&L) provide important theoretical contributions to psychology and Bayesian modeling. Especially illuminating is their discussion of whether Bayesian models are agnostic about psychology, serving mainly as useful scientific and mathematical tools, or instead make substantive claims about cognition.

Despite its many strengths, the target article succumbs to some common misconceptions about evolutionary psychology (EP) (Confer et al. 2010). The first is an erroneous characterization of EP's approach to *optimality and constraints*. Although the article acknowledges the importance of constraints in evolutionary theory, it lapses into problematic statements such as "evolutionary pressures tune a species' genetic code such that the observed phenotype gives rise to optimal behaviors" (sect. 5, para. 3). J&L suggest that evolutionary psychologists reinterpret behavioral phenomena as "optimal" by engaging in a post hoc adjustment of their view of the relevant selection pressures operating in ancestral environments.

These statements imply that a key goal of EP is to look for optimality in human behavior and psychology. On the contrary, the existence of optimized mechanisms is rejected by evolutionary psychologists, as this passage from Buss et al. (1998) illustrates:

> [T]ime lags, local optima, lack of genetic variation, costs, and limits imposed by adaptive coordination with other mechanisms all constitute major constraints on the design of adaptations. . . . Adaptations are not optimally designed mechanisms. They are . . . jerry-rigged, meliorative solutions to adaptive problems . . ., constrained in their quality and design by a variety of historical and current forces. (Buss et al. 1998, p. 539)

J&L argue that "it is not [simply] any function that is optimized by natural selection, but only those functions that are relevant to fitness" (sect. 5, para. 4). We agree with the implication that psychologists must consider the fitness-relevance of the mechanisms they choose to investigate. Identifying adaptive function is central. Nonetheless, natural selection is better described as a "meliorizing" force, not an optimizing force (see Dawkins 1982, pp. 45–46) – and thus even psychological mechanisms with direct relevance to fitness are not optimized. As J&L correctly note elsewhere, selection does not favor the best design in some global engineering sense, but rather features that are *better* than competing alternatives extant in the population at the time of selection, within existing constraints (Buss et al. 1998; Dawkins 1982).

Despite occasional problems with the target article's depiction of EP's views on optimality, we fully agree with J&L that (a) adaptationist accounts place significant constraints on explanation, (b) evolution proceeds by "survival of the best current

design, not survival of the globally optimal design" (sect. 5.3, para. 3), (c) human cognition is not optimally designed, and (d) the "rational program" in Bayesian modeling has an overly narrow focus on optimally functioning adaptations.

J&L present a partly accurate and partly inaccurate characterization of the relevance of *mechanism* in evolutionary approaches. They correctly acknowledge the importance of elucidating the specific mechanistic workings of adaptations. However, the target article compares EP to Bayesian Fundamentalism and Behaviorism by claiming that all three approaches eschew the investigation of mechanism. We disagree with this latter assessment.

In our view, it is difficult or impossible to study function without investigating form or mechanism. The central logic of adaptationism makes the inextricable link between form (or mechanism) and function clear: An adaptation must necessarily be characterized by a good fit between form and function – between an adaptation and the adaptive problem it was "designed" to solve .The key point is that evolutionary approaches to psychology *necessarily* involve the joint investigation of mechanism and function. Evolutionary psychology generates hypotheses about "design features," or particular mechanistic attributes, that adaptations either must have or might have in order to successfully solve the adaptive problems that they evolved to solve. Indeed, mechanism is one of Tinbergen's (1963) four explanatory levels – mechanism, ontogeny, function, and phylogeny. Ideally, all should be analyzed in order to achieve a complete understanding of any behavior or psychological phenomenon, and all are central to core aims of EP. Of course, not every scientist explores all four questions; every empirical study has delimited aims; and the field is certainly far from a complete understanding of all of the design features of any mechanism, whether it be the human visual system or incest-avoidance adaptations.

As a single example of mechanistic EP research, adaptationist analyses of fear have uncovered social inputs that elicit the emotion, nonsocial inputs that trigger the emotion, the adaptive behavioral output designed to solve the problem, the perceptual processes involved in detecting threats and reacting fearfully, the developmental trajectory of human fears, and the physiological and endocrinological mechanisms driving the fear response (see, e.g., Bracha 2004; Buss 2011; Neuhoff 2001; Öhman et al. 2001). Analogous progress has been made in understanding other evolved mechanisms, such as mating adaptations, perceptual biases, and adaptive social inference biases (Buss 2011).

Most human adaptations are only just beginning to be subjected to scientific investigation, and many mechanistic details have certainly not yet been elucidated. EP could profitably increase its use of formal mechanistic modeling in this endeavor. Fusing the strengths of mathematical and computational modelers with those of evolutionary psychologists would enrich both fields.

Finally, the target article depicts EP as occasionally falling into "backward-looking" hypotheses (sect. 5.2, para. 3) or engaging in "just so" storytelling (sect. 5.2, para. 1; Gould & Lewontin 1979). By this, the authors mean that evolutionary psychologists sometimes note a behavior or psychological mechanism, and then construct a conceivable function for it and simply stop there. We agree with J&L that this practice would be highly problematic if it were the end point of scientific analysis.

Fortunately, leading work in EP proceeds using both the *forward* method in science (theory leads directly to hypothesis, which then leads to empirical predictions, which are then tested) as well as the *backward* method (observed phenomenon leads to hypothesis, which in turn leads to novel empirical predictions, which are then tested) (see Buss 2011; Tooby & Cosmides 1992). Much of evolutionary psychology uses the forward method, and here it is not even *possible* to level the "just-so story" criticism. When evolutionary psychologists employ the backward method, they typically avoid the problem by taking the additional necessary step of deriving *novel* and *previously untested* predictions from the hypothesis (for numerous examples, see Buss 2011). We concur with the implication that there are better and poorer practitioners of the rigors of science, and that all should be held to the highest standards for more rapid progress.

In sum, we view an accurately characterized modern evolutionary psychology as largely avoiding the conceptual pitfalls J&L note, and we look forward to a richer and more successful integration of Bayesian modeling and evolutionary psychology.

The myth of computational level theory and the vacuity of rational analysis

doi:10.1017/S0140525X11000185

Barton L. Anderson

School of Psychology, University of Sydney, Sydney, NSW 2006, Australia.
barta@psych.usyd.edu.au
http://www.psych.usyd.edu.au/staff/barta/

Abstract: I extend Jones & Love's (J&L's) critique of Bayesian models and evaluate the conceptual foundations on which they are built. I argue that: (1) the "Bayesian" part of Bayesian models is scientifically trivial; (2) "computational level" theory is a fiction that arises from an inappropriate programming metaphor; and (3) the real scientific problems lie outside Bayesian theorizing.

The Bayesian framework asserts that problems of perception, action, and cognition can be understood as (approximations to) ideal rational inference. Bayes' rule is a direct consequence of the definition of conditional probability, and is reasonably captured as the simple "vote counting" procedure outlined in the target article by Jones & Love (J&L). This is clearly not where the interesting science lies. The real scientific problems for a Bayesian analysis arise in defining the appropriate hypothesis space (the "candidates" for whom votes will be cast), and a principled means of assigning priors, likelihoods, and cost functions that will, when multiplied, determine the distribution of votes (and the ultimate winner[s]).

Bayesian models of cognition begin by asserting that brains are devices that compute, and that it is possible to dissociate *what* they compute from *how* they compute. David Marr's (1982) now infamous dissociation of the computational, algorithmic, and implementation "levels of analysis" is usually invoked to justify this belief, and inspires attempts to "reverse engineer" the mind (Tenenbaum et al. 2011). It is no coincidence that Marr's levels resemble the stages of someone writing a computer program, which are granted some (unspecified) form of ontological status: A problem is defined, code is written to solve it, and a device is employed to run the code. But unlike the computational devices fashioned by man, the brain, like other bodily organs, emerged as the consequence of natural processes of self-organization; the complexity of its structure and function was not prescribed in some top-down manner as solutions to pre-specified computational problem(s). The only "force" available to construct something ideal is natural selection, which can only select the best option from whatever is available, even if that is nothing more than a collection of hacks. As for the "computational level" theory, it is far from evident that brains can be accurately characterized as *performing* computations any more than one can claim that planets compute their orbits, or rocks rolling down hills compute their trajectory. Our formal models are the language that we use to try to capture the causal entailments of the natural world with the inferential entailments embodied in the formal language of mathematics (Rosen 1991). The assertion

that a computational level of analysis exists is merely an assertion grounded on the latest technological metaphor. If this assertion is false, then models that rely on its veracity (like Bayesian models of cognition) are also false.

But even if we accept Marr's levels of analysis, we have gained very little theoretical leverage into how to proceed. We must now guess what the computational problems are that brains solve. There is no principled Bayesian method to make *these* guesses, and this is not where the guessing game ends. Once a computational problem is identified, we must guess a hypothesis space, priors, likelihoods, and cost functions that, when multiplied together, supply the problem's solution. The majority of this guess-work is shaped by the very data that cognitive models are attempting to explain: The "richly structured representations" of cognition often seem like little more than a re-description of the structure in the data, recast as post hoc priors and likelihoods that now pose as theory.

Similar problems arise in Bayesian models of perception, although some of the guess-work can be constrained by generative models of the input and the statistics of natural environments. The Bayesian approach has been cast into "natural systems analysis" (Geisler & Ringach 2009), which asserts that perceptual systems should be analyzed by specifying the natural tasks an animal performs and the information used to perform them. The specification of "natural tasks" plays essentially the same role as Marr's computational level analysis. Once a task is defined, an ideal observer is constructed: a hypothetical device that performs a task optimally given the "available information." It is difficult to argue with the logic of this approach, as it is equivalent to stating that we should study the perceptual abilities that were responsible for our evolutionary survival. But how do we discriminate between the natural tasks that were the product of natural selection from those that merely came along for the ride? This problem is not unique to the analysis of psychological systems; it is a general problem of evolutionary biology (i.e., distinguishing products of adaptive selection from "spandrels" – by-products of the selective adaptation of some other trait).

It is unclear how natural systems analysis provides any theoretical leverage into any of these deep problems (i.e., how the modifier "natural" constrains "systems analysis"). We must first guess what counts as the "natural tasks" to determine the appropriate objects of study. We then guess what information is available (and used) to perform that task ideally. The ideal observers so articulated are only "ideal" to the extent that we have correctly identified both the available information and a task that the perceptual system actually performs. We then construct an experiment to compare biological performance with ideal performance. And although natural systems analysis begins by considering properties of natural scenes, the majority of the experimental paradigms that assess natural tasks are largely indistinguishable from the larger body of perceptual research. In order to achieve adequate experimental control, most of the complexity of natural scenes has been abstracted away, and we are left with displays and methods that could have (and typically were) invented without any explicit reference to, or consideration of, Bayes theorem.

In the end, we are left trying to understand what animals do and how they do it. The hard problems remain inaccessible to the tools of Bayesian analysis, which merely provide a means to select an answer once the hard problem of specifying the list of possible answers has been solved (or at least prescribed). Bayesian analysis assures us that there is some way to conceive of our perceptual, cognitive, and motor abilities as "rational" or "ideal." Like J&L, I fail to experience any insight in this reassurance. And I am left wondering how to reconcile such views with the seemingly infinite amount of irrationality I encounter in my daily life.

Maybe this old dinosaur isn't extinct: What does Bayesian modeling add to associationism?

doi:10.1017/S0140525X11000203

Irina Baetu,[a] Itxaso Barberia,[b] Robin A. Murphy,[c] and A. G. Baker[b]

[a]*Department of Experimental Psychology, University of Cambridge, Cambridge CB2 3EB, England, United Kingdom;* [b]*Department of Psychology, McGill University, Montréal, QC, H3A 1B1, Canada;* [c]*Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, United Kingdom.*

irina.baetu@mail.mcgill.ca **itxaso.barberiafernandez@mail.mcgill.ca**
andy.baker@mcgill.ca **robin.murphy@psy.ox.ac.uk**
http://web.me.com/robinmurphy/oxford/Welcome.html

Abstract: We agree with Jones & Love (J&L) that much of Bayesian modeling has taken a fundamentalist approach to cognition; but we do not believe in the potential of Bayesianism to provide insights into psychological processes. We discuss the advantages of associative explanations over Bayesian approaches to causal induction, and argue that Bayesian models have added little to our understanding of human causal reasoning.

Jones & Love (J&L) reveal the shortcomings of current Bayesian approaches to psychological processes. We agree with the authors on this but believe they still show too much faith in the potential of the Bayesian approach as an alternative to connectionist modeling. We illustrate the problems with a discussion of research on human causal induction. Bayesian fundamentalism has taken firm hold in this field; however, there is also a tradition of associative analysis, allowing us to consider the added value of Bayesianism.

Causal reasoning in naturalistic environments quickly becomes a mathematically complex problem requiring important auxiliary processes beyond the math. According to a Bayesian analysis, people start by representing and selecting from multiple causal structures. For example, three events A, B, and C can be described by several classes of causal model, including a causal chain in which A causes B which then causes C ($A \rightarrow B \rightarrow C$) and a common cause model in which A causes B and C ($C \leftarrow A \rightarrow B$). Observations are used to decide which is most likely. A simple Bayesian analysis has problems because some of the models can be Bayesian equivalents. For example, a deterministic chain model ($A \rightarrow B \rightarrow C$) and a deterministic common cause model ($C \leftarrow A \rightarrow B$) produce the same empirical probabilities. Either all three events are present on a trial or none are.

One way to reduce this ambiguity is to do an intervention. Removing the effect of B enables one to discriminate between chain and common cause models. With a chain ($A \rightarrow B \rightarrow C$), C will no longer occur, but in a common cause model ($C \leftarrow A \rightarrow B$), C will occur. People can decide between models by using interventions (Steyvers et al. 2003). However, the critical feature of this Bayesian modeling is that the Bayesian component does not do the crucial intervention but simply assesses the result of the intervention. A *supervisor* or other mechanism decides which intervention to make (Hagmayer et al. 2007; Steyvers et al. 2003). It is this decision process that is outside the Bayesian frame that is critical. The dependencies in the causal graphs may be generated by Bayesian processes, but the Bayesian analysis does not do the psychological work that solves the problem. As J&L suggest, it merely does the complex counting. Associative or connectionist models can also do this work. For example, a traditional associative net has nodes and connections of varying strengths between them (e.g., Baker et al. 1996). And it is common practice for scientists to remove nodes in order to discover what the associative nets have learned (e.g., Cowell et al. 2006; Harm & Seidenberg 1999; Plaut 1995). If the *supervisor* were to perform similar actions, this would be an associative analogue to graph surgery. An associative analysis is not only very similar to the Bayesian

computation, it also has some advantages which we discuss. Therefore, it is disingenuous to claim that this type of research in causal induction arises from or differentially supports a Bayesian perspective.

The second way to disambiguate the causal models, temporal precedence, has been the object of much associative analysis (e.g., Pavlov 1927). The causal arrows in the models can be replaced with "leads to" and, if the observer can discriminate order, they could easily discriminate the common cause, A leads to B and C, from the chain, A leads to B that leads to C models. Bayesian fundamentalists researching causal induction (e.g., Lagnado & Sloman 2006) have, indeed, shown that with temporal information people can discriminate the models. However, the timing information does not arise directly from Bayesian computations. Again it is the *supervisor* that disambiguates the data by using timing. However, learning orders emerges easily from a connectionist perspective. Associative nets, like neuronal processes, have activations that decay with time. In (A → B → C), when C finally comes along, A activation will have decayed more than B activation, so a stronger link between B and C will form. Baetu and Baker (2009) have reported a series of experiments in which they have studied how people form causal chains from experience with the individual links (generative and preventative). People are good at assembling these independently learned links into chains. But, most important, a modification of a simple associative model (the autoassociator; McClelland & Rumelhart 1988) generates representations of causal structure, and predicts participants' behavior.

Finally, the associative structures have a certain face validity for psychological processes that the Bayesian frames do not. In associative terms, causal chains are represented as associative strengths and not likelihood ratios. Order and timing can flow naturally from them. They represent causes and effects of different magnitudes, and not just binary (absent/present) events. Causes and activations may be weak or strong, and not just present or absent.

What does this say about J&L's thesis? First, we agree that Bayesianism must progress beyond fundamentalism. Indeed, much of the research concerning the *supervisor* can lead the unwary believer to the unwarranted conclusion that this work discriminates Bayesian computations from others. Second, J&L argue that the Bayesian analysis can prosper at all levels. In our rather simple case, it certainly does not account for the *supervisor*. It is not clear how it could in a principled way. Third, they argue that Bayesian analyses should become closely linked to psychological mechanism, and we agree; but we argue that associative structure may already be there. For instance, we are now closer to understanding how a prediction-error algorithm (e.g., Rescorla & Wagner 1972) might be implemented in the brain (Kim et al. 1998; Waelti et al. 2001).

In conclusion, we agree with J&L that Agnostic Bayesian nets offer a powerful method for artificial intelligence and that, if elaborated, can learn about or represent any finite data set – but so could an associative net. However, the question for psychological process is one of parsimony and mechanistic plausibility, and we are not convinced that J&L have demonstrated this contribution. We would be more convinced if they had described a single instance where a Bayesian analysis produced a realistic psychological mechanism or empirical result.

Integrating Bayesian analysis and mechanistic theories in grounded cognition

doi:10.1017/S0140525X11000197

Lawrence W. Barsalou

Department of Psychology, Emory University, Atlanta, GA 30322.
barsalou@emory.edu
http://www.psychology.emory.edu/cognition/barsalou/index.html

Abstract: Grounded cognition offers a natural approach for integrating Bayesian accounts of optimality with mechanistic accounts of cognition, the brain, the body, the physical environment, and the social environment. The constructs of *simulator* and *situated conceptualization* illustrate how Bayesian priors and likelihoods arise naturally in grounded mechanisms to predict and control situated action.

In the spirit of Bayesian Enlightenment, as suggested by Jones & Love (J&L), grounded cognition offers architectural mechanisms that naturally afford Bayesian analysis. In particular, the constructs of *simulator* and *situated conceptualization* illustrate the potential for integrating explanations across Marr's (1982) computational, algorithmic, and implementation levels. Many other grounded constructs also undoubtedly offer similar potential.

In *perceptual symbol systems*, a *simulator* is a dynamical system distributed across the modalities that process a category's properties, aggregating information about diverse instances, comparable to a concept in traditional theories (Barsalou 1999; 2003a). The *beer* simulator, for example, aggregates information about how beer looks, smells, and tastes, how it is consumed, how we feel afterwards, and so on. If someone experiences a wide variety of beers (e.g., American, Belgian, English, German, Czech, Indian, Thai, etc.), the beer simulator captures diverse multi-modal states about the category (modeled naturally with neural net architectures; e.g., Pezzulo et al. 2011). On a given occasion, the beer simulator dynamically produces one of many specific beer simulations, from an infinite set possible. A natural way of thinking about the space of possible simulations within a simulator is as a space of Bayesian priors, with some simulations being more likely than others. Furthermore, the strength of a given prior can be assessed empirically (rather than simply assumed), reflecting well-established and readily measured factors, including how frequently and recently category instances have been experienced, how ideal or preferred they are, their similarity to other instances, and so forth (Barsalou 1985; 1987).

In grounded approaches to the conceptual system, a *situated conceptualization* is a situation-relevant simulation from a simulator embedded in the representation of a likely background situation (Barsalou 2003b; 2008c; Yeh & Barsalou 2006). One situated conceptualizations of *chair*, for example, represents a chair on a *jet*, embedded in a jet setting, accompanied by relevant actions and mental states. A natural way of thinking about a situated conceptualization is as representational structure that captures and produces Bayesian likelihoods. Entering a jet setting, for example, may activate the situated conceptualization for jet chairs, producing the expectancy that this specific type of chair will be experienced shortly. Similarly, seeing a jet chair activates expectancies about how to interact with it, how it will feel to operate, and so on. Again, the statistical structure of situated conceptualizations can be assessed empirically, through objective assessments of the environment, subjective estimates of co-occurrence, et cetera. In general, much evidence demonstrates that situated conceptualizations produce part-whole inferences to support diverse forms of reasoning (e.g., Barsalou et al. 2003; 2005; Wilson-Mendenhall et al. 2011).

Together, simulators and situated conceptualizations naturally produce Bayesian inferences. Before entering a building for the first time, priors associated with the *chair* simulator produce situation-independent inferences about chairs likely to be encountered (e.g., kitchen chairs, easy chairs, office chairs), whereas likelihoods emerging from relevant situated conceptualizations produce inferences about likely chairs to be found in this particular context (e.g., living rooms of Buddhist friends). From the perspective of grounded cognition, priors and likelihoods are combined to produce simulations that prepare agents for what is likely to exist in the world, for how to act on the world, and for the mental states likely to result (Barsalou 2009; Barsalou et al. 2007). Because such simulations utilize the modalities for perception, action, and internal states, representations in these modalities become primed, thereby facilitating expected interaction with the environment. Although architectures remain to be developed

that combine these sources of Bayesian information to produce simulations, neural net architectures that reactivate previously experienced states have much potential for doing so.

Because simulators and situated conceptualizations occur in nonhumans, they offer a natural account of conceptual processing across species (Barsalou 2005). If so, the kind of Bayesian analysis just described applies comparatively, perhaps via somewhat common forms of optimality arising continuously across evolution. Where humans are likely to differ is in the linguistic control of this architecture, with words activating simulators, and larger linguistic structures specifying situated conceptualizations compositionally and productively (Barsalou 1999; 2008b).

Bayesian analysis can also be applied to linguistic forms, similarly to how it can be applied to simulators and situated conceptualizations. On activating a word, the probability that other words become active reflects a distribution of priors over these words, constrained by likelihoods, given other words in the context. As research shows increasingly, the statistical structure of linguistic forms mirrors, to some extent, the structure of conceptual knowledge grounded in the modalities (e.g., Andrews et al. 2009; Barsalou et al. 2008; Louwerse & Connell 2011). Of interest is whether similar versus different factors optimize the retrieval of linguistic forms and conceptual knowledge, and what sorts of factors optimize their interaction.

Finally, the grounded perspective assumes that cognition relies inherently on the body, the physical environment, and the social environment, not just on classic cognitive mechanisms (Barsalou 2008a). Because cognition does not occur independently of these other systems, characterizing their structure is essential, analogous to the importance of characterizing the physical environment in Bayesian analysis.

For all these reasons, grounded cognition offers a natural approach for practicing and achieving Bayesian Enlightenment. As cognition emerges from bodily and neural mechanisms through interactions with physical and social environments, numerous forms of optimization undoubtedly occur at many levels. Fully understanding these optimizations seems difficult – not to mention unsatisfying – unless all relevant levels of analysis are taken into account. Indeed, this is the epitome of cognitive science.

Mechanistic curiosity will not kill the Bayesian cat

doi:10.1017/S0140525X11000215

Denny Borsboom,[a] Eric-Jan Wagenmakers,[a] and Jan-Willem Romeijn[b]

[a]*Department of Psychology, University of Amsterdam, 1018 WB Amsterdam, The Netherlands;* [b]*Department of Philosophy, University of Groningen, 9712 GL Groningen, The Netherlands.*

dennyborsboom@gmail.com ej.wagenmakers@gmail.com j.w.romeijn@rug.nl http://sites.google.com/site/borsboomdenny/ dennyborsboom http://www.ejwagenmakers.com/ http://www.philos.rug.nl/~romeyn

Abstract: Jones & Love (J&L) suggest that Bayesian approaches to the explanation of human behavior should be constrained by mechanistic theories. We argue that their proposal misconstrues the relation between process models, such as the Bayesian model, and mechanisms. While mechanistic theories can answer specific issues that arise from the study of processes, one cannot expect them to provide constraints in general.

Jones & Love (J&L) argue that Bayesian approaches to human behavior should attend more closely to cognitive and neural mechanisms. Because mechanisms play such an important role in their target article, it is important to get a clear idea of what mechanisms are and what they are good for. J&L unfortunately do not clarify the term. They get closest when, in section 5.1, they mention the "notion of *mechanism* (i.e., process or representation)" (para. 3, emphasis J&L's). This treatment is, in our view, less accurate than would be needed to support the strong claims the target article makes with regard to the status of Bayesian approaches to cognition. When the concepts of mechanism and process are fleshed out, these claims might well turn out to be untenable.

Roughly, processes and mechanisms relate as follows. A *process* concerns the change of a system over time. The easiest way to think about this is as a path through a set of possible states the system can be in. A *process model* is a description of this path, detailing how each new state (or its probability) depends on its previous state(s). In the behavioral sciences, such a model can often be represented by a flowchart. A *mechanism*, by contrast, is not a process but a system. It typically has parts that work together to implement an input-output relation. For instance, smoking (input) robustly produces lung cancer (output), through a causal mechanism (smoke brings tar into the lungs which leads to mutations). A *mechanistic model* is a representation of the way the parts of the system influence one another. Typically, this is represented as a directed graph or a circuit diagram. Mechanisms are closely tied to the notion of *function*, because they are often studied and discovered by pursuing questions of the "how does this work?" variety (e.g., "how does smoke cause cancer?").

Now, a Bayesian model is a process model, not a mechanistic model. This is not, as J&L believe, because "the Bayesian metaphor is tied to a mathematical ideal and thus eschews mechanism altogether" (sect. 2.2, para. 3), but simply because it describes how a rational agent moves through an abstract state-space of beliefs (probabilities of hypotheses) when confronted with evidence (data): all the model says is how a rational agent is to move to new belief state at $t + 1$, given the prior belief state and evidence available at time t. This has nothing to do with the fact that the model is mathematically formalized. Mechanistic and causal models have mathematical formalizations just as well (e.g., see Pearl 2000). The Bayesian model is simply not a mechanistic model because it is a process model. To argue that the Bayesian model fails to capture mechanisms is much like arguing against relativity theory because it provides no mechanistic detail on how clocks slow down when moved.

Clearly there have to be mechanisms that allow the belief-updating process to run, and these mechanisms are likely to reside in our brain. One may profitably study these mechanisms and even provide support for Bayesian models with that. A good question, for instance, that may receive a mechanistic answer is, "How do people implement belief updating?" (Ma et al. 2006). Note that, by a suitable choice of variables and probabilistic relations, any sequence of belief states can be viewed as resulting from a Bayesian update (cf. Albert 2001). But say that we have independently motivated our starting points and found a convincing fit with the behavioral data of the belief dynamics (e.g., Brown et al. 2009). J&L then seem to suggest how this model might be given a mechanistic underpinning when they say that "belief updating of Bayes' Rule [amounts] to nothing more than vote counting" (sect. 7, para. 3). To us, the vote-counting idea seems just about right, since vote counting is about all that neurons can do if they are supposed to be ultimately implementing the process. We would add that mechanisms might also support the Bayesian account by providing independent motivations for choosing the variables and relations that make up the model.

Another good question is, "Why do people deviate from optimality in circumstance X?" The Bayesian model cannot explain such deviations directly, since it presupposes optimality. However, without a clear definition of optimality, as given by the Bayesian model, it would be impossible to detect or define such deviations in the first place: Without the presence of rationality, the concept of bounded rationality cannot exist. What's

more, suboptimal behavior can be elucidated by giving it the semblance of optimality within a Bayesian model. Those models then suggest what potentially irrational assumptions real agents make; the Bayesian models of reasoning behavior (Oaksford & Chater 2007) are a case in point.

J&L are not satisfied by this type of mechanistic support for Bayesian models; they argue that mechanistic theories should *constrain* the Bayesian model. However, it is unclear why exactly we should believe this. Surely, it does not matter for the empirical adequacy of the Bayesian process models whether peoples' beliefs are physically realized as activation networks in their frontal lobe, as global properties of their brain states, or as bursts of currents running in their big left toe. What matters is that the behavioral data are fitted within an independently motivated and predictively accurate model. In fact, if it turned out that dualism were correct after all, and belief revision actually went on in Cartesian mental stuff, that would not hurt the Bayesian analysis one bit – as long as the mental stuff updated its beliefs properly. Thus, the relation between Bayesian explanation and mechanistic accounts is asymmetric: While the finding that there is a mechanistic realization of Bayesian belief revision supports the Bayesian view, not finding such a mechanistic realization does not refute the theory. The only facts that can refute the Bayesian explanation are empirical facts about human behavior.

More varieties of Bayesian theories, but no enlightenment

doi:10.1017/S0140525X11000227

Jeffrey S. Bowers[a] and Colin J. Davis[b]

[a]*School of Psychology, University of Bristol, Clifton, Bristol BS8 1TU, United Kingdom;* [b]*Department of Psychology, Royal Holloway University of London, Egham Hill TW20 0EX, and School of Psychology, University of Bristol, Clifton, Bristol BS8 1TU, United Kingdom.*

j.bowers@bris.ac.uk **c.davis@rhul.ac.uk**

http://psychology.psy.bris.ac.uk/people/jeffbowers.htm

http://www.pc.rhul.ac.uk/staff/c.davis/

Abstract: We argue that Bayesian models are best categorized as *methodological* or *theoretical*. That is, models are used as tools to constrain theories, with no commitment to the processes that mediate cognition, or models are intended to approximate the underlying algorithmic solutions. We argue that both approaches are flawed, and that the Enlightened Bayesian approach is unlikely to help.

We agree with many points raised by Jones & Love (J&L) in the target article, but do not think that their taxonomy captures the most important division between different Bayesian approaches; and we question their optimism regarding the promise of the Enlightened Bayesian approach.

In our view, the critical distinction between Bayesian models is whether they are being used as a tool or a theory, what we have called the Methodological and Theoretical Bayesian approaches, respectively (Bowers & Davis, submitted). According to the Methodological approach, Bayesian models are thought to provide a measure of optimal performance that serves as a benchmark against which to compare actual performance. Researchers adopting this perspective highlight how often human performance is near optimal, and such findings are held to be useful for constraining a theory. (Whatever algorithm the mind uses, it needs to support behaviour that approximates optimal performance.) But there is no commitment to the claim that the algorithms that support perception, cognition, and behaviour approximate Bayesian computations.

By contrast, according to the Theoretical approach, the mind is claimed to carry out (or closely approximate) Bayesian analyses at the algorithmic level; this perspective can be contrasted with the view that the mind is a rag-bag of heuristics. For example, when describing the near-optimal performance of participants in making predictions about uncertain events, Griffiths and Tenenbaum (2006) write: "These results are inconsistent with claims that cognitive judgments are based on non-Bayesian heuristics" (p. 770).

Unfortunately, it is not always clear whether theorists are adopting the Methodological or the Theoretical approach, and at times, the same theorists endorse the different approaches in different contexts. Nevertheless, this is the key distinction that needs to be appreciated in order to understand what claims are being advanced, as well as to evaluate theories. That is, if Bayesian models are used as a tool to constrain theories, then the key question is whether this tool provides constraints above and beyond previous methods. By contrast, if the claim is that performance is supported by Bayesian-like algorithms, then it is necessary to show that Bayesian theories are more successful than non-Bayesian theories.

In our view there are two main problems with the Methodological Bayesian approach. First, measures of optimality are often compromised by the fact Bayesian models are frequently constrained by performance. For instance, Weiss et al. (2002) developed a Bayesian model of motion perception that accounts for an illusion of speed: Objects appear to move more slowly under low-contrast conditions. In order to accommodate these findings, Weiss et al. assumed that objects tend to move slowly in the world, and this prior plays a more important role under poor viewing conditions. One problem with this account, however, is that there are other conditions under which objects appear to move more quickly than they really are (Thompson et al. 2006). Stocker and Simoncelli's (2006) response to this problem is to note that their Bayesian theory of speed perception could account for the latter phenomenon as well:

> [I]f our data were to show increases in perceived speed for low-contrast high-speed stimuli, the Bayesian model described here would be able to fit these behaviors with a prior that increases at high speeds. (Stocker & Simoncelli 2006, p. 583)

The modification of Bayesian models in response to the data is widespread, and this renders the models more as descriptions of behaviour than as tools with which to measure optimality.

Second, even if a Bayesian model provides a good measure of optimal performance, it is not clear how the tool contributes to constraining theories. Under these conditions, a model can be supported or rejected because it does or does not match optimal performance, or more simply, a model can be supported or rejected because it does or does not capture human performance. The match or mismatch to data is sufficient to evaluate the model – the extra step of comparing to optimal performance is superfluous.

With regard to the Theoretical Bayesian approach, the key question is whether a Bayesian model does a better job in accounting for behaviour compared to non-Bayesian alternatives. However, this is rarely considered. Instead, proponents of this approach take the successful predictions of a Bayesian model as support for their approach, and often ignore the fact that non-Bayesian theories might account for the data just as well. We are not aware of any psychological data that better fit a Bayesian as compared to a non-Bayesian alternative.

What about the promise of the Bayesian Enlightenment approach? On our reading, this perspective encompasses both the theories that we would call Methodological (e.g., the adaptive heuristic approach of Gigerenzer), and the theories that we would call Theoretical (e.g., demonstrations that Bayesian computations can be implemented in neural wetware are considered Enlightened). Thus, the above criticisms apply to the Bayesian Enlightenment approach as well.

With regard to Enlightened theories that take the form of heuristics, it is not clear that Bayesian models are providing any constraints. For example, we are not aware of any instance

in which Gigerenzer and colleagues used a Bayesian model in order to constrain their heuristic solution, and we are not sure how in practice this could help in the future. The underlying processes of Bayesian models and heuristics are as different as could be, and unless there are cases in which a Bayesian model provides important constraints on heuristic theories above and beyond the data, we do not see the point.

With regard to Enlightened models of neural computation, there is no evidence that neurons actually compute in a Bayesian manner. Almost all the evidence taken to support this view is behavioural, with the computational neuroscience largely devoted to providing existence proofs that Bayesian computations in brain are possible. Accordingly, alternative computational solutions might equally account for the relevant data. More generally, J&L argue that an Enlightened Bayesian model looks for optimal solutions, given a set of representations and processes. However, we are unclear how this approach adds to the more traditional approach to science, namely, evaluating how well a specific implemented model accounts for performance.

The imaginary fundamentalists: The unshocking truth about Bayesian cognitive science

doi:10.1017/S0140525X11000239

Nick Chater,[a] Noah Goodman,[b] Thomas L. Griffiths,[c] Charles Kemp,[d] Mike Oaksford,[e] and Joshua B. Tenenbaum[f]

[a]*Behavioural Science Group, Warwick Business School, University of Warwick, Coventry CV4 7AL, United Kingdom;* [b]*Department of Psychology, Stanford University, Stanford, CA 94305;* [c]*Department of Psychology, University of California, Berkeley, CA 94720-1650;* [d]*Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213;* [e]*Department of Psychological Sciences, Birkbeck College, University of London, London WC1E 7HX, United Kingdom;* [f]*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139.*

Nick.chater@wbs.ac.uk ngoodman@stanford.edu
tom_griffiths@berkeley.edu ckemp@cmu.edu
m.oaksford@bbk.ac.uk jbt@mit.edu
http://www.wbs.ac.uk/faculty/members/Nick/Chater
http://stanford.edu/~ngoodman/
http://psychology.berkeley.edu/faculty/profiles/tgriffiths.html
http://www.charleskemp.com
http://www.bbk.ac.uk/psychology/our-staff/academic/mike-oaksford http://web.mit.edu/cocosci/josh.html

Abstract: If Bayesian Fundamentalism existed, Jones & Love's (J&L's) arguments would provide a necessary corrective. But it does not. Bayesian cognitive science is deeply concerned with characterizing algorithms and representations, and, ultimately, implementations in neural circuits; it pays close attention to environmental structure and the constraints of behavioral data, when available; and it rigorously compares multiple models, both within and across papers. J&L's recommendation of Bayesian Enlightenment corresponds to past, present, and, we hope, future practice in Bayesian cognitive science.

The Bayesian Fundamentalist, as described by Jones & Love (J&L), is an alarming figure. Driven by an unshakeable assumption that all and every aspect of cognition can be explained as optimal, given the appropriate use of Bayes' rule, the fearsome fundamentalist casts aside questions about representation and procesFs, pays scant attention to the environment, and is relatively unconcerned with empirical data or model comparison; the fundamentalist launches an assault on the mind, armed only with complex mathematics and elaborate computational models. J&L suggest that cognitive science should shun this extreme and bizarre position, and instead embrace Bayesian Enlightenment. The latter is a moderate doctrine, which sees Bayesian computational explanation as one of a number of mutually constraining levels of explanation of the mind and brain, pays attention to representation and process and the structure of the environment, and compares explanatory models with empirical data and each other.

Readers new to Bayesian cognitive science may find the argument persuasive. The curious doctrine of Bayesian Fundamentalism is surely a "bad thing," and Bayesian Enlightenment is clearly preferable. Such readers will, while being grateful to J&L for forewarning them against the perils and pitfalls of Bayesian Fundamentalism, also wonder how a viewpoint as radical and peculiar as Bayesian Fundamentalism ever became established in the first place.

The truth is that it didn't. To our knowledge, Bayesian Fundamentalism is purely a construct of J&L's imagination. There are no Bayesian Fundamentalists and never have been. There is, to be sure, a large literature on Bayesian cognitive science. Bayesian Fundamentalists appear nowhere within it. This is where the reader of J&L new to Bayesian cognitive science is liable to be led astray.

We agree with J&L that Enlightened Bayesians are commendable; and that Fundamentalist Bayesians, if they existed, would be deplorable. But Bayesian Enlightenment, rather than Bayesian Fundamentalism, is and has always been the norm in Bayesian cognitive science.

Our discussion has four parts. First, we clarify some technical inaccuracies in J&L's characterization of the Bayesian approach. Second, we briefly note that Bayesian Fundamentalism differs from the actual practice of cognitive science along a number of dimensions. Third, we outline the importance and potential explanatory power of Bayesian computational-level explanation. Fourth, we suggest that one characteristic of the Bayesian approach to cognitive science is its emphasis on a top-down, function-first approach to psychological explanation.

1. What is Bayes? In the target article, J&L worry that Bayesian inference is conceptually trivial, although its consequences may be complex. The same could be said of all mathematical science: The axioms are always "trivial"; the theorems and implications are substantive, as are the possibilities for engineering nontrivial systems using that mathematics as the base.. J&L focus their attention on Bayes' rule, but this is just the starting point for the approach, not its core. The essence of Bayes is the commitment to representing degrees of belief with the calculus of probability. By adopting appropriate representations of a problem in terms of random variables and probabilistic dependencies between them, probability theory and its decision-theoretic extensions offer a unifying framework for understanding all aspects of cognition that can be properly understood as inference under uncertainty: perception, learning, reasoning, language comprehension and production, social cognition, action planning and motor control, as well as innumerable real-world tasks that require the integration of these capacities. The Bayesian framework provides a principled approach to solving basic inductive challenges that arise throughout cognition (Griffiths et al. 2008a; Tenenbaum et al. 2011), such as the problem of trading off simplicity and fit to data in model evaluation, via the Bayesian Occam's razor (MacKay 2002) or the problem of developing appropriate domain-specific inductive biases for constraining learning and inference, via hierarchical Bayesian models (Gelman et al. 2003).

Bayes' rule is the most familiar and most concrete form in which psychologists typically encounter Bayesian inference, so it is often where Bayesian modelers start as well. But interpreted literally as the form of a computational model – what we take to be J&L's target when they refer to Bayes' rule as a "simple vote-counting scheme" (sect. 3, para. 9) – the form of Bayes' rule J&L employ applies to only the simplest tasks requiring an agent to evaluate two or more mutually exclusive discrete hypotheses posited to explain observed data. Some of the earliest

Bayesian models of cognition did focus on these cases; starting with the simplest and most familiar settings is often a good research strategy. But most of cognition cannot be directly cast in such a simple form, and this has been increasingly reflected in Bayesian cognitive models over the last decade. Indeed, the form of Bayes' rule that J&L discuss hardly figures in many contemporary Bayesian cognitive models.

What does it mean in practice for a computational model of cognition to be Bayesian, if not to literally implement Bayes' rule as a mechanism of inference? Typically, it means to adopt algorithms for generating hypotheses with high posterior probabilities based on Monte Carlo sampling, or algorithms for estimating the hypothesis with highest posterior probability (i.e., maximum a posteriori probability [MAP]) using local message-passing schemes (MacKay 2002). The outputs of these algorithms can be shown, under certain conditions, to give reasonable approximations to fully Bayesian inference, but can scale up to much larger and more complex problems than could be solved by exhaustively scoring all possible hypotheses according to Bayes' rule (J&L's "simple vote-counting scheme"). A little further on we briefly discuss several examples of how these approximate inference algorithms have been explored as models of how Bayesian computations might be implemented in the mind and brain.

2. Bayesian Fundamentalism versus Bayesian cognitive science. J&L charge Bayesian Fundamentalists with a number of failings. The practice of Bayesian cognitive science is largely free of these, as we will see.

(i) J&L suggest in their Introduction that "[i]t is extremely rare to find a comparison among alternative Bayesian models of the same task to determine which is most consistent with empirical data" (sect. 1, para. 6). Yet such comparisons are commonplace (for a tiny sample, see Goodman et al. 2007; Griffiths & Tenenbaum 2009; Kemp & Tenenbaum 2009; Oaksford & Chater 2003); Goo. Nonetheless, of course, Bayesian authors do sometimes press for a single model, often comparing against non-Bayesian alternative accounts (e.g., Goodman et al. 2008b). This is entirely in line with practice in other modeling frameworks.

(ii) J&L are concerned that Bayesians downplay the structure of the environment. This is a particularly surprising challenge given that Anderson's path-breaking Bayesian rational analyses of cognition (e.g., Anderson 1990; 1991a; Oaksford & Chater 1998b) are explicitly based on assumptions about environmental structure. Similarly, Bayesian approaches to vision essentially involve careful analysis of the structure of the visual environment – indeed, this defines the "inverse problem" that the visual system faces (e.g., Yuille & Kersten 2006); and Bayesian models of reasoning are crucially dependent on environmental assumptions, such as "rarity" (Oaksford & Chater 1994). Finally, in the context of language acquisition, there has been substantial theoretical and empirical progress in determining how learning depends on details of the "linguistic environment," which determine the linguistic structures to be acquired (Chater & Vitányi 2007; Foraker et al. 2009; Hsu & Chater 2010; Hsu et al., in press; Perfors et al. 2010; 2011).

(iii) J&L claim (in sect. 4) that Bayesian Fundamentalism is analogous to Behaviorism, because it "eschews mechanism" (sect. 2.2, para. 3). But, as J&L note, Bayesian cognitive science, qua cognitive science, is committed to *computational* explanation; behaviorists believe that no such computations exist, and further that there are no internal mental states over which such computations might be defined. Assimilating such diametrically opposing viewpoints obscures, rather than illuminates, the theoretical landscape.

(iv) J&L suggest, moreover, that Bayesians are unconcerned with representation and process, and that the Bayesian approach is driven merely by technical advances in statistics and machine learning. This seems to us completely backwards: Most of the technical advances have precisely been to enrich the range of representations over which Bayesian methods can operate (e.g., Goodman et al. 2011; Heller et al. 2009; Kemp et al. 2010a; 2010b) and/or to develop new computational methods for efficient Bayesian inference and learning. These developments have substantially expanded the range of possible hypotheses concerning representations and algorithms in human inference and learning. Moreover, some of these hypotheses have provided new mechanistic accounts. For example, Sanborn et al. (2010a, p.1144) have argued that "Monte Carlo methods provide a source of 'rational process models' that connect optimal solutions to psychological processes"; related approaches are being explored in a range of recent work (e.g., Vul et al. 2009a; 2009b). Moreover, there has been considerable interest in how traditional psychological mechanisms, such as exemplar models (Shi et al. 2010) and neural networks (e.g., McClelland 1998; Neal 1992), may be viewed as performing approximate Bayesian inference. Such accounts have been applied to psychological data on, for example, conditional reasoning (Oaksford & Chater 2010).

We have argued that Bayesian cognitive science as a whole is closely involved both with understanding representation and processes and with specifying environmental structure. Of course, individual Bayesian projects may not address all levels of explanation, and so forth. We believe it would be unnecessary (and pernicious) to require each project to embrace all aspects of cognition. (For instance, we would not require all connectionist models to make explicit the bridge to biological neural networks.) Indeed, according to the normal canons of scientific inference, the more that can be explained, with the fewer assumptions, the better. Thus, contra J&L, we see it as a strength, rather than weakness, of the Bayesian approach that some computational-level analyses have broad applications across cognition, independent of specific representational, processing, or environmental assumptions, as we now explore.

3. The power of Bayesian computational-level explanation: The case of explaining away. Consider the Bayesian analysis of *explaining away* (Pearl 1988). Suppose two independent causes (e.g., *no petrol* or *dead battery*) can cause a car not to start. Learning that the car did not start then raises the probability of both *no petrol* and *dead battery*: they both provide potential explanations for the car not starting. But if we then learn that the battery was dead, the probability of *no petrol* falls back to its original value. The battery explains the car not starting; so the apparent evidence that there might be no petrol is "explained away."

Experiments have found that, when given reasoning problems with verbal materials, people do, indeed, follow this, and related, patterns of reasoning (e.g., Ali et al. 2011), although this pattern is clearer in young children (Ali et al. 2010), with adults imposing additional knowledge of causal structure (Walsh & Sloman 2008; Rehder & Burnett 2005). Moreover, the same pattern is ubiquitous in perception: If a piece of sensory input is explained as part of one pattern, it does not provide evidence for another pattern. This principle emerges automatically from Bayesian models of perception (Yuille & Kersten 2006).

Furthermore, explaining away also appears to help understand how children and adults *learn* about causal regularities (e.g., Gopnik et al. 2004; Griffiths & Tenenbaum 2009). If a "blicket detector" is triggered whenever *A* and *B* are present, there is a prima facie case that *A* and/or *B* causes the detector to sound. But if the detector also sounds when preceded only by *A*, then this regularity explains away the sounding of the detector and reduces the presumed causal powers of *B*. In animal learning, a related pattern is known as *blocking* (Kamin 1969).

Blocking can also be explained using connectionist-style mechanistic models, such as the Rescorla-Wagner model of error-driven associative learning (Rescorla & Wagner 1972). But such explanations fail to capture the fact that partial reinforcement (i.e., where the putative effect only sometimes

follows the putative cause) extinguishes more slowly than total reinforcement. Indeed, partial reinforcement should induce a weak link which should more easily be eliminated. From a Bayesian point of view, extinction in partial reinforcement is slower, because the lack of effect must occur many times before there is good evidence that the state of the world has really changed (e.g., a causal link has been broken). This type of Bayesian analysis has led to a wide range of models of human and animal learning, which are both compared extensively with each other and with empirical data (for a review, see Courville et al. 2006). Associative learning accounts of blocking also cannot explain the rapid and complex dynamics observed in adults' and children's causal learning: the fact that causal powers may be identified from just one or a few observed events in the presence of appropriate background knowledge about possible causal mechanisms, and the strong dependence of the magnitude of causal discounting on the base rates of causes in the environment (Griffiths & Tenenbaum 2009). In contrast, these phenomena are not only explained by, but were predicted by and then experimentally verified from, the dynamics of explaining away in Bayesian analyses of causal learning.

We have seen that a general qualitative principle, *explaining away*, which follows directly from the mathematics of probability, has broad explanatory power across different areas of cognition. This generality is possible precisely because the Bayesian analysis abstracts away from mechanism – which presumably differs in detail between verbal reasoning, perception, and human and animal learning. Thus, contra J&L, the Bayesian approach is not merely closely tied with empirical data; it provides a synthesis across apparently unconnected empirical phenomena, which might otherwise be explained by using entirely different principles.

Framing explanations of some phenomena at this high level of abstraction does not imply commitment to any kind of Bayesian Fundamentalism. Rather, Bayesian cognitive scientists are merely following the standard scientific practice of framing explanation at the level of generality appropriate to the phenomena under consideration. Thus, the details, across computational, algorithmic, and implementation levels, of accounts of animal learning, perception, or causal reasoning will differ profoundly – but the phenomenon of "explaining away" can insightfully be seen as applying across domains. This aspect of explanation is ubiquitous across the sciences: For example, an abstract principle such as the conservation of energy provides a unified insight across a wide range of physical phenomena; yet the application of such an abstract principle in no way detracts from the importance of building detailed models of individual physical systems.

4. Bayesian cognitive science as a top-down research strategy. Bayesian cognitive scientists strongly agree with J&L that it is vital to create mutually constraining accounts of cognition across each of Marr's computational levels of explanation. We stress that what is distinctive about the Bayesian approach, in distinction from many traditional process models in cognitive psychology, is a top-down, or "function-first" research strategy, as recommended by Marr (1982): from computational, to algorithmic, to implementational levels (see, e.g., Anderson 1990; Chater et al. 2003; Griffiths et al. 2010).

The motivation for this approach is tactical, rather than ideological. Consider attempting to understand an alien being's pocket calculator that uses input and output symbols we don't understand. If we realize that an object is doing addition (computational level), we have some chance of discerning which type of representations and algorithms (algorithmic level) might be in play; it is hard to see how any amount of study of the algorithmic level alone might lead to inferences in the opposite direction. Indeed, it is difficult to imagine how much progress could be made in understanding an algorithm, without an understanding of what that algorithm is computing.

Thus, the problem of *reverse engineering* a computational system, including the human mind, seems to inevitably move primarily from function to mechanism. Of course, constraints between levels will flow in both directions (Chater & Oaksford 1990). The hardware of the brain will place strong constraints on what algorithms can be computed (e.g., Feldman & Ballard 1982), and the possible algorithms will place strong constraints on what computational-level problems can be solved or approximated (Garey & Johnson 1979). Yet, from this reverse-engineering perspective, the first task of the cognitive scientist is to specify the nature of the computational problem that the cognitive system faces, and how such problems might, in principle, be solved. This specification typically requires, moreover, describing the structured environment, the goal of the cognitive system, and, frequently, computational constraints or representational commitments (Anderson 1990; Oaksford & Chater 1998b).

The appropriate mathematical frameworks used for this description cannot, of course, be determined a priori, and will depend on the nature of the problem to be solved. Analyzing the problem of moving a multi-jointed motor system might, for example, require invoking, among other things, tensor calculus and differential geometry (which J&L mention as important to developments in physics). A rational analysis of aspects of early auditory and visual signal processing might invoke Fourier analysis or wavelet transforms. A computational-level analysis of language use might involve the application of symbolic grammatical and computational formalism. In each case, the appropriate formalism is also open to challenge: For example, researchers differ widely concerning the appropriate grammatical or logical formalism required to represent language and thought; or, indeed, as to whether symbolic formalism is even required at all (e.g., McClelland 2010).

Within this diversity, there is an important common mathematical thread. A wide range of cognitive problems, from motor control to perception, language processing, and commonsense reasoning, involve (among other things) making inferences with uncertain information, for which probability theory is a natural mathematical framework. For example, the problem of finding an underlying pattern in a mass of sensory data – whether that pattern be the layout of the environment, a set of causal dependencies, the words, syntactic structure, or meaning of a sentence, or even the grammatical structure of a language – is naturally framed in terms of probabilistic (or Bayesian) inference. This explains why probability is a common theme in Bayesian modeling, and why engineering approaches to solving these and many other problems often take a Bayesian approach (though there are important alternatives) (Bishop 1996; Manning & Schütze, 1999; Russell & Norvig 2011). Indeed, Bayesian cognitive scientists have themselves contributed to extending the boundaries of engineering applications in some domains (e.g., Griffiths & Ghahramani 2006; Johnson et al. 2007; Kemp & Tenenbaum 2008; Kemp et al. 2006; Goodman et al. 2008a).

J&L are concerned that a close relationship between hypotheses in Bayesian cognitive science and technical/mathematical developments in engineering (broadly construed to include statistics and computer science) may amount to a confusion of "technical advances with theoretical progress" (sect. 1, para. 3). We suggest, by contrast, that theoretical approaches in cognitive science that are not tied to rich technical developments have little chance of success. Indeed, given that the human mind/brain is the most complex mechanism known, and that its information-processing capacities far outstrip current artificial intelligence, it is surely inevitable that, in the long term, successful reverse engineering will be possible only in the light of spectacular technical developments, alongside careful use of empirical data. We suggest that Bayesian cognitive science promises to be a small forward step along this path.

Keeping Bayesian models rational: The need for an account of algorithmic rationality

doi:10.1017/S0140525X11000240

David Danks[a] and Frederick Eberhardt[b]

[a]*Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213;* [b]*Philosophy–Neuroscience–Psychology Program, Washington University in St. Louis, St. Louis, MO 63130.*

ddanks@cmu.edu **eberhardt@wustl.edu**
http://www.hss.cmu.edu/philosophy/faculty-danks.php
http://www.artsci.wustl.edu/~feberhar/

Abstract: We argue that the authors' call to integrate Bayesian models more strongly with algorithmic- and implementational-level models must go hand in hand with a call for a fully developed account of algorithmic rationality. Without such an account, the integration of levels would come at the expense of the explanatory benefit that rational models provide.

We commend Jones & Love's (J&L's) call for a greater integration of "Fundamentalist" Bayesian models with algorithmic- and implementational-level models in psychology. We agree that such integrations would significantly improve the empirical testability of Bayesian models, particularly in light of the currently unsatisfactory fit of Bayesian models with normative accounts of rational inference and choice. At the same time, we believe that this call for integration must be accompanied by a call for the development of new, complementary accounts of rationality for the algorithmic and implementational levels.

The target article might be read (mistakenly, in our view) as a complete surrender of the rational or computational level of modeling based on the difficulty of testing Bayesian models empirically. J&L suggest that rational Bayesian models can be tested only if the models are subject to constraints that span across the three levels – constraints which are not provided by extant Bayesian models. The problem is exacerbated by the apparent gap between standard accounts of rationality supporting the Bayesian paradigm (such as diachronic Dutch book arguments) and the empirical evidence used to confirm Bayesian models (see, e.g., Eberhardt & Danks [2011] for a concrete example). The appropriate remedy, J&L suggest, is to tie the Bayesian models to the algorithmic- and implementational-level models. But on its own, such an integration would provide only a reduction of the computational level to the algorithmic and implementational levels, and so would relinquish the explanatory benefits the computational level provides.

Bayesian models have achieved widespread popularity in part because, as computational-level models, they describe an inference procedure that is *rational* or *optimal* given the agent's task. Bayesian models purport to provide a certain type of explanation – an optimality explanation, rather than a mechanistic one – while not being committed to a particular algorithm or implementation in the learner's brain. As already noted, however, these "explanations" are undermined by the misfit between empirical results and standard accounts of rationality. The common response in the so-called Fundamentalist Bayesian literature (to the extent that there is one) has been that rational behavior is defined on a case-by-case basis, where the appropriate principles of rationality depend on the learner's particular context (see, e.g., Oaksford & Chater 2007). Obviously, from a normative perspective, such claims are unsatisfactory: We need an account of rational principles that are fixed independently of the learner's behavior.

We contend that the target article's call for a closer integration of the computational and algorithmic levels of modeling provides an ideal opportunity: Rather than attempting to develop an account of rational behavior independent of the learner's cognitive infrastructure, an integration of levels should go hand in hand with the development of an account of *algorithmic rationality*. That is, we require an account of rationality that is informed by the learner's constraints at the implementational and algorithmic levels. By providing such an integrated account of rationality, we can resist the purely reductionist view in which a computational-level model is ultimately assessed only on the basis of its success at the algorithmic and implementational levels, while preserving the (additional, distinctive) explanatory power provided by rational models for *all* levels. (J&L's discussion of a possible "Bayesian Enlightenment" is, we think, closely related.)

What would an account of algorithmic rationality look like? Often, suggestions along these lines are made under the heading of *bounded rationality*. In many cases, however, models of bounded rationality (e.g., Simon's "satisficing," or Gigerenzer's "fast and frugal heuristics") propose boundedness constraints that are not informed by *particular* implementational constraints of the learner, but rather are motivated by abstract limitations or efficiencies of the computational procedure. Nevertheless, they are on the right track. Accounts of algorithmic rationality would have to provide support for models that are empirically adequate, and must approximate behavior that is deemed ideally rational (e.g., maximization of expected utility, avoidance of Dutch book, etc.). At the same time, accounts of algorithmic rationality must integrate known implementational (i.e., biological) constraints. The resulting models would be neither just computational nor just algorithmic, but would instead provide an account that is both descriptively adequate and normatively supported. Confirmation of such a model would require both empirical confirmation at multiple levels (e.g., matching the right input-output relation, predicting that neurons will fire in particular places at particular times), as well as an account of why this behavior is the rational or optimal thing to do in light of the given task *and* the cognitive and neural infrastructures.

We thus argue that there needs to be not only a "tying down" of the rational models to algorithmic- and implementational-level constraints, but also a "pulling up" of algorithmic-level models to go beyond purely descriptive accounts (that occasionally come paired with evolutionary "how-possibly" stories) to explanations of why the particular algorithm or implementation is the right one for the task. The fundamental insight of Newell and Simon, Pylyshyn, and Marr, that there are different levels of explanation, has led to enormous productivity at each of the different levels. J&L are right to point out that these efforts have drifted apart: Models at different levels are frequently no longer informative about other levels. However, a reintegration that simply reduces one of the levels to the others would abandon one of the distinctive forms of explanation identified in that original insight. We urge instead that any reintegration must include the development of an account of algorithmic-level rationality, which then permits us to determine whether Bayesian models remain successful, both empirically and theoretically.

Survival in a world of probable objects: A fundamental reason for Bayesian enlightenment

doi:10.1017/S0140525X11000252

Shimon Edelman and Reza Shahbazi

Department of Psychology, Cornell University, Ithaca, NY 14853.

edelman@cornell.edu **rs689@cornell.edu**
http://kybele.psych.cornell.edu/~edelman

Abstract: The only viable formulation of perception, thinking, and action under uncertainty is statistical inference, and the normative way of statistical inference is Bayesian. No wonder, then, that even seemingly non-Bayesian computational frameworks in cognitive science ultimately draw their justification from Bayesian considerations, as enlightened theorists know fully well.

Setting up "Bayesian Fundamentalism" as a straw man for criticism, which is what Jones & Love (J&L) do in their target article, is counterproductive for two related reasons. The first reason is both substantive and didactic. The only viable general formulation of perception, cognition, and action is statistical inference under uncertainty, which proceeds from prior experience to probabilistic models that can support principled decision-making and control, and thereby make informed behavior possible. Branding the realization that the modus operandi of any mind is fundamentally Bayesian as "fundamentalist" is akin to condemning Hamiltonian mechanics for its reliance on a single overarching principle. This is not a good idea, if even a single reader decides after reading the target article that some kind of pluralism ideal should trump a genuine insight into the workings of the world.

In the Enlightenment-era natural philosophy and in the sciences that grew out of it, the insight that singles out statistics as the only possible conceptual foundation for a sound theory of how the mind works can be traced back to Hume's *Treatise*: "All knowledge resolves itself into probability" (Hume 1740, Part IV, sect. I). More recently, J. J. Gibson (1957) noted in a book review, from which we borrowed the title of the present commentary, that "perception is always a wager." By now, it has become incontrovertibly clear that not only perception, but thinking and action too, indeed take the form of statistical inference (Glimcher et al. 2008; Heit 2000; Knill & Richards 1996; Körding & Wolpert 2006; for a brief overview, see Chater et al. 2006).

Statistical inference is indispensable in cognition because informed behavior requires a degree of foresight; that is, anticipating and planning for the future. Foresight, in turn, is possible because the world is well-behaved (the past resembles the future often enough to support learning from experience), but can be only probabilistic because the regularities in the world's behavior manifest themselves as merely statistical patterns in sensorimotor data (Edelman 2008b).

The rich theory of statistical learning from experience developed over the past several decades, all of which ultimately relies on this view of cognition, is very diverse. Many of the techniques that it encompasses have been inspired by particular classes of problems, or solution methods, that on the face of it have no bearing on the Bayesian debate. For instance, insofar as perception is ill-posed in the formal sense that a given problem typically admits multiple solutions, it needs to be regularized by adding extra constraints (Poggio 1990; cf. Tikhonov & Arsenin 1977). Likewise, modeling problems in motor control can be approached via function approximation (Mussa-Ivaldi & Giszter 1992) and control theory (Kawato 1999).

At the same time, because the *normative* way of performing statistical inference is Bayesian (Howson & Urbach 1991; Schervish 1995; Wasserman 2003), one would expect that all such methods would ultimately reduce to Bayesian inference, and indeed they do. In particular, perceptual regularization can be given a statistical formulation (Marroquin et al. 1987), and so can function approximation – be it classification or regression (Bishop 2006; Hastie et al. 2003/2009). More generally, various popular theoretical approaches in cognitive science, too numerous to mention here, which may appear refreshingly pluralistic when contrasted to the big bad Bayes, do in fact draw their justification from Bayesian considerations (Bishop 2006; Hastie et al. 2003/2009).

This brings us to the second reason to shun the distinction between Bayesian Fundamentalism and Bayesian Enlightenment, which is that the latter builds upon the former rather than supplanting it. It used to be possible to produce good work in computer vision or natural language engineering without giving much thought to the ultimate justification of the methods one uses, and it may still be the case in some isolated corners of those fields. Likewise, in cognitive modeling, where theoretical confusion – especially conflation of Marr's levels of understanding – is often endemic (cf. Edelman 2008a; 2008c), modelers may happily use "backpropagation networks" without giving much thought to the fact that they are doing function approximation, which in turn rests on Bayesian principles. Although it is always good to know exactly why your model works as it does (or does not), in practice focusing on the more problematic levels, such as coming up with a good theory of the hypothesis space that applies to the learning task at hand, may actually lead to progress that would elude someone who is concerned exclusively with the fundamentals.

Thus, Bayesian theorists worth their salt (do we need to worry about others?) know that a full-fledged theory has many aspects to it that must be addressed, including, in addition to the just-mentioned issue of the structure of the hypothesis space, various questions regarding the nature of inference and decision-making algorithms that have to be computationally tractable and biologically feasible. Crucially, they also know that the very need for statistical inference and the constraint of having to deal incrementally with a stream of environmental data are non-negotiable. If this earns them the label of fundamentalism, we should all be eager to share it with them, normatively. After all, we owe to it our survival.

Don't throw out the Bayes with the bathwater

doi:10.1017/S0140525X11000264

Philip M. Fernbach and Steven A. Sloman
Department of Cognitive, Linguistic and Psychological Sciences, Brown University, Providence, RI 02912.
philip_fernbach@brown.edu **steven_sloman@brown.edu**
http://sites.google.com/site/philfernbachswebpage/
http://sites.google.com/site/slomanlab/

Abstract: We highlight one way in which Jones & Love (J&L) misconstrue the Bayesian program: Bayesian models do not represent a rejection of mechanism. This mischaracterization obscures the valid criticisms in their article. We conclude that computational-level Bayesian modeling should not be rejected or discouraged a priori, but should be held to the same empirical standards as other models.

There is an important point in this target article by Jones & Love (J&L); unfortunately it is hard to see as it is swimming below some very red herrings. Bayesians do not reject mechanism, mental representation, or psychological or neural process. Comparisons to behaviorism are thus misleading. Moreover, the utility of particular Bayesian models does not rest on whether they engage with mechanistic issues, nor even whether they are concerned with issues of mechanism at all; it rests on what new light the Bayesian analysis sheds on the problem, and this depends on the same principles of evaluation that should be applied to any theory. There is a meaningful debate to be had about the value of Bayesian modeling, but ironically, given the title of the article, we fear that J&L's wholesale rejection of "Bayesian Fundamentalism" will lead to arguments about whose religion is better. In geopolitics, these kinds of disputes often produce "intractable conflicts" and do not lead anywhere productive (Bar-Tal 2002). In this commentary we first highlight one way in which J&L's analysis goes awry and clarify what we believe the debate should *not* be about. We then turn to what the debate should be about and suggest some prescriptions for good Bayesian practice.

A central pillar of J&L's argument is the claim that Bayesian modeling at the computational level represents a rejection of mechanism and psychological process in favor of a purely mathematical analysis. This is a straw man that conflates the rejection of psychological representations and processes with what Bayesian models really do, which is abstract over cognitive processes. Rejecting mechanism would indeed imply that there is nothing to a Bayesian model except for the "conceptually trivial" (according to J&L; see target article, Abstract) mathematical relation between inputs and outputs. This idea is reflected in the puzzling language that the authors use to explain the nature of the hypothesis space. They write, "In general, a hypothesis is nothing more than a probability distribution" (sect. 3, para. 2, repeated in para. 9). It is true that any Bayesian model is committed to the assumption that it is coherent to represent a probability distribution over the hypothesis space, but this does not mean that a hypothesis is a probability distribution. Rather, in the Bayesian models we are familiar with, hypotheses are interpretable as psychological constructs, such as partitions of conceptual space (e.g., Tenenbaum & Griffiths 2001) or causal models (e.g., Steyvers et al. 2003). For that reason, a Bayesian model for a particular task is not conceptually trivial, but rather embodies an analysis specific to the mental models presumed to be relevant to that task. It is true that modelers are not always as clear as they should be about whether these hypotheses represent psychological entities or merely a conceptual analysis of the task (or both), and the import of the model does depend critically on that. But either way, such a model can be extremely useful. In the former case, it provides a descriptive hypothesis. In the latter, it may shed new light on how we should think about the task. Critically, none of this is contingent on whether the model engages with questions about reaction time, neurophysiological information, or other "mechanistic" data.

An important criticism that we believe should be taken away from the target article is that some Bayesian modelers fail to clearly distinguish normative from descriptive claims. Vagueness about this issue has led to confusion between two very different ideas: "rational" and "computational" (Sloman & Fernbach 2008). A *rational* model is one that explains data by showing how it reflects optimal behavior, given the task at hand (Anderson 1990). A *computational* model is one that describes the function that people are engaged in, whether or not that function is what they *should* be engaged in. We agree completely with J&L on this issue: The fact that a model is at the computational level is not an excuse for its failure to fit data. Violations of a model's predictions should be taken seriously and not explained away as due to the approximate way the optimal computation is implemented. And a rational analysis does not demonstrate rationality if people do not abide by it. The beauty of rational analysis is its ability to explain behavior by appealing to the structure of the world. When a model fails to be crystal clear about which of its parts reflect the world and which reflect mental or neural facts, its contribution becomes obscure.

This issue is not a problem with Bayesianism per se, but it does suggest principles for good Bayesian practice. For one, model comparisons and sensitivity analyses should be de rigueur to show whether the phenomenon falls uniquely out of the Bayesian model or whether the model is merely consistent with it. It is often helpful to a reader to explain precisely what in the model is generating the phenomenon of interest, and whether that aspect is inherent to the model's structure and would emerge from any parameterization, or whether it is due to an assumption. Modelers should be clearer about whether their model is a psychological theory or a normative analysis, and they should consider contradictory data conscientiously as opposed to explaining it away. Of course, these are just recommendations for good scientific practice and thus they hardly seem controversial. J&L are right to point out that Bayesian modelers sometimes ignore them.

Were these practices adopted more systematically, Bayesian models would become more falsifiable. This is a double-edged sword, of course: Given the reams of evidence that cognition is fallible, Bayesians are fighting an uphill battle (at least in high-order cognition). But only falsifiable theories can be truly revolutionary.

Osiander's psychology

doi:10.1017/S0140525X11000276

Clark Glymour

Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213.
cg09@andrew.cmu.edu

Abstract: Bayesian psychology follows an old instrumentalist tradition most infamously illustrated by Osiander's preface to Copernicus's masterpiece. Jones & Love's (J&L's) criticisms are, if anything, understated, and their proposals overoptimistic.

In his preface to Copernicus's masterpiece, Osiander wrote, "For these hypotheses need not be true or even probable; if they provide a calculus consistent with the observations, that alone is sufficient" (Copernicus 1995).

So far as I can tell, Bayesian psychologists are doing the kind of thing Osiander said astronomers should be doing. Bayesian psychologists collect data from experiments and in each case provide a model for the data. In these models, subjects have prior probabilities for various hypotheses and likelihoods for the data on each hypothesis. Stimuli change their degrees of belief according to Bayes' Rule, and then the experimental subjects somehow use the updated probabilities to perform some action. The minimal Bayesian psychological claim is just that various experiments on learning behavior can be accommodated in this way. That also seems to be the maximal claim. Bayesian psychologists – at least those I read and talk with – do not claim that their subjects are deliberately, consciously, carrying out a computation of posterior degrees of belief, and deliberately, consciously, using the result in a decision rule. In many cases that would be implausible, and in any case, if that were the claim, a bit of protocol elicitation would confirm or disconfirm it. But neither do Bayesian psychologists exactly say that their participants are carrying out the Bayesian computations and applying the decision rule unconsciously.

That some Bayesian model or other succeeds in accommodating learning data is no surprise: The experimenters get to screen the subjects, to postulate the priors, and to postulate post hoc the rules – however irrational they may appear by decision theoretical standards – mapping posterior probabilities to behavior. Caution and modesty are virtues in science and in life, and agnosticism may be the best theology, but vacuity is not a scientific virtue. Here is a list of further complaints:

First, for a computationally bounded agent, there is neither a theoretical nor a practical rationale for learning and judgment exclusively in accord with Bayesian principles.

Second, the Bayesian framework itself is empirically vacuous: Anything, or almost anything, could be explained by adjusting priors and likelihoods.

Third, experiments showing "probability matching" in forced choice learning tasks are prima facie in conflict with Bayesian criteria for rationality. Bayesian explanations require an ad hoc error theory that is seldom, if ever, justified empirically.

Fourth, scientific practice shows that Bayesian principles are inadequate when novel hypotheses are introduced to explain previously accepted data. Old evidence, long established, is taken to be evidence for novel hypotheses. (The most famous example: The anomalous advance of the perihelion of Mercury, established in 1858, was evidence for the novel general theory of relativity, announced in 1915.) Similar behavior should be expected in psychological experiments, although I don't know of any test of

this expectation. By Bayesian principles, established evidence, fully believed, cannot confirm or support novel hypotheses.

Fifth, these points, as well as many of Jones & Love's (J&L's) in the target article, have already been made in the philosophical literature (Eberhardt & Danks, in press; Glymour 2007), which Bayesian advocates have largely ignored. As J&L suggest in their title, some Bayesian cognitive modelers do seem to be fundamentalists.

What could be the point of all this ritualistic mathematical modeling? Ptolemy, for all of the infinite flexibility of his modeling devices, at least predicted future eclipses and positions of the planets, the sun, and the moon. I know of no Bayesian psychological prediction more precise than "We can model it." Ptolemy at least found robust regularities of the observed variables. I know of no essentially Bayesian psychological discovery of a robust behavioral regularity – that is of a prediction of a phenomenon that would not be expected on contrary grounds. Sometimes the Bayesian claim is that a Bayesian model saves the phenomena better than specific alternatives do, but that depends on the phenomena and the alternatives. Since Bayesian models came into fashion, psychological hands are sometimes waved toward Bayesian accounts when simple heuristics would account for the subjects' judgments as well or better.

J&L propose bigamous weddings, one of Bayesian psychology to the rest of conventional cognitive psychology, and one of Bayesian psychology to cognitive neuropsychology. The latter pairing may prosper, but I strongly suspect it will be a marriage of unequal partners: The neuropsychologists will be the breadwinners. In any case, the interesting recent and current work on neural encodings that can be interpreted to represent probabilities and conditional probabilities is at an astronomical distance from Bayesian psychological models.

At least Bayesian psychology at least does no harm. It does not obfuscate like psychoanalysis, Hegelian metaphysics, or postmodern literary studies. Its real cost is in the lost opportunities for good minds to advance our understanding of how the brain produces thought, emotion, and action.

Probabilistic models as theories of children's minds

doi:10.1017/S0140525X11000288

Alison Gopnik

Department of Psychology, University of California at Berkeley, Berkeley, CA 94720.

Gopnik@berkeley.edu **Alisongopnik.com**

Abstract: My research program proposes that children have representations and learning mechanisms that can be characterized as causal models of the world – coherent, structured hypotheses with consistent relationships to probabilistic patterns of evidence. We also propose that Bayesian inference is one mechanism by which children learn these models from data. These proposals are straightforward psychological hypotheses and far from "Bayesian Fundamentalism."

So who exactly are these Bayesian Fundamentalists? Since I don't know or know of any researchers who fit Jones & Love's (J&L's) description, I can't defend them. Instead, let me describe what my colleagues and students and I have actually done over the last ten years. Some 10 years ago, Clark Glymour and I proposed that children have representations and learning mechanisms that can be characterized as causal models of the world – coherent, structured hypotheses with consistent and predictable relationships to probabilistic patterns of evidence – and that Bayesian inference is one mechanism by which children may learn these models from data.

This proposal has exactly the same status as any other claim in psychology. Since the cognitive revolution, it has been a commonplace that we can explain the mind in terms of representations and rules. It is also a commonplace of the philosophy of cognitive science that these representations and rules can involve many levels of abstraction, from high-level representations such as intuitive theories to neural representations of three-dimensional structure. In vision science, for example, the poster child of successful cognitive science, explanations at many levels from ideal observer theories to neural implementations have been mutually illuminating.

Bayesian learning is just one part of a broader approach, better called the "probabilistic model" approach (see Gopnik & Schulz 2007; Griffiths et al. 2010). (For applications to cognitive development, see Gopnik et al. [2004] and special sections of *Developmental Science* [Schulz et al. 2007b] and *Cognition* [Buchsbaum et al., in press]). The central advance has not been Bayes' law itself, but the ability to formulate structured representations, such as causal graphical models, or "Bayes nets" (Pearl 2000; Spirtes et al. 2000), or hierarchical causal models, category hierarchies, and grammars, that can be easily combined with probabilistic learning, such as Bayesian inference.

I agree with J&L that formal models are useful only if they can help address empirical questions. For developmental psychologists, the great question is how children can learn as much as they do about the world around them. In the past, "theory theorists" proposed that children learn by constructing hypotheses and testing them against evidence. But if this is a deterministic process, then the "poverty of the stimulus" problem becomes acute. In contrast, if the child is a probabilistic learner, weighing the evidence to strengthen support for one hypothesis over another, we can explain how children are gradually able to revise their initial theories in favor of better ones.

Empirically, we have discovered – as a result of extensive and often challenging experiments – that young children do indeed behave like probabilistic learners, rather than simply using associationist mechanisms to match the patterns in the data or fiddling with details of innate core knowledge. The framework lets us make precise predictions about how children will behave; for example, that they will infer a common cause structure with one set of evidence, but infer a causal chain in an almost identical task with slightly different evidence (Schulz et al. 2007a). From a more specifically Bayesian perspective, it also allows us to make predictions about the probability of particular responses. The empirical likelihood that a child will make one prediction or another is closely matched to the posterior distribution of hypotheses that support those predictions (e.g., Schulz et al. 2007b).

The ultimate test of any perspective is whether it generates new and interesting empirical research. Researchers inspired by this approach have already begun to make important developmental discoveries – discoveries that don't fit either a connectionist/dynamic or nativist picture. Nine-month-olds, for example, can make causal inferences that go beyond association (Sobel & Kirkham 2006); 20-month-olds can infer a person's desire from a non-random sampling pattern (Kushnir et al. 2010); and 4-year-olds discover new abstract variables and rules from only a few data points (Lucas et al. 2010; Schulz et al. 2008), integrate new evidence and prior knowledge (Kushnir & Gopnik 2007; Schulz et al. 2007b; Sobel et al. 2004), and rationally experiment to uncover new causal structure (Schulz & Bonawitz 2007).

The framework has also enabled us to identify cases where children behave in ways that are interestingly different from ideal inference engines. For example, Kushnir and Gopnik (2005) showed that children weight their own actions more heavily than they normatively should for purposes of causal inference. Bonawitz et al. (2010) have shown that 2-year-olds have difficulty integrating purely correlational information and information about the outcomes of actions.

The precision of these models also enables us to test competing hypotheses. For example, we recently described formal

models of learning by imitation. These models gave different weights to statistical evidence and a demonstrator's intentions (Buchsbaum et al., in press). Comparing the models with the data allowed us to show that young children make prior assumptions about why people act, and that these assumptions can explain why they sometimes "overimitate," reproducing everything that the teacher does, and also predict the degree to which overimitation will occur.

It would certainly be helpful to relate these more abstract accounts to algorithmic-level instantiations. This has already been done to some extent (Danks 2003), and we are working on precisely this problem in my lab and others at the moment (Bonawitz et al. 2010). Ultimately, we would also like to relate these representations to neural implementations. It might also be interesting to relate these findings to memory or attention. But this is simply the usual scientific work of connecting different research programs, and we can't tell which connections will be illuminating beforehand. There is nothing privileged about information processing or neuroscience that means that such research is about "mechanisms" while research into representations is not. This would be like rejecting the explanation that my computer saved the file when I pushed ctrl-s because ctrl-s is the save command in Windows, since this explanation doesn't refer to the computer's working memory or wiring diagrams.

I would be happy if J&L saw this research program as a part of the Bayesian Enlightenment. If so, however, it is an enlightenment that has been in place for at least ten years.

The uncertain status of Bayesian accounts of reasoning

doi:10.1017/S0140525X1100029X

Brett K. Hayes and Ben R. Newell

School of Psychology, University of New South Wales, Sydney, NSW, 2052, Australia.

B.hayes@unsw.edu.au **Ben.newell@unsw.edu.au**

Abstract: Bayesian accounts are currently popular in the field of inductive reasoning. This commentary briefly reviews the limitations of one such account, the Rational Model (Anderson 1991b), in explaining how inferences are made about objects whose category membership is uncertain. These shortcomings are symptomatic of what Jones & Love (J&L) refer to as "fundamentalist" Bayesian approaches.

The tension between what Jones & Love (J&L) refer to as "fundamentalist" and "enlightened" Bayesian approaches to cognition is well illustrated in the field of human reasoning. Historically, Bayesian models have played an important role in trying to answer an intriguing but difficult problem in inductive reasoning: how people make predictions about objects whose category membership is uncertain. This problem can be illustrated by the example of a person hiking through a forest who hears a rustling in the scrub near her feet and wishes to predict whether they are in danger. Although the hiker knows of many animals (e.g., snakes, small mammals, birds) that could produce the noise, she cannot be certain about the actual source. The main question is to what extent does she consider the various category alternatives when making a prediction about the likelihood of danger?

One influential approach to this problem is Anderson's (1991b) Rational Model (for a more recent instantiation, see Sanborn et al. 2010a). This Bayesian model assumes that "categorization reflects the derivation of optimal estimates of the probability of unseen features of objects" (Anderson 1991b, p. 409). In the case of inductive prediction with uncertain categories, the model assumes that predictions will be based on a normative consideration of multiple candidate animal categories and the conditional probability of a target feature (e.g., whether or not the animal is likely to cause injury) within each category.

One of the appeals of the model is that it can account for cases of inductive prediction where category membership of the target object is uncertain, as well as cases where the object's category membership has been established with certainty. In contrast, most non-Bayesian accounts of induction (e.g., Osherson et al. 1990; Sloman 1993) only deal with the latter case.

Despite this promise, the Rational Model suffers from many of the shortcomings that J&L attribute to "fundamentalist" Bayesian approaches. These include the following:

1. *Lack of attention to psychological processes such as selective attention.* A considerable body of empirical evidence shows that, contrary to the Bayesian account, people do not consider all relevant category alternatives when making feature predictions (for a review, see Murphy & Ross 2007). Instead, they generally make predictions based only on the category that a target object *is most likely* to belong to. In other words, in most cases of uncertain induction people ignore the uncertainty and selectively attend to the most likely category alternative. Although this leads to non-normative predictions, it may be a useful heuristic, leading to predictions that are approximately correct while avoiding much of the complex computation involved in integrating probabilities across categories (Ross & Murphy 1996).

2. *Implausible or incorrect assumptions about representation.* Like many other Bayesian accounts, the Rational Model makes assumptions about feature and category representation that are not well-grounded in psychological theory and data. The model assumes that people treat features as conditionally independent when making inductive predictions. This means that the target object's known features (e.g., the rustling sound) are only used to identify the categories to which it might belong. These features are then ignored in the final stage of feature prediction. This assumption ignores a wealth of evidence that people are sensitive to correlations between features in natural categories and that such feature correlations influence categorization (Malt & Smith 1984; Murphy & Ross 2010; Rosch & Mervis 1975). Moreover, we have shown that people frequently base their inductive predictions on such feature correlations (Griffiths et al., in press; Newell et al. 2010; Papadopoulos et al. 2011).

3. *Failure to consider the impact of learner's goals and intent.* The extent to which inductive prediction conforms to Bayesian prescriptions often reflects the goals of the reasoner. The same individual can show more or less consideration of category alternatives when making an inductive prediction, depending on a variety of task-specific factors such as the degree of association between category alternatives and the to-be-predicted feature and the cost of ignoring less likely alternatives (Hayes & Newell 2009; Ross & Murphy 1996). When people do factor category alternatives into their predictions, it is not necessarily because they are following Bayesian prescriptions but because of changes in what J&L refer to as "mechanistic considerations" such as changes in the relative salience of the categories (Griffiths et al., in press; Hayes & Newell 2009).

4. *Disconnect between representation and decision processes.* Bayesian algorithms like those proposed by Anderson (1991b) are best interpreted as models of the decision process. As such, they often make assumptions about (rather than examine) how people represent the category structures involved in induction. In the case of uncertain induction this is a problem because the neglect of less probable category alternatives may occur prior to the final decision, when people are still encoding evidence from the candidate categories (Griffiths et al., in press; Hayes et al. 2011). It remains an open question whether Bayesian models of induction can capture such biases in evidence-sampling and representation (e.g., through the appropriate adjustment of priors).

In sum, the Rational Model has not fared well as an account of induction under category uncertainty. Many of the model's shortcomings reviewed here are common to other Bayesian models of

cognition. These shortcomings do not necessarily mean that we should abandon Bayesian approaches. But progress in fields like inductive inference is more likely to be achieved if Bayesian models are more grounded in psychological reality.

In praise of secular Bayesianism

doi:10.1017/S0140525X11000306

Evan Heit and Shanna Erickson

School of Social Sciences, Humanities, and Arts, University of California at Merced, Merced, CA 95343.

eheit@ucmerced.edu **serickson@ucmerced.edu**
http://faculty.ucmerced.edu/eheit

Abstract: It is timely to assess Bayesian models, but Bayesianism is not a religion. Bayesian modeling is typically used as a tool to explain human data. Bayesian models are sometimes equivalent to other models, but have the advantage of explicitly integrating prior hypotheses with new observations. Any lack of representational or neural assumptions may be an advantage rather than a disadvantage.

The target article by Jones & Love (J&L) is timely, given the growing importance of Bayesian models and the need to assess these models by the same standards applied to other models of cognition. J&L's comments about the pitfalls of behaviorism as well as connectionist modeling are well made and give good reason to be cautious in the development of Bayesian models. Likewise, the comments about recent trends in economics pose an interesting challenge for Bayesian modeling.

However, the religious metaphor is not apt. Most researchers cited under the heading of Bayesian Fundamentalism also have interests in psychological mechanisms and/or neural implementation and employ other modeling approaches in addition to Bayesian modeling. Hence, Bayesian modeling is less of a new religion and more of a new tool for researchers who study cognition by using both empirical and modeling methods (cf. Lewandowsky & Heit 2006). Although there is merit to J&L's point that Bayesian modeling often does not include empirical measurement of the environment, it would be a mistake to say that Bayesian modeling is mainly a normative approach or that it does not involve data collection. Most Bayesian research cited by J&L considers the adequacy of Bayesian models as descriptive models and compares their predictions to human data.

Therefore, we would say that Bayesians are just like everyone else. Bayesian models are often very similar, if not equivalent, to other kinds of models. For example, in memory research, a biologically inspired connectionist model (Norman & O'Reilly 2003) makes many of the same predictions as earlier Bayesian models of memory (McClelland & Chappell 1998; Shiffrin & Steyvers 1997). In this case, the Bayesian models fit a large number of results and could no more be dismissed on the basis of experimental data than connectionist models. Indeed, connectionist models can implement approximate Bayesian inference (McClelland 1998), and exemplar models likewise have this capability (Heit 1995; Shi et al. 2010). At a more general level, connectionist models such as mixture of experts models (Heit & Bott 2000; Heit et al. 2004; Jacobs 1997), as well as exemplar models (Heit 2001), can implement the key notion of Bayesian models – that prior knowledge is integrated with new observations to form a posterior judgment. Thus, it is difficult to dismiss Bayesian models when alternative models are fundamentally doing the same work or even making the same predictions.

J&L do not place sufficient weight on one of the main benefits of Bayesian models; namely, in explaining how prior hypotheses are put together with new observations. Cognitive activities take place in knowledge- and meaning-rich environments; for example, expectations, stereotypes, and theories affect memory, reasoning, and categorization (e.g., Bartlett 1932; Dube et al. 2010; Hayes et al. 2010; Heit 1997; Heit & Bott 2000; Murphy & Medin 1985). The role of knowledge is often obscured in experimental research using abstract stimuli. It would be a sterile model of cognition indeed that could only deal with these blank-slate situations involving meaningless materials. The approach of Bayesian models, incorporating prior hypotheses, is crucial to their success as models of many cognitive activities, and is a distinctive advantage over models that assume cognition does not incorporate prior hypotheses.

The target article criticizes Bayesian models for a lack of representational assumptions. This point is questionable; hypothesis spaces are arguably representations of the world and could be treated as psychological representations (Heit 1998; 2000; Kemp & Tenenbaum 2009). However, a lack of representational assumptions could be a virtue, because representational assumptions are usually wrong, or at best, untestable. For example, in reasoning research, there has been a long-running debate between mental model representation and mental rule representation; in categorization research, there has been a related debate between exemplar representation versus abstract representation (Rips 1990). Presumably, in each case at least one assumed form of representation is wrong, although decades of research has not resolved the matter. It has been further argued that sufficiently powerful representational systems, along with appropriate processing assumptions, are indistinguishable (Anderson 1978; Barsalou 1990). On these grounds, any lack of strong representational assumptions in Bayesian models seems a wise approach.

A similar point can be made about a lack of predictions for the neural level. Not making predictions about the brain could actually be a virtue, if the alternative is to force post hoc predictions that do not follow distinctively from the model. For example, the mental model theory of reasoning has often been tested by looking for activation in the left hemisphere of the brain. The origin of this prediction appears to be Johnson-Laird (1994), and it has been tested in many brain-imaging studies of reasoning. Although reasoning tasks are typically associated with left hemisphere activation, the results have actually been mixed, with many studies showing activation in both hemispheres, and on the whole, these brain-imaging studies do not distinguish between current theories of reasoning (see Goel [2007] for a review). An absence of neural predictions for Bayesian modeling could reflect a correct understanding of how best to use this tool. It would be better to make as few neural or representational assumptions as possible, than to make unfounded ones.

The fact that Bayesian models are rational does not imply that researchers who use them are irrational fundamentalists. Bayesian modeling is simply a new tool with advantages and disadvantages like any other kind of modeling. J&L make a vivid case for some disadvantages, but greatly understate the advantages of models that explain how prior hypotheses are put together with new observations, using a minimum number of unfounded assumptions. Still, J&L deserve credit for fostering a debate on these important models.

Relating Bayes to cognitive mechanisms

doi:10.1017/S0140525X11000318

Mitchell Herschbach and William Bechtel

Department of Philosophy, University of California, San Diego, La Jolla, CA 92093-0119.

mherschb@ucsd.edu **bill@mechanism.ucsd.edu**
http://mechanism.ucsd.edu/~mitch
http://mechanism.ucsd.edu/~bill

Abstract: We support Enlightenment Bayesianism's commitment to grounding Bayesian analysis in empirical details of psychological and neural mechanisms. Recent philosophical accounts of mechanistic science illuminate some of the challenges this approach faces. In particular, mechanistic decomposition of mechanisms into their component parts and operations gives rise to a notion of levels distinct from and more challenging to accommodate than Marr's.

We find attractive *Enlightenment Bayesianism*'s commitment to grounding Bayesian analysis in knowledge of the neural and psychological mechanisms underlying cognition. Our concern is with elucidating what the commitment to mechanism involves. While referring to a number of examples of mechanistic accounts in cognitive science and ways that Bayesians can integrate mechanistic analysis, Jones & Love (J&L) say little about the details of mechanistic explanation. In the last two decades, several philosophers of science have provided accounts of mechanistic explanation and mechanistic research as these have been practiced in biology (Bechtel & Abrahamsen 2005; Bechtel & Richardson 1993/2010; Machamer et al. 2000) and the cognitive sciences (Bechtel 2008; Craver 2007). Drawing on these can help illuminate some of the challenges of integrating mechanistic analysis into Bayesian accounts.

At the core of mechanistic science is the attempt to explain how a mechanism produces a phenomenon by decomposing it into its parts and operations and then recomposing the mechanism to show how parts and operations are organized, such that when the mechanism is situated in an appropriate environment, it generates the phenomenon. One of the best-developed examples in cognitive science is the decomposition of visual processing into a variety of brain regions, each of which is capable of processing different information from visual input. When organized together, they enable individuals to acquire information about the visible world. Decomposition can be performed iteratively by treating the parts of a given mechanism (e.g., V1) as themselves mechanisms and decomposing them into their parts and operations.

A hierarchical ordering in which parts are at a lower level than the mechanism is thus fundamental to a mechanistic perspective. This notion of levels is importantly different from that advanced by Marr (1982), to which J&L appeal, which does not make central the decomposition of a mechanism into its parts and operations. To illustrate the mechanistic conception of levels in terms of mathematical accounts, it is often valuable to provide a mathematical analysis of the phenomenon for which the mechanism is responsible. In such an account (e.g., the Haken-Kelso-Bunz [HKB] model of bimanual cordination described by Kelso 1995), the variables and parameters refer to characteristics of the mechanism as a whole and aspects of the environment with which the mechanism interacts. But to explain how such a mechanism functions one must identify the relevant parts and their operations. The functioning of these parts and operations may also require mathematical modeling (especially when the operations are nonlinear and the organization non-sequential; see Bechtel & Abrahamsen 2010). These models are at a lower level of organization and their parts and operations are characterized in a different vocabulary than that used to describe the phenomenon (as the objective is to show how the phenomenon is produced by the joint action of parts that alone cannot produce it).

We can now pose the question: At what level do Enlightenment Bayesian accounts operate? Do they, like Bayesian Fundamentalist accounts, operate at the level of the whole person, where the hypothesis space reflects people's actual beliefs? Beliefs are most naturally construed as doxastic states of the person that arise from the execution of various operations within the mind/brain. J&L's invocation of Gigerenzer's work on cognitive heuristics (e.g., Gigerenzer & Todd 1999) suggests this is a perspective they might embrace – the heuristics are inference strategies of agents and do not specify the operations that enable agents to execute the heuristics. The resulting Bayesian model may reflect but does not directly embody the results of decomposing the mind into the component operations that enable it to form beliefs.

Another possibility is that the Bayesian hypothesis space might directly incorporate details of the operations performed by components (e.g., brain regions identified in cognitive neuroscience research). Now an additional question arises – with respect to what environment is optimization evaluated? Since we are working a level down from the whole mechanism, one might think that the relevant environment is the internal environment of the local component (comprising other neural components). But this seems not to be the strategy in the research J&L cite (Beck et al. 2008; Wilder et al. 2009). Rather, optimization is still with respect to the task the agent performs. In Beck et al.'s account, a brain region (lateral intraparietal cortex: LIP) is presented as computing a Bayesian probability. This directly links the Bayesian account to parts of the mechanism, but if this approach is to be generalized, it requires that one find brain components that are computing Bayesian probabilities in each instance one applies a Bayesian analysis.

Although we find the prospect of integrating mechanistic and Bayesian approaches attractive, we are unclear how the results of mechanistic decomposition – which often leave the agent-level representations behind to explain how they are realized through a mechanism's parts and operations characterized in a different vocabulary than that which characterizes the agent's beliefs – are to be incorporated into a Bayesian account. We suspect that the most promising strategy is more indirect: Mechanistic research at lower levels of organization helps constrain the account of knowledge possessed by the agent, and Bayesian inference then applies to such agent-level representations.

A further challenge for understanding how mechanism fits into Bayesian analysis stems from the fact that Bayesian analyses are designed to elicit optimal hypotheses. As J&L note, mechanisms, especially when they evolve through descent with modification, are seldom optimal. What then is the point of integrating mechanistic accounts into normative Bayesian models? One possibility is that the normative accounts serve as discovery heuristics – mismatches between the normative model and cognitive agents' actual behavior motivate hypotheses as to features of the mechanism that account for their limitations. While this is plausible, we wonder about its advantages over investigating the nature of the mechanism more directly, by studying its current form or by examining how it evolved through a process of descent with modification. Often, understanding descent reveals how biological mechanisms have been kludged to perform a function satisfactorily but far from optimally.

What the Bayesian framework has contributed to understanding cognition: Causal learning as a case study

doi:10.1017/S0140525X1100032X

Keith J. Holyoak[a] and Hongjing Lu[a,b]

Departments of [a]Psychology and [b]Statistics, University of California, Los Angeles, CA 90095-1563.

holyoak@lifesci.ucla.edu **hongjing@ucla.edu**
http://www.reasoninglaboratory.dreamhosters.com
http://cvl.psych.ucla.edu/

Abstract: The field of causal learning and reasoning (largely overlooked in the target article) provides an illuminating case study of how the modern Bayesian framework has deepened theoretical understanding, resolved long-standing controversies, and guided development of new and more principled algorithmic models. This progress was guided in large part by the systematic formulation and empirical comparison of multiple alternative Bayesian models.

Jones & Love (J&L) raise the specter of Bayesian Fundamentalism sweeping through cognitive science, isolating it from algorithmic models and neuroscience, ushering in a Dark Ages dominated

by an unholy marriage of radical behaviorism with evolutionary "just so" stories. While we agree that a critical assessment of the Bayesian framework for cognition could be salutary, the target article suffers from a serious imbalance: long on speculation grounded in murky metaphors, short on discussion of actual applications of the Bayesian framework to modeling of cognitive processes. Our commentary aims to redress that imbalance.

The target article virtually ignores the topic of causal inference (citing only Griffiths & Tenenbaum 2009). This omission is odd, as causal inference is both a core cognitive process and one of the most prominent research areas in which modern Bayesian models have been applied. To quote a recent article by Holyoak and Cheng in *Annual Review of Psychology*, "The most important methodological advance in the past decade in psychological work on causal learning has been the introduction of Bayesian inference to causal inference. This began with the work of Griffiths & Tenenbaum (2005, 2009; Tenenbaum & Griffiths 2001; see also Waldmann & Martignon 1998)" (Holyoak & Cheng 2011, pp. 142–43). Here we recap how and why the Bayesian framework has had its impact.

Earlier, Pearl's (1988) concept of "causal Bayes nets" had inspired the hypothesis that people learn causal models (Waldmann & Holyoak 1992), and it had been argued that causal induction is fundamentally rational (the power PC [probabilistic contrast] theory of Cheng 1997). However, for about a quarter century, the view that people infer cause-effect relations from non-causal contingency data in a fundamentally rational fashion was pitted against a host of alternatives based either on heuristics and biases (e.g., Schustack & Sternberg 1981) or on associative learning models, most notably Rescorla and Wagner's (1972) learning rule (e.g., Shanks & Dickinson 1987). A decisive resolution of this debate proved to be elusive in part because none of the competing models provided a principled account of how *uncertainty* influences human causal judgments (Cheng & Holyoak 1995).

J&L assert that, "Taken as a psychological theory, the Bayesian framework does not have much to say" (sect. 2.2, para. 3). In fact, the Bayesian framework says that the assessment of causal strength should not be based simply on a point estimate, as had previously been assumed, but on a probability distribution that explicitly quantifies the uncertainty associated with the estimate. It also says that causal judgments should depend jointly on prior knowledge and the likelihoods of the observed data. Griffiths and Tenenbaum (2005) made the critical contribution of showing that different likelihood functions are derived from the different assumptions about cause-effect representations postulated by the power PC theory versus associative learning theory. Both theories can be formulated within a common Bayesian framework, with each being granted exactly the same basis for representing uncertainty about causal strength. Hence, a comparison of these two Bayesian models can help identify the fundamental representations underlying human causal inference.

A persistent complaint that J&L direct at Bayesian modeling is that, "Comparing multiple Bayesian models of the same task is rare" (target article, Abstract); "[i]t is extremely rare to find a comparison among alternative Bayesian models of the same task to determine which is most consistent with empirical data" (sect. 1, para. 6). One of J&L's concluding admonishments is that, "there are generally many Bayesian models of any task. . . . Comparison among alternative models would potentially reveal a great deal" (sect. 7, para. 2). But as the work of Griffiths and Tenenbaum (2005) exemplifies, a basis for comparison of multiple models is exactly what the Bayesian framework provided to the field of causal learning.

Lu et al. (2008b) carried the project a step further, implementing and testing a 2×2 design of Bayesian models of learning causal strength: the two likelihood functions crossed with two priors (uninformative vs. a preference for sparse and strong causes). When compared to human data, model comparisons established that human causal learning is better explained by the assumptions underlying the power PC theory, rather than by those underlying associative models. The sparse-and-strong prior accounted for subtle interactions involving generative and preventive causes that could not be explained by uninformative priors.

J&L acknowledge that, "An important argument in favor of rational over mechanistic modeling is that the proliferation of mechanistic modeling approaches over the past several decades has led to a state of disorganization" (sect. 4.1, para. 2). Perhaps no field better exemplified this state of affairs than causal learning, which had produced roughly 40 algorithmic models by a recent count (Hattori & Oaksford 2007). Almost all of these are non-normative, defined (following Perales & Shanks 2007) as not derived from a well-specified computational analysis of the goals of causal learning. Lu et al. (2008b) compared their Bayesian models to those which Perales and Shanks had tested in a large meta-analysis. The Bayesian extensions of the power PC theory (with zero or one parameter) accounted for up to 92% of the variance, performing at least as well as the most successful non-normative model (with four free parameters), and much better than the Rescorla-Wagner model (see also Griffiths & Tenenbaum 2009).

New Bayesian models of causal learning have thus built upon and significantly extended previous proposals (e.g., the power PC theory), and have in turn been extended to completely new areas. For example, the Bayesian power PC theory has been applied to analogical inferences based on a single example (Holyoak et al. 2010). Rather than blindly applying some single privileged Bayesian theory, alternative models have been systematically formulated and compared to human data. Rather than preempting algorithmic models, the advances in Bayesian modeling have inspired new algorithmic models of sequential causal learning, addressing phenomena related to learning curves and trial order (Daw et al. 2007; Kruschke 2006; Lu et al. 2008a). Efforts are under way to link computation-level theory with algorithmic and neuroscientific models. In short, rather than monolithic Bayesian Fundamentalism, normal science holds sway. Perhaps J&L will happily (if belatedly) acknowledge the past decade of work on causal learning as a shining example of "Bayesian Enlightenment."

Come down from the clouds: Grounding Bayesian insights in developmental and behavioral processes

doi:10.1017/S0140525X11000331

Gavin W. Jenkins, Larissa K. Samuelson, and John P. Spencer

Department of Psychology and Delta Center, University of Iowa, Iowa City, IA 52242-1407.

gavin-jenkins@uiowa.edu larissa-samuelson@uiowa.edu john-spencer@uiowa.edu

http://www.psychology.uiowa.edu/people/gavin_jenkins

http://www.psychology.uiowa.edu/people/larissa_samuelson

http://www.psychology.uiowa.edu/people/john_spencer

Abstract: According to Jones & Love (J&L), Bayesian theories are too often isolated from other theories and behavioral processes. Here, we highlight examples of two types of isolation from the field of word learning. Specifically, Bayesian theories ignore emergence, critical to development theory, and have not probed the behavioral details of several key phenomena, such as the "suspicious coincidence" effect.

A central failing of the "Bayesian Fundamentalist" perspective, as described by Jones & Love (J&L), is its isolation from other theoretical accounts and the rich tradition of empirical work in psychology. Bayesian fundamentalists examine phenomena exclusively at the computational level. This limits contact with other theoretical advances, diminishing the relevance and impact of Bayesian models. This also limits Bayesians' concern

with the processes that underlie human performance. We expand upon the consequences of these senses of isolation within the context of word learning research.

One of the most striking shortcomings of Bayesian word learning approaches is a lack of integration with developmental theory. J&L put this quite starkly: In the Bayesian perspective, "Nothing develops" (see sect. 5.4). We agree, but believe that this would be more aptly put as, "Nothing *emerges*." Why? Emergence – the coalescing of useful complexity out of simple inputs – is a key element of *any* developmental theory and a key concept in modern theories of word learning (see Smith 2000). Without emergence, existing knowledge can only be shuffled around or re-weighted; no qualitatively new psychological progress can be made (see Smith & Thelen 2003; Spencer et al. 2009).

Critically, Bayesian models leave no room for emergence in their hypothesis space, the priors, or the Bayes' rule itself. Recent approaches using hierarchical Bayesian models (HBMs) show an impressive ability to discover structure in data (e.g., Tenenbaum et al. 2011), giving a surface feel of emergence. However, because this ability rests on the modeler building in multiple hypothesis spaces and priors in advance, it is not deeply emergent. These models do not build something new that was not there before (see Spencer & Perone 2008).

Bayesian disregard for emergence and development is clearly seen in the Kemp et al. (2007) model of the shape bias discussed by J&L. This model does not add any quantitative or predictive value over Smith and colleagues' earlier alternatives (Smith et al. 2002). Indeed, by modeling children's behavior with static hypotheses about word meanings, they failed to capture the Smith group's crucial arguments about the *emergence* of this word learning bias. In effect, Kemp et al. presented a model of the phenomenon but without the development. This is not forward theoretical progress.

A second shortcoming of the Bayesian perspective is a failure to probe the inner workings of empirical phenomena in greater than a computational level of detail. Our recent work in the area of word learning does exactly this and reveals severe limitations of Bayesian interpretations.

In one set of experiments, we have demonstrated that a well-known Bayesian phenomenon – the suspicious coincidence (Xu & Tenenbaum 2007b) – falls apart when several key empirical details are manipulated. The "suspicious coincidence" refers to adults' and children's more narrow interpretation of a word when taught using multiple, identical exemplars than when taught with a single exemplar. Spencer et al. (2011) showed that when the multiple exemplars are presented *sequentially* rather than simultaneously – as is the case in many real-world learning situations – adults no longer show a suspicious coincidence effect. This result has no specific contact to the concepts used in the Bayesian model, yet it intuitively maps onto concepts with a rich history in psychology: Simultaneous presentations encourage multiple comparisons over objects, leading to an emphasis on specific featural details, while sequential presentations afford a more global interpretation of similarity (see, e.g., Samuelson et al. 2009). Clearly, a theoretical account of the suspicious coincidence must address such facts.

In a separate experiment, we replicated the suspicious coincidence effect with $3\frac{1}{2}$- to 5-year-old children when exemplars were labeled three times. When, however, we increased the number of labeling events, children no longer showed a suspicious coincidence effect (Jenkins et al., in press). Once again, this manipulation falls outside the scope of the concepts used in the Bayesian model, but it is a factor that most theories of word learning and categorization would naturally consider. And, critically, children's performance is robustly modulated by such details.

Xu and Tenenbaum (2007b) also neglected to probe the details of the knowledge children bring to the word learning task (in Bayesian terms, their hypothesis spaces and priors). Instead of measuring knowledge directly, Xu and Tenenbaum substituted adult data from a separate adult experiment. By contrast, we gathered data from children by using a table-top similarity ratings task (Perry et al., in preparation; see also, Goldstone 1994). Results showed dramatic, qualitative differences in the structure of children's and adults' category knowledge. Moreover, children with above-median prior knowledge of the object categories, as measured by parental report, failed to show a suspicious coincidence effect, whereas below-median children showed a strong suspicious coincidence effect. This is the opposite of what Bayesian models predict.

One empirical detail of the suspicious coincidence that Bayesians *have* probed is its dependence on whether exemplars are chosen by a knowledgeable teacher. Bayesians claim a sample is representative to word learners if it is chosen by a knowledgeable teacher but potentially biased, and therefore less informative, otherwise (Xu & Tenenbaum 2007a). We attempted – and failed – to replicate the behavioral evidence supporting this dependence. Xu and Tenenbaum found a striking difference between teacher-informed adults ("teacher-driven" in Figure 1A) and adults who partially

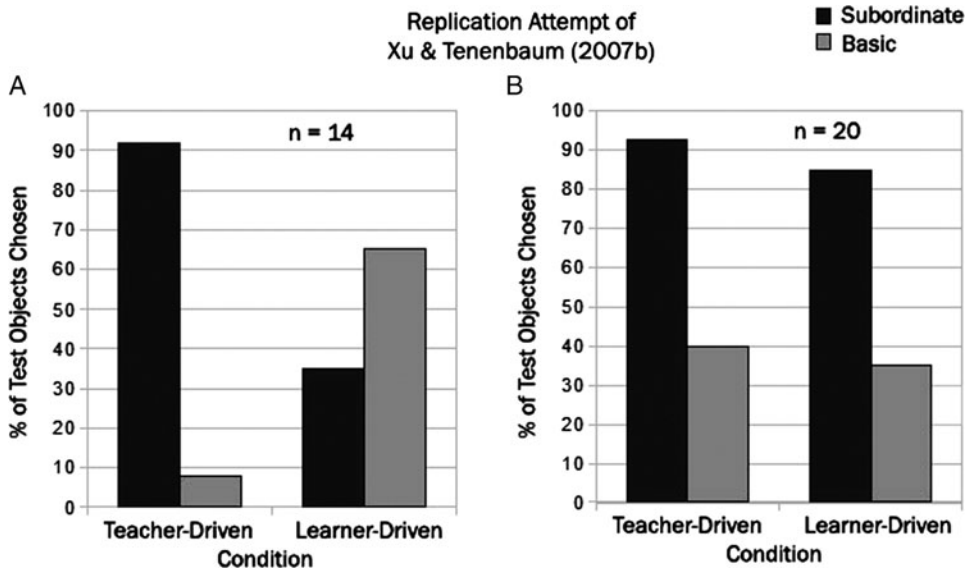

Figure 1 (Jenkins et al.). Replication attempt by Xu and Tenenbaum (2007a). A: Xu and Tenenbaum's results. B: Our exact replication attempt.

chose their own exemplars ("learner-driven" in Figure 1A). Our adult subjects showed no such effect (Figure 1B). It is possible that the Xu and Tenenbaum data were influenced by the low number of participants ($N = 14$ in Figure 1A; $N = 20$ in Figure 1B).

The foregoing examples demonstrate a general fragility of one prominent line of Bayesian word learning research. We believe this fragility to be both a characteristic and direct consequence of the Bayesian tendency to isolate theory from the details of mechanism and process.

In summary, we concur with J&L that there are serious limitations in the Bayesian perspective. Greater integration with other theoretical concepts in psychology, particularly in developmental science, and a grounded link to the details of human performance are needed to justify the continued excitement surrounding this approach.

In praise of Ecumenical Bayes

doi:10.1017/S0140525X11000343

Michael D. Lee

Department of Cognitive Sciences, University of California, Irvine, CA 92697-5100.

mdlee@uci.edu **www.socsci.uci.edu/~mdlee**

Abstract: Jones & Love (J&L) should have given more attention to *Agnostic* uses of Bayesian methods for the statistical analysis of models and data. Reliance on the frequentist analysis of Bayesian models has retarded their development and prevented their full evaluation. The *Ecumenical* integration of Bayesian statistics to analyze Bayesian models offers a better way to test their inferential and predictive capabilities.

In the target article, Jones & Love (J&L) argue that using Bayesian statistics as a theoretical metaphor for the mind is useful but, like all metaphors, limited. I think that is a sensible position. Bayesian methods afford a complete and coherent solution to the problem of drawing inferences over structured models from sparse and noisy data. That seems like a central challenge faced by the mind, and so it is not surprising the metaphor has led to insightful models of human cognition. But it will never be the only useful metaphor.

I certainly agree with the target article that using Bayesian methods as a statistical framework – that is, as a means to connect models of cognition with data – is the right thing to do (Lee 2008; 2011). This "Agnostic" approach is not discussed much in the target article, which focuses on "Fundamentalist" uses of Bayes as a theoretical metaphor. The argument is that Fundamentalist approaches can lead to Enlightenment through reintegrating processes and representations into Bayesian cognitive models.

What I think is missing from this analysis is the central role of Agnostic Bayes on the path to enlightenment. I think Bayesian models of cognition, including potentially more process and representation rich ones, need to use Bayesian methods of analysis if they are to realize their full potential. The target article does not say very much about the Bayesian analysis of Bayesian models. It does sound favorably disposed when discussing the need to evaluate the complexity of cognitive models, which is a natural property of Bayesian model selection. But the argument for Bayesian statistical analysis is never made as forcefully as it should be.

Using Bayesian statistics to analyze Bayesian models might be called "Ecumenical" Bayes, since it integrates the two uses of Bayesian methods in studying human cognition. As best I know, there are very few examples of this integrative approach (e.g., Huszar et al. 2010; Lee & Sarnecka 2010; in press). But I think it is theoretically and practically important.

It has always struck me (e.g., Lee 2010; 2011), and others (e.g., Kruschke 2010) that there is a sharp irony in many papers presenting Bayesian models of cognition. Often the rationality of Bayesian inference is emphasized when discussing how people might make optimal use of available information. But, when the authors want to test their model against data, and hence face the same inferential problem, the solution is suddenly different. Now they revert to irrational statistical methods, like frequentist estimation and null hypothesis tests, to draw conclusions about their model.

This complaint is not just statistical nit-picking. Non-Bayesian analysis has retarded the development of Bayesian models of cognition, by limiting the sorts of Bayesian models that can be considered, and the depth to which they have been understood and used.

I think it is possible to illustrate this claim by using Lee and Sarnecka's (2010; in press) work on modeling children's development of number concepts. The target article is dismissive of this work, saying it is done "at the expense of identifying general mechanisms and architectural characteristics . . . that are applicable across a number of tasks" (sect. 5, para. 5). This is a strange critique, since the main point of Lee and Sarnecka (2010; in press) is to argue for specific types of constrained representations, in the form of knower-levels, and show how those representations explain observed behavior on multiple tasks. But, that confusion aside, I want to use the work as an example of the benefits of using Bayesian statistics to analyze Bayesian models.

A key part of Lee and Sarnecka's (2010; in press) model is a base rate for behavioral responses, which corresponds to the child's prior. It is a probability distribution over the numbers 0 to 15, and is difficult to handle with frequentist estimation. If the model were being analyzed in the manner usually adopted to evaluate Bayesian cognitive models, my guess is the following would have been done. The base-rate prior would have been hand-tuned to a reasonable set of values, and the model would have been used to generate behavior. These "predictions" would then have been compared to experimental data, perhaps accompanied by a simple summary statistic measuring the agreement, and compared to "straw" models that, for example, did not have base-rate priors. The conclusion would have been drawn that the Bayesian machinery had the right properties to explain key patterns in data showing how children acquire number concepts.

I find this sort of approach unsatisfying. One of the main reasons for developing sophisticated models of cognition, like Bayesian models, is to be able to draw inferences from data, and make predictions and generalization to future and different situations. A high-level demonstration that a model is, in principle, capable of generating the right sorts of behavioral patterns falls a long way short of best-practice model-based empirical science.

What Lee and Sarnecka (2010; in press) were able to do, using Bayesian instead of frequentist statistical methods, was *infer* the base-rate prior from behavioral data, together with all of the other psychological variables in the model. This is a much more mature application of Bayesian modeling, because it makes full contact with the data. It allows the descriptive and predictive adequacy of the model to be assessed (e.g., through standard posterior predictive analysis). It allows the Bayesian model to be used to learn about parameters from data, since it gives the full joint posterior distribution over the (complicated) parameter space. And it enables the same representational model to be applied to data from multiple developmental tasks simultaneously, within a hierarchical framework.

I think these sorts of Bayesian statistical capabilities have the potential to address many of the concerns raised by the target article about the currently demonstrated success of Bayesian

models of cognition. Bayesian statistical methods are important, useful, and should play a central role in analyzing all models of cognition, including Bayesian ones. The target article views this as a side issue, but I think it is a fundamental element of the path to enlightenment.

Cognitive systems optimize energy rather than information

doi:10.1017/S0140525X11000355

Arthur B. Markman and A. Ross Otto

Department of Psychology, University of Texas, Austin, TX 78712.
markman@psy.utexas.edu **rotto@mail.utexas.edu**
http://homepage.psy.utexas.edu/homepage/Faculty/Markman/PSY394/kreps11.html

Abstract: Cognitive models focus on information and the computational manipulation of information. Rational models optimize the function that relates the input of a process to the output. In contrast, efficient algorithms minimize the computational cost of processing in terms of time. Minimizing time is a better criterion for normative models, because it reflects the energy costs of a physical system.

Two parallel developments in the 1940s set the stage both for the cognitive revolution of the 1950s and for the discussion presented in the target article. The development of information theory explored ways to characterize the information content of a message and ways to consider how to best pass messages (Shannon 1949). At the same time, the architecture for digital computing led to advances in discrete mathematics that facilitated the analysis of the efficiency of algorithms (Turing 1950).

One consequence of the cognitive revolution was that that it became common to characterize the mind as a computational device. Thus, researchers began to formulate theories of mental processes in computational terms. As Marr (1982) points out, a process can be defined at either a computational level or an algorithmic level of description. At the computational level, the process is defined by a mapping between information available at the start and end of the process. For example, Anderson (1990) advocates a Bayesian, "rational-level" analysis of the information relationship between inputs and outputs of a system. At the algorithmic level, a process is specified in terms of a set of steps that implements this computational-level description. Any given algorithm can be analyzed for its efficiency in time. The efficiency of a cognitive process can be established at either the computational level of description or at the algorithmic level. The Bayesian approaches described in the target article are focused on defining the optimality of a cognitive process at the computational level (Anderson 1990; Tenenbaum & Griffiths 2001). Anderson (1990) does point out that computational costs can also play a role in determining a rational model, but, in practice, these considerations did not have a significant influence on the structure of his rational models.

The danger in casting optimality purely at the computational level is that human cognition is implemented by a physical system. Indeed, it has been proposed that any characterization of the optimality of actions or beliefs should take into account the resource-limited nature of the human cognitive apparatus (Cherniak 1986; Stanovich & West 1998). As the target article points out, the brain consumes a significant amount of energy. Thus, energy minimization is likely to be an important constraint on cognitive processing.

The idea that energy-minimization is an important constraint on cognitive processing is implicit in the focus on efficient computational procedures. We do not suppose that the metabolic cost of cognition is completely invariant of the type of thinking that people are engaged in, but marginal changes in metabolic rates attributed to different types of cognition pale in comparison to the metabolic cost of simply keeping the brain running. Thus, the time taken by a process is a good proxy for energy conservation. On this view, for example, habits minimize energy, because they allow a complex behavior to be carried out quickly (e.g., Logan 1988; Schneider & Shiffrin 1977).

Of course, effort-minimization is not the only constraint on cognitive processing. It is crucial that a process be carried out to a degree sufficient to solve the problem faced by the individual. This view was central to Simon's (1957b) concept of *satisficing*. This view suggested that cognitive processes aim to expend the minimal amount of effort required to solve a problem. On this view, the costs of additional effort outweigh the gains in decision accuracy. This idea was elaborated in the effort accuracy framework developed by Payne et al. (1993). Their work examined the variety of strategies that people utilize in order to balance decision accuracy with effort – the cognitive costs of gathering and integrating information about choice attributes – in decision-making. Payne et al. point out that these strategies differ both in the effort required to carry them out as well as in their likelihood of returning an accurate response. People negotiate the trade-off between effort and accuracy by selecting decision strategies that minimize the effort required to yield an acceptable outcome from a choice.

A key shortcoming, then, of the Bayesian Fundamentalist approach is that it optimizes the wrong thing. The ideal observer or actor defined purely in terms of information is quite useful, but primarily as a point of comparison against human cognitive or sensory abilities rather than as a statement of what is optimal as a cognitive process (e.g., Geisler 1989). A definition of optimal behavior needs to take energy minimization into account. Thus, the key limitation of Bayesian Fundamentalism is that it focuses selectively on optimality of information processing rather than on the combination of information and time.

Enlightenment grows from fundamentals

doi:10.1017/S0140525X11000367

Daniel Joseph Navarro and Amy Francesca Perfors

School of Psychology, University of Adelaide, Adelaide, SA 5005, Australia.
daniel.navarro@adelaide.edu.au **amy.perfors@adelaide.edu.au**
http://www.psychology.adelaide.edu.au/personalpages/staff/danielnavarro/
http://www.psychology.adelaide.edu.au/personalpages/staff/amyperfors/

Abstract: Jones & Love (J&L) contend that the Bayesian approach should integrate process constraints with abstract computational analysis. We agree, but argue that the fundamentalist/enlightened dichotomy is a false one: Enlightened research is deeply intertwined with – and to a large extent is impossible without – the basic, fundamental work upon which it is based.

Should Bayesian researchers focus on "enlightened" modelling that seriously considers the interplay between rational and mechanistic accounts of cognition, rather than a "fundamentalist" approach that restricts itself to rational accounts only? Like many scientists, we see great promise in the "enlightened" research program. We argue, however, that enlightened Bayesianism is deeply reliant on research into Bayesian fundamentals, and the fundamentals cannot be abandoned without greatly affecting more enlightened work. Without solid fundamental work to extend, enlightened research will be far more difficult.

To illustrate this, consider the paper by Sanborn et al. (2010a), which Jones & Love (J&L) consider to be "enlightened" as it seeks to adapt an ideal Bayesian model to incorporate insights about psychological process. To achieve this, however, it relies heavily upon work that itself would not have counted as

"enlightened." The comparison between Gibbs sampling and particle filtering as rival process models grew from "unenlightened" research that used these algorithms purely as methodological tools. As such, without this "fundamentalist" work the enlightened paper simply would not have been written.

Enlightened research can depend on fundamentals in other ways. Rather than adapt an existing Bayesian model to incorporate process constraints, Navarro and Perfors (2011) used both Bayesian fundamentals (an abstract hypothesis space) and process fundamentals (capacity limitations on working memory) as the foundations of an analysis of human hypothesis testing. Identifying a conditionally optimal learning strategy, given the process constraint, turned out to reproduce the "positive test strategy" that people typically employ (Wason 1960), but only under certain assumptions about what kinds of hypotheses are allowed to form the abstract hypothesis space. This analysis, which extended existing work (Klayman & Ha 1987; Oaksford & Chater 1994) and led us to new insights about what kinds of hypotheses human learners "should" entertain, could not have been done without "fundamentalist" research into *both* the statistical and the mechanistic basis of human learning.

Not only do "enlightened" papers *depend* on fundamental ones, we suggest that they are a natural *outgrowth* of those papers. Consider the early work on Bayesian concept learning, which contained a tension between the "weak sampling" assumption of Shepard (1987) and the "strong sampling" assumption of Tenenbaum and Griffiths (2001). When strong sampling was introduced, it would presumably have counted as "fundamentalism," since the 2001 paper contains very little by way of empirical data or consideration of the sampling structure of natural environments. Nevertheless, it served as a foundation for later papers that discussed exactly those issues. For instance, Xu and Tenenbaum (2007a) looked at how human learning is shaped by explicit changes to the sampling model. This in turn led Navarro et al. (in press) to propose a more general class of sampling models, and to pit them all against one another in an empirical test. (It turned out that there are quite strong individual differences in what people use as their "default" sampling assumption.) The change over time is instructive: What we observe is a gradual shift from simpler "fundamentalist" papers that develop the theory in a reduced form, towards a richer framework that begins to capture the subtleties of the psychology in play.

Even J&L's own chosen examples show the same pattern. Consider the Kemp et al. (2007) article, which J&L cite as a prime example of "fundamentalist" Bayesianism, since it introduces no new data and covers similar ground to previous connectionist models (Colunga & Smith 2005). Viewing the paper in isolation, we might agree that the value added is minor. But the framework it introduced has been a valuable tool for subsequent research. An extension of the model has been used to investigate how adults learn to perform abstract "second order" generalizations (Perfors & Tenenbaum 2009) and to address long-debated issues in verb learning (Perfors et al. 2010). A related model has even been used to investigate process-level constraints; Perfors (in press) uses it to investigate whether or not memory limitations can produce a "less is more" effect in language acquisition. It is from the basic, fundamental research performed by Kemp et al. (2007) that these richer, more enlightened projects have grown.

Viewed more broadly, the principle of "enlightenment growing from fundamentals" is applicable beyond Bayesian modelling; our last example is therefore an inversion. We suggest that J&L understate the importance of computational considerations in good process modelling. For instance, one of their key examples comes from Sakamoto et al. (2008), who consider mechanistic models of category learning. That paper might be characterized as a "fundamentalist" work in process modelling, insofar as it gives no consideration to the computational level issues that pertain to their choice of learning problem. As consequence of this "process fundamentalism," the "rational" model that paper employs is not actually a rational model. It is highly mis-specified for the problem of learning time-inhomogeneous categories. In recent work (Navarro & Perfors 2009), we discuss this concern and introduce extensions to the experimental framework aimed at highlighting the computational considerations involved; at present, we are working on model development to build on this. However, the goal in our work is *not* to deny the importance of process, but to learn which aspects of human behaviour are attributable to computational level issues and which aspects reflect process limitations. In this case, that goal is met by building on fundamental work on the process level (i.e., Sakamoto et al.'s 2008 paper) and adding computational considerations. In general, attaining the goal of "enlightened" research is possible only if fundamentals on both levels are taken seriously – if researchers deny neither psychological mechanism *nor* ideal computation.

Like J&L, we believe that it is the *interaction* between the twin considerations of computation and process that leads us to learn about the mind. However, this should not lead us to abandon work that focuses on only one of these two components. Enlightened research is constructed from the building blocks that fundamental work provides.

The illusion of mechanism: Mechanistic fundamentalism or enlightenment?

doi:10.1017/S0140525X11000379

Dennis Norris

MRC Cognition and Brain Sciences Unit, Cambridge CB2 7EF, United Kingdom.

dennis.norris@mrc-cbu.cam.ac.uk

http://www.mrc-cbu.cam.ac.uk/people/dennis.norris

Abstract: Rather than worrying about Bayesian Fundamentalists, I suggest that our real concern should be with Mechanistic Fundamentalists; that is, those who believe that concrete, but frequently untestable mechanisms, should be at the heart of all cognitive theories.

Jones & Love (J&L) suggest that we should reject Bayesian Fundamentalism in favour of Bayesian Enlightenment, thus combining Bayesian analysis with mechanistic-level models. This raises two questions: Who are these Bayesian Fundamentalists and what is a mechanistic-level model?

First, let us go in search of Bayesian Fundamentalists. As I read the target article, I began to wonder how it could be that I'd never encountered a Bayesian Fundamentalist. If these ideas are so pervasive, then surely J&L could quote at least one author who has made a clear statement of the Bayesian Fundamentalist programme? From the first line of the abstract it appears that the main proponent of Bayesian Fundamentalism must be Anderson (1990) with his Rational Analysis framework, and his suggestion that behaviour can often be explained by assuming that it is optimally adapted to its purpose and the environment. In criticising rational analysis, J&L argue that "Rather than the globally optimal design winning out, often a locally optimal solution . . . prevails. . . . Such non-behavioral factors are enormously important to the optimization process, but are not reflected in rational analyses, as these factors are tied to a notion of mechanism, which is absent in rational analyses" (sect. 5.3, paras. 3 and 5).

A similar concern about the limitations of rational analysis can be found in the following quotation: "My guess is that short-term memory limitations do not have a rational explanation. . . . [T]hey reflect the human trapped on some local optimum of evolution" (Anderson 1990, pp. 91–92). These cautionary words on the dangers of relying entirely on rational explanations were written by the arch-Fundamentalist himself. Is there really a difference, then, between these two positions?

Let's now move on to the second question: What is a mechanistic-level model? If we need to develop mechanistic models, then we need to know what such entities might look like. Nowhere do J&L define what they mean by a mechanistic-level theory. Perhaps we can get some clues from their other writings. Sakamoto et al. (2008) describe a "mechanistic model that principally differs from the aforementioned rational models in that the mechanistic model does not have perfect memory for the training items" (p. 1059). But here the mechanism is simply a specification of the computations that the model performs. That is, although the model is not as abstract as Marr's (1982) computational level, it is not as concrete as his algorithmic level, and certainly says nothing about implementation.

One might call this a *process model*, or a *functional-level* explanation. It specifies the functions, computations, and processes in a way that allows the model to be implemented as a computer program and to simulate behavioural data. The program performing the simulations must compute mathematical functions such as square roots, but presumably the exact algorithm or implementation used to compute a square root is not part of the theory. If this kind of functional explanation is indeed what J&L mean by a mechanistic theory, then I am wholeheartedly in favour of their approach. But I am not entirely sure that this is exactly what they have in mind. Elsewhere they talk about mechanistic issues "of representation, timing, capacity, anatomy, and pathology" (sect. 4.1, para. 3). If this is simply to echo Marr in wishing to bring together multiple levels of description and explanation, then few would disagree. However, I worry that J&L may be encouraging *Mechanistic Fundamentalism*: the belief that a good cognitive theory must do more than just describe processes and computations, and must also specify concrete mechanisms in terms of mechanical components such as *nodes, activations, weights*, and *buffers*. This view easily leads to the *illusion of mechanism*, whereby the mechanisms are mistaken for explanations.

Let's illustrate this by considering interactive activation networks, which are still at the core of many contemporary models. In these networks the activation of each node increases as a result of weighted input, and decreases as a result of inhibition from competing nodes. Activations roughly reflect the evidence for each node or hypothesis as a proportion of the evidence for all hypotheses. Although it is hard to specify exactly what computational function such networks perform, the general principle seems very much like Bayes' theorem. However, for many psychologists the network model is to be preferred over a Bayesian explanation because the former seems to say something about mechanism. But this is the illusion of mechanism. Unless the precise implementation of the network is intended as a theoretical claim about how processes are implemented in the brain, the mechanism itself makes no contribution to the explanation. If it happened to be the case that the data could be fit by any "mechanism" that could compute Bayes' theorem, then the explanation would be that the system behaves in an approximately optimal manner.

This immediately raises the problem of model equivalence. Unless candidate mechanisms produce testably different behaviours, the implementational details are not part of the explanation. To quote yet again from Anderson (1990), "If two theorists propose two sets of mechanisms in two architectures that compute the same function, then they are proposing the same theory" (p. 26). One might protest that at least when studying the mind and the brain, there will always be some neurobiological data that could definitively distinguish between alternative mechanisms. Even ignoring the fact that such a view would imply that there is no distinctly psychological level of explanation, in practice such optimism is misplaced. Even the most productive cognitive theories rarely make any definitive commitment to implementational details. Again, this is apparent in connectionist models based on artificial neurons whose properties bear little resemblance to real neurons. But connectionist modellers are fully aware of this deliberate limitation, and it is hard to see that any of the insights from connectionist modelling are undermined by this simplification.

In conclusion, then, I suggest that most Bayesians are already enlightened; it is the Mechanistic Fundamentalists we should worry about.

Reverse engineering the structure of cognitive mechanisms

doi:10.1017/S0140525X11000380

David Pietraszewski and Annie E. Wertz

Yale University, Department of Psychology, Yale University, New Haven, CT 06520-8205.

david.pietraszewski@yale.edu **annie.wertz@yale.edu**

Abstract: Describing a cognitive system at a mechanistic level requires an engineering task analysis. This involves identifying the task and developing models of possible solutions. Evolutionary psychology and Bayesian modeling make complimentary contributions: Evolutionary psychology suggests the types of tasks that human brains were designed to solve, while Bayesian modeling provides a rigorous description of possible computational solutions to such problems.

Because of their mathematical formalism, Bayesian models of cognition have the potential to infuse greater rigor into psychological models of how the mind works. Any theoretical framework committed to specifying (1) the class of cues that a mechanism is a sensitive to, (2) the operations it performs in response to those cues, and (3) the resultant outputs, is to be heartily welcomed into the theoretical toolbox of psychology.

Jones & Love (J&L) argue that, to be successful, Bayesian modelers should increase their focus on a mechanistic level of analysis, and use the examples of behaviorism and evolutionary psychology to warn them against the pitfalls of theoretical approaches that ignore psychological mechanisms and instead move directly from behavior to the environment. In the case of evolutionary psychology, this critique is simply mistaken. In fact, the field was founded specifically in response to previous evolutionary approaches, such as ethology, that ignored this middle level of analysis (e.g., Cosmides & Tooby 1987). The goal of evolutionary psychology is the same as any branch of cognitive science: to describe the information-processing structure of psychological mechanisms. What is distinct about evolutionary psychology is that principles of natural selection are used to predict the structure of cognitive mechanisms. These models generate testable predictions that can be adjudicated by empirical data.

The history of psychology suggests that well-specified task analyses (Marr 1982) are the most tractable way of reverse engineering the structure of cognitive mechanisms. As J&L discuss, the challenge for any psychologist is to (1) identify the task being solved, and (2) develop models of possible solutions. Through this lens, evolutionary psychology and Bayesian modeling make complimentary contributions. Evolutionary psychology, properly applied, is a deductive framework for generating predictions about the types of tasks cognitive mechanisms were designed to solve. This constrains the possibility space for the structure of a cognitive mechanism – what class of cues mechanisms are likely to use and what their resultant output and criterion for success should be. It rests on the premise that natural selection builds deterministic cognitive mechanisms that take as inputs aspects of the world that were invariant over phylogenetic time and generate outputs that would have led to the intergenerational differential reproduction of such systems. It is therefore a way to deductively generate hypotheses about the existence of previously unknown cognitive mechanisms. What evolutionary psychology is not – even in principle – is a description of any

particular engineering solution. In contrast, Bayesian modeling is a description of an engineering solution: Cognitive mechanisms whose function requires holding and updating probabilities will – constraints aside – behave according to Bayes' theorem. What Bayesian modeling is not – even in principle – is a way to generate hypotheses or predictions about the range of cues cognitive systems use and their criteria for success.

Natural selection builds cognitive mechanisms around phylogenetic invariances. Organisms' cognitive mechanisms reflect the dynamics of multiple generations of individuals interacting with recurrent features of the natural environment (i.e., phylogenetic dynamics not visible within an individual lifetime). For example, after copulation with a female, a male house mouse will commit infanticide on any pups born for the duration of the typical mouse gestational period, after which point they will rear any pups born; manipulations demonstrated that males achieve this by tracking the number of light/dark cycles (Perrigo et al. 1991; 1992). The cognitive mechanisms in the male mouse that mediate this relationship between light/dark cycles and killing versus caring behaviors are a result of the dynamics of differential reproductive success over multiple generations. Invariant relationships in the world – the duration of light/dark cycles in a natural terrestrial environment, that copulation leads to offspring, the duration of gestation, and so forth – are "seen" by natural selection, which in turn engineers biological mechanisms that instantiate input/output relationships. In this example, not only are the input cues based around intergenerational invariances, but the generated outputs are those which would lead to differential reproductive success within the context of those intergenerational invariances (i.e., mechanisms that discriminately kill or rear pups as a function of actuarial relatedness will do differentially better over multiple generations than mechanisms that do not).

As this example demonstrates, differential reproductive success (i.e., natural selection) operating over phylogenetic invariances determines input/output relationships in cognitive systems (see Tooby et al. [2008] for examples of using the deductive logic of phylogenetic invariances to predict and test novel cognitive mechanisms in humans). Of course, once a task is identified and a relevant mechanism proposed, the computational structure of that mechanism must still be described. Any one particular computational engineering solution is not entailed by the fact that natural selection designed a certain cognitive mechanism – in principle, there are many possible engineering solutions. In some cases, the computational solution to handling a particular set of invariances will be a Bayesian system. Integrating a phylogenetic perspective (in addition to an ontogenetic one) can provide Bayesian modelers with clear, deductive ways to determine the hypothesis space for a computational system and to set priors.

Going forward: Engineering task analyses. Historical accident aside, Bayesian modeling and evolutionary psychology are not in fact alternative approaches to understanding psychology. Rather, both make necessary but distinct contributions to the process of reverse engineering the mind at a mechanistic level. We are confident that both evolutionary psychology and Bayesian modeling could productively pool their efforts. Evolutionary psychology can provide the framing of task analyses – descriptions of the problem and tasks that cognitive systems must in principle solve. Bayesian models of cognition can provide rigorous, mathematical descriptions of certain types of engineering solutions. We look forward to a time when psychologists choose ecologically valid task analyses and posit fully mechanistic accounts of how the solution to those problems could be implemented by a fully mechanistic system without trying to shoe-horn each reverse engineering task analysis into any common overarching meta-theoretical framework. In the future, we hope there are no evolutionary psychologists or Bayesian modelers, just psychologists who reverse engineer the mind at a mechanistic level, using any and all deductive theoretical tools at their disposal.

Taking the rationality out of probabilistic models

doi:10.1017/S0140525X11000422

Bob Rehder

Department of Psychology, New York University, New York, NY 10003.
bob.rehder@nyu.edu
www.psych.nyu.edu/rehder/

Abstract: Rational models vary in their goals and sources of justification. While the assumptions of some are grounded in the environment, those of others – which I label *probabilistic models* – are induced and so require more traditional sources of justification, such as generalizability to dissimilar tasks and making novel predictions. Their contribution to scientific understanding will remain uncertain until standards of evidence are clarified.

The Jones & Love (J&L) target article begins what is hopefully an extended discussion of the virtues of rational models in psychology. Such discussion is sorely needed because the recent proliferation of such models has not been accompanied by the meta-theoretical understanding needed to appreciate their scientific contribution. When rational models are presented at conferences, the speaker always receives polite applause, but casual conversation afterwards often reveals that many listeners have little idea of how scientific understanding of the topic has been advanced. Even the practitioners (including myself) are often unable to fluently answer the question: "How has the field's understanding of the psychology of X been advanced?" This state of affairs needs to change.

J&L's article may help by clarifying how rational models can vary in their purpose and source of justification. At one extreme, there are models that fall into the "Bayesian Fundamentalism" category and yet are not susceptible to J&L's criticisms. One only need look at a model as old and venerable as signal detection theory (SDT) for an example. SDT specifies optimal behavior, given certain assumptions about the representation of perceptual input, priors, and a cost function. Importantly, the priors (the probability of a signal) and costs (e.g., of a false alarm) can be tied to features of the SDT experiment itself (for a review, see Maloney & Zhang 2010). There are many examples of such models in the domains of perception and action.

But the apparent target of J&L's article are models in which priors are assumed rather than tied to features of an experimental (or any other) context and for which costs of incorrect decisions are unspecified. For example, numerous models specify how one should learn and reason with categories; that is, they assume some sort of prior distribution over systems of mutually exclusive categories (e.g., Kemp & Tenenbaum 2009; Sanborn et al. 2010a). But although this assumption may seem uncontroversial, it is not. Notoriously, even biological species (the paradigmatic example of categories) fail to conform to these assumptions, as there are cases in which the males of one "species" can successfully breed with the females of another, but not vice versa (and cases of successful breeding between As and Bs, and Bs and Cs, but not As and Cs) (Dupre 1981). In what sense should a model that accounts for human categorical reasoning be considered rational when its prior embodies assumptions that are demonstrably false? Of course, the *costs* associated with such ungrounded priors may be small, but models that fail to explicitly consider costs are common. Many rational models in higher-order cognition have this character.

My own modest proposal is that we should drop the label "rational" for these sorts of models and call them what they are, namely, *probabilistic models*. I suggest that freeing probabilistic models from the burden of rationality clarifies both their virtues and obligations. Considering obligations, J&L correctly observe that, if not grounded in the environment, justification for a model's priors must be found elsewhere. But the history of science provides numerous examples of testing whether

postulated hidden variables (e.g., priors in a probabilistic model) exist in the world or in the head of the theorist, namely, through *converging operations* (Salmon 1984). For example, one's confidence in the psychological reality of a particular prior is increased when evidence for it is found across multiple, dissimilar tasks (e.g., Maloney & Mamassian 2009; Rehder & Kim 2010). It is also increased when the probabilistic model not only provides post hoc accounts of existing data but is also used to derive and test *new* predictions. For instance, the case for the psychological reality of SDT was strengthened when perceivers responded in predicted ways to orthogonal manipulations of stimulus intensity and payoff structure. This is how one can treat the assumptions of a probabilistic model as serious psychological claims and thus be what J&L describe as an "enlightened" Bayesian.

Taking the rationality out of probabilistic models also shifts attention to their other properties, and so clarifies for which tasks such models are likely to be successful. By using Bayes' law as the only rule of inference, one's "explanation" of a psychological phenomenon, divided between process and knowledge in classic information-processing models, is based solely on knowledge (priors) instead. Said differently, one might view Bayes' law as supporting a *programming language* in which to express models (a probabilistic analog of how theorists once exploited the *other* normative model of reasoning – formal logic – by programming in PROLOG [programming logic]; Genesereth & Nilsson 1987). These models will succeed to the extent that task performance is determined primarily by human reasoners' prior experience and knowledge. Probabilistic models also help identify variables that are likely to be critical to behavior (i.e., they provide an old-fashioned task analysis; Card et al. 1983); in turn, this analysis will suggest critical ways in which people may differ from one another. Finally, by making them susceptible to analysis, probabilistic models are directing researchers' attention towards entirely new sorts of behaviors that were previously considered too complex to study systematically.

My expectation is that the analysis conducted by J&L will help lead to an appreciation of the heterogeneity among rational/probabilistic models and to clarity regarding the standards to which each should be held. This clarity will not only help conference-goers understand why they are clapping, it will promote the other sorts of virtuous model testing practices that J&L advocate. There *are* examples of Bayesian models being compared with competing models, both Bayesian (Rehder & Burnett 2005) and non-Bayesian ones (e.g., Kemp & Tenenbaum 2009; Rehder 2009; Rehder & Kim 2010), but more are needed. Such activities will help the rational movement move beyond a progressive research program (in Lakatos's terms; see Lakatos 1970) in which research activities are largely confirmatory, to a more mature phase in which the scientific contribution of such models is transparent.

Distinguishing literal from metaphorical applications of Bayesian approaches

doi:10.1017/S0140525X11000434

Timothy T. Rogers and Mark S. Seidenberg

Department of Psychology, University of Wisconsin–Madison, Madison, WI 53726.

ttrogers@wisc.edu **http://concepts.psych.wisc.edu**
seidenberg@wisc.edu **http://lcnl.wisc.edu**

Abstract: We distinguish between literal and metaphorical applications of Bayesian models. When intended literally, an isomorphism exists between the elements of representation assumed by the rational analysis and the mechanism that implements the computation. Thus, observation of the implementation can externally validate assumptions underlying the rational analysis. In other applications, no such isomorphism exists, so it is not clear how the assumptions that allow a Bayesian model to fit data can be independently validated.

Jones & Love's (J&L's) attempt to differentiate uses of Bayesian models is very helpful. The question is, what distinguishes the useful tools from the "fundamentalist" applications? We think one factor is whether Bayesian proposals are intended literally or metaphorically, something that is not usually made explicit. The distinction is exemplified by the different uses of Bayesian theories in studies of vision versus concepts.

In vision, computational analyses of the statistics of natural scenes have yielded hypotheses about representational elements (a class of basis functions) that provide a putatively optimally efficient code (Simoncelli & Olshausen 2001). The fact that neurons in visual cortex have receptive fields that approximate these basis functions was a major discovery (Olshausen & Field 1996). Thus, there is a direct, rather than metaphorical, relation between a rational hypothesis about a function of the visual system and its neurobiological basis. It is easy to see how the firing activity of a visual neuron might literally implement a particular basis function, and thus, how the pattern of activation over a field of such neurons might provide an efficient code for the statistics of the visual scene. This isomorphism is not merely coincidental.

In metaphorical applications, no such mapping exists between the proposed function and implementation. People are assumed to compute probability distributions over taxonomic hierarchies, syntactic trees, directed acyclic graphs, and so on, but no theorist believes that such distributions are directly encoded in neural activity, which, in many cases, would be physically impossible. For instance, Xu and Tenenbaum (2007b) have proposed that, when learning the meaning of a word, children compute posterior probability distributions over the set of all possible categories. If there were only 100 different objects in a given person's environment, the number of possible categories (2^{100}, or $\sim 1.27 \times 10^{30}$) would exceed the number of neurons in the human brain by about 19 orders of magnitude. Thus, theorists working in this tradition disavow any direct connection to neuroscience, identifying the work at Marr's computational level (Marr 1982). The idea seems to be that, although the brain does not (and cannot) actually compute the exact posterior probability distributions assumed by the theory, it successfully approximates this distribution via some unknown process. Since any method for approximating the true posterior distribution will achieve the same function, there is no need to figure out how the brain does it.

The problem is that this approach affords no way of externally validating the assumptions that enable the Bayesian theory to fit data, including assumptions about the function being carried out, the structure of the hypothesis space, and the prior distributions. This limitation is nontrivial. Any pattern of behavior can be consistent with some rational analysis if the underlying assumptions are unconstrained. For instance, given any pattern of behavior, one can always work backward from Bayes' rule to find the set of priors that make the outcomes look rational. Thus, good fit to behavioral data does not validate a Bayesian model if there is no independent motivation for the priors and other assumptions. The strongest form of independent motivation would be external validation through some empirical observation not directly tied to the behavior of interest, as in the vision case: Conclusions from the rational analysis (i.e., that a particular basis function provides an optimally efficient code, so vision must make use of such basis functions) were validated through empirical observation of the receptive fields of neurons in visual cortex. But this kind of external validation is not available in cases where the mapping between the rational analysis and neural implementation is unknown.

Much of this is familiar from earlier research on language. Bayesian cognitive theories are competence theories in Chomsky's (1965) sense. Like Chomskyan theories, they make strong a priori commitments about what the central functions are and how knowledge is represented, and they idealize many aspects of performance in the service of identifying essential truths. The links between the idealization and how it is acquired, used, or represented in the brain are left as promissory notes – still largely unfulfilled in the case of language. But the language example

suggests that the idealizations and simplifications that make a competence (or "computational") theory possible also create non-isomorphisms with more realistic characterizations of performance and with brain mechanisms (Seidenberg & Plaut, in press). The situation does not materially change because Bayesian theories are nominally more concerned with how specific tasks are performed; the result is merely *competence theories of performance*.

As J&L note in the target article, similar issues have also arisen for connectionism over the years, with critics arguing that connectionist models can be adapted to fit essentially any pattern of data. There is a key difference, however: The connectionist framework is intended to capture important characteristics of neural processing mechanisms, so there is at least the potential to constrain key assumptions with data from neuroscience. This potential may not be realized in every instantiation of a connectionist model, and models invoking connectionist principles without connection to neural processes are subject to the same concerns we have raised about Bayesian models. But it is becoming increasingly common to tie the development of such models to observations from neuroscience, and this marriage has produced important and productive research programs in memory (Norman & O'Reilly 2003; O'Reilly & Norman 2002), language (Harm & Seidenberg 2004; McClelland & Patterson 2002), cognitive control (Botvinick et al. 2001), routine sequential action (Botvinick & Plaut 2004), and conceptual knowledge (Rogers & McClelland 2004; Rogers et al. 2004) over the past several years. Bayesian approaches will also shed considerable light on the processes that support human cognition in the years to come, when they can be more closely tied to neurobiological mechanisms.

Bayesian computation and mechanism: Theoretical pluralism drives scientific emergence

doi:10.1017/S0140525X11000392

David K. Sewell,[a] Daniel R. Little,[a] and Stephan Lewandowsky[b]

[a]*Department of Psychological Sciences, The University of Melbourne, Melbourne, VIC 3010, Australia;* [b]*School of Psychology, The University of Western Australia, Crawley, WA 6009, Australia.*

dsewell@unimelb.edu.au **daniel.little@unimelb.edu.au**
lewan@psy.uwa.edu.au
http://www.psych.unimelb.edu.au/people/staff/SewellD.html
http://www.psych.unimelb.edu.au/research/labs/knowlab/index.html
http://www.cogsciwa.com/

Abstract: The breadth-first search adopted by Bayesian researchers to map out the conceptual space and identify what the framework can do is beneficial for science and reflective of its collaborative and incremental nature. Theoretical pluralism among researchers facilitates refinement of models within various levels of analysis, which ultimately enables effective cross-talk between different levels of analysis.

The target article by Jones & Love (J&L) is another entry to the recent debate contrasting the merits of Bayesian and more mechanistic modeling perspectives (e.g., Griffiths et al. 2010; McClelland et al. 2010). Regrettably, much of this debate has been tainted by a subtext that presupposes the approaches to be adversarial rather than allied (see, e.g., Feldman 2010; Kruschke 2010). J&L are correct in asserting that research agendas pitched at different levels of analysis will investigate different research questions that lead to different theoretical solutions (e.g., Dennett 1987; Marr 1982/2010). However, any complete psychological theory must account for phenomena at multiple levels of analysis and, additionally, elucidate the relations between levels (e.g., Schall 2004; Teller 1984). We also note that the various levels of analysis are causally interrelated and are thus mutually constraining (Rumelhart & McClelland 1985). It follows that refinement of a model at one level of analysis focuses the search for theoretical solutions at another. We therefore view theoretical pluralism among researchers as an efficient means of developing more complete psychological theories.

We suggest that findings from the so-called "Bayesian Fundamentalist" perspective have highlighted core issues in developing more complete psychological theories, and that discoveries by individual "Fundamentalist" researchers may actually facilitate discipline-wide "Enlightenment" by sharpening questions and generating novel insights that stimulate research (e.g., Shiffrin et al. 2008). J&L's admonishment of Bayesian Fundamentalism, depending on whether it is directed at psychological science as a whole, or to individual researchers, is either a) powerful but directed at a largely non-existent opponent, or (b) misguided insofar that the collaborative nature of scientific progress offsets the narrow focus of individual scientists.

Contrary to J&L, we argue the "breadth-first" approach adopted by many Bayesian theorists, rather than stifling theoretical progress, actually facilitates cross-talk between levels of analysis. That contemporary Bayesian theorists are aware of, and aspire to resolve this tension, is reflected in recent work that has sought to reconcile rational accounts with more traditional process models. For example, to the extent that models of cognitive processing implement sampling algorithms to approximate full Bayesian inference, models at different levels of analysis can be mutually informative. Shi et al. (2010) illustrate how exemplar models (e.g., Nosofsky 1986) can be interpreted as an importance sampling algorithm, and, similarly, Sanborn et al.(2010a) explored the particle filter algorithm as a way of leveraging a process interpretation of Anderson's (1991b) rational model. Lewandowsky et al.(2009) used *iterated learning* (Griffiths & Kalish 2007; Kalish et al. 2007), an experimental paradigm motivated by technological advances in sampling techniques used to approximate Bayesian posteriors, to decisively reject a sparse-exemplar model of predicting the future. Kruschke (2006; 2008) contrasted globally and locally Bayesian approaches to associative learning, the latter of which can be construed as carrying very direct process implications concerning selective attention. J&L acknowledge the potential of these approaches for transcending computational level theories but do not acknowledge the role of the computational theories for driving research in this direction.

One area where Bayesian perspectives appear particularly more illuminating than mechanistic approaches is in explaining individual differences. For example, work from within the knowledge partitioning framework has repeatedly found large differences in transfer performance in tasks that can be decomposed into a number of simpler sub-tasks (e.g., Lewandowsky et al. 2002; 2006; Yang & Lewandowsky 2003). Mechanistic modeling of these results has highlighted the importance of modular architecture (Kalish et al. 2004; Little & Lewandowsky 2009), selective attention (Yang & Lewandowsky 2004), and their interaction (Sewell & Lewandowsky 2011) in accounting for such individual differences. However, a significant limitation of a mechanistic approach is that the solutions have been built into the models. By contrast, recent Bayesian modeling of knowledge partitioning has showed that many aspects of the individual differences observed empirically emerge naturally if one assumes that people are trying to learn about their environment in a rational manner (Navarro 2010).

J&L draw uncharitable parallels between "Bayesian Fundamentalism" on the one hand, and Behaviorism, connectionism, and evolutionary psychology on the other. In response, we note that theoretical setbacks in those paradigms have clarified our understanding of how the mind does and does not work. Consequently, cognitive science has emerged with a more refined theoretical toolkit and new, incisive research questions. For Behaviorism, a restrictive theoretical stance solidified the need

to consider more than just the history of reinforcement in explaining behavior (Neisser 1967). The inability of the perceptrons to handle nonlinearly separable problems forced connectionists to consider more powerful model architectures (Thomas & McClelland 2008). Likewise, controversies that have erupted in evolutionary psychology over the propagation of cognitive modules have forced theorists to refine and reevaluate classical notions of modularity (cf. Barrett & Kurzban 2006; Fodor 1983). Thus, the failures of the precedents chosen by J&L actually constitute successes for the field; for example, the cognitive revolution was propelled and accelerated by the spectacular failure of Behaviorism.

We close by considering how J&L's critique of Bayesian Fundamentalism relates to scientific activity in practice. If they address the scientific community as a whole, their criticism is powerful, but lacks a real target. Alternatively, if J&L's concerns are directed at individual scientists, their plea overlooks the fact that scientific progress, being inherently distributed across multiple research groups, "averages out" individual differences in theoretical dispositions. That is, the aggregate outcomes produced by the scientific community are unlikely to be reflected in the individual outcomes produced by a given scientist (Kuhn 1970).

Whereas a complete level-spanning theory will always be the goal of science, the approach toward that collective goal will be incremental, and those pursuing it will tend to focus on a particular level of analysis. The important question for any individual researcher is whether an adopted theoretical framework sharpens questions, provides insight, and guides new empirical inquiry (Shiffrin et al. 2008); recent Bayesian modeling of cognition undoubtedly fulfills these requirements.

Is everyone Bayes? On the testable implications of Bayesian Fundamentalism

doi:10.1017/S0140525X11000409

Maarten Speekenbrink and David R. Shanks

Division of Psychology and Language Sciences, University College London, London WC1E 6BT, United Kingdom.

m.speekenbrink@ucl.ac.uk d.shanks@ucl.ac.uk

http://www.psychol.ucl.ac.uk/m.speekenbrink

http://www.psychol.ucl.ac.uk/david.shanks/Shanks.html

Abstract: A central claim of Jones & Love's (J&L's) article is that Bayesian Fundamentalism is empirically unconstrained. Unless constraints are placed on prior beliefs, likelihood, and utility functions, all behaviour – it is proposed – is consistent with Bayesian rationality. Although such claims are commonplace, their basis is rarely justified. We fill this gap by sketching a proof, and we discuss possible solutions that would make Bayesian approaches empirically interesting.

Although the authors are perhaps attacking a straw-man, we agree with many points raised in Jones & Love's (J&L's) critique of "Bayesian Fundamentalism." It is our objective here to strengthen their claim that Bayesian Fundamentalism is empirically unconstrained; although such claims are often made, their basis is not usually fleshed out in any detail. This is such a key part of the case that we sketch a proof and discuss possible solutions.

Without placing constraints on prior beliefs, likelihood, and utility functions, claims of Bayesian rationality are empirically empty: any behaviour is consistent with that of some rational Bayesian agent. To illustrate this point, consider a simple probability learning task in which a participant has two response options (e.g., press a left or a right button), only one of which will be rewarded. On each trial t, the participant gives a response $x_t = \{0,1\}$, and then observes the placement of the reward $y_t = \{0,1\}$, which is under control of the experimenter. The question is whether the assumption of Bayesian rationality places any restrictions on the response sequence for a given reward sequence.

In Bayesian inference, the prior distribution and likelihood (model of the task) assign a probability $P(y_t = S_j)$ to each possible reward sequence. Without further constraints, we can take this probability to be proportional to a value $v_j \geq 0$. After observing y_1, some of the rewarded sequences are impossible, and learning consists of setting the probability of these sequences to 0 and then renormalizing. For example, consider a task with three trials. The possible reward (and response) sequences are given in Table 1. Assume the sequence of rewards is $y = S_1$. After observing $y_1 = 0$, S_5 to S_8 are impossible and the posterior probabilities become $P(S_j|y_1) = v_j/\Sigma_k v_k$, for $j, k = 1, \ldots, 4$, and $P(S_j|y_1)=0$ for $j = 5, \ldots, 8$. After observing $y_2 = 0$, S_3 and S_4 are also impossible, and the posterior probabilities become $P(S_j|y_1) = v_j/\Sigma_j v_k$, for $j, k = 1, 2$, and $P(S_j|y_1) = 0$, for $j = 3, 4$. After observing $y_3 = 0$, only S_1 remains with a probability 1.

A rational Bayesian agent gives responses which maximise his or her subjective expected utility, conditional upon the previously observed rewards. For simplicity, assume the utility of a correct prediction is $u(y_t = x_t)=1$ and that of an incorrect prediction is $u(y_t \neq x_t) = 0$, so that the expected utilities correspond to the posterior predicted probabilities of the next reward. The crucial point is that in this general setup, we can always choose the values v_j to make any sequence of responses x_t conform to that of a maximizer of subjective expected utility. For example, suppose the sequence of rewards is S_1 and the sequence of responses is S_8. The first response $x_1=1$ implies that $v_1 + v_2 + v_3 + v_4 v_5 v_6 v_7 v_8$; the second response $x_2 1$ implies that $v_1 v_2 v_3 v_4$; the third response $x_3\ 1$ implies that $v_1 v_2$. One choice of values consistent with this is v_j. For any response sequence, we can choose values which adhere to such implied inequalities, so behaviour is always consistent with a rational Bayesian agent. Although we have considered a rather simple situation with a small number of trials, this result generalizes readily to other sequential learning tasks such as category learning (for a related, more general and formal proof, see, e.g., Zambrano 2005). The problem becomes even more severe if we allow the utilities to depend on previous outcomes, which may not be entirely implausible (e.g., a third misprediction in a row may be more unpleasant than the first).

One may object that the particular method of Bayesian inference sketched here is implausible: Would someone really assign probabilities to all possible reward sequences? Maybe not explicitly, but in an abstract sense, this is what Bayesian modelling boils down to. Granted, the values assigned have been arbitrary, but that is exactly the point: Bayesian rationality is silent about the rationality of priors and likelihoods, yet some of these seem more rational than others. Thus, rationality hinges on more than adherence to Bayesian updating and utility maximization.

Is the claim of Bayesian inference and decision making always empirically empty? No. For instance, the assumption that rewards are *exchangeable* (that they can be reordered without affecting the probabilities) places equivalence restrictions on the values v such that, given a sufficient number of trials, some response sequences would violate utility maximization. Exchangeability is crucial to the convergence of posterior probabilities and the decisions based on them. Another option would be to let participants make multiple decisions while keeping their

Table 1 *(Speekenbrink & Shanks). Possible reward and response sequences (S_j) in a simple learning task with three trials (t)*

t	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
1	0	0	0	0	1	1	1	1
2	0	0	1	1	0	0	1	1
3	0	1	0	1	0	1	0	1

information base (posterior probabilities) constant, so that intransitive decisions become possible. More generally, testable conditions of Bayesian rationality can be found in the axioms of subjective expected utility theory (e.g., Savage 1954). Empirically meaningful claims of Bayesian rationality should minimally ensure the possibility that the data can falsify these axioms. Axiomatic tests are "model-free" in the sense that they do not rely on a particular choice of prior distribution and utility function. Such tests should be a first step in rational analysis; if the assumption of Bayesian rationality is not rejected, one can then look for priors and utilities which match the observed behaviour. Given rich-enough data, this search can be guided by conjoint measurement procedures (e.g., Wallsten 1971).

To conclude, while "Bayesian Fundamentalism" is generally unconstrained, by placing appropriate restrictions, the assumption of Bayesian rationality is subject to empirical testing and, when not rejected, can help guide model building.

Post hoc rationalism in science

doi:10.1017/S0140525X11000410

Eric Luis Uhlmann

HEC Paris—School of Management, Management and Human Resources Department, 78351 Jouy-en-Josas, France.

eric.luis.uhlmann@gmail.com

www.socialjudgments.com

Abstract: In advocating Bayesian Enlightenment as a solution to Bayesian Fundamentalism, Jones & Love (J&L) rule out a broader critique of rationalist approaches to cognition. However, Bayesian Fundamentalism is merely one example of the more general phenomenon of Rationalist Fundamentalism: the tendency to characterize human judgments as rational and optimal in a post hoc manner, after the empirical data are already known.

Jones & Love (J&L) are right to criticize what they term "Bayesian Fundamentalism" as not empirically grounded, uninformed by psychological data, open to multiple rational accounts of a task or decision, and conducive to post hoc explanations. However, in advocating Bayesian Enlightenment as a solution, they appear to rule out a broader critique of rationalist approaches to human cognition. Specifically, Bayesian Fundamentalism is one example of the more general phenomenon of *Rationalist Fundamentalism*: the tendency to characterize a given judgment as rational and optimal in a post hoc manner, after the empirical data are already known. Few researchers would argue that human behavior is perfectly optimal and rational. However, a desire to see the human mind as operating rationally, and the use of post hoc justifications to reach this conclusion, suggest we should be skeptical of after-the-fact "rational" explanations.

Decades of empirical studies show people are strongly motivated to see themselves as rational and objective (for reviews, see Armor 1999; Pronin et al. 2004; Pyszczynski & Greenberg 1987; Ross & Ward 1996). Decision makers engage in motivated reasoning and psychological rationalizations designed to preserve this "illusion of objectivity" (Armor 1999; Pronin et al. 2002) – for instance, changing their definition of what an optimal judgment is after the fact (Dunning & Cohen 1992; Epstein et al. 1992; Kunda 1987; Norton et al. 2004; Uhlmann & Cohen 2005). Evidence that general psychological processes are not rational or optimal represents a threat to this cherished illusion. Fundamentalist resistance to evidence of human irrationality further stems from economics and related disciplines, in which optimality and the maximization of utility are widely perceived as necessary assumptions about human behavior.

A rationalist defense can involve constructing a post hoc Bayesian account of an empirical finding predicted a priori from theories grounded in psychological limitations and motives. Consider the phenomenon of *biased assimilation*, in which participants rate a scientific study that supports their political beliefs (e.g., about the deterrent effects of capital punishment) as methodologically superior to a study that refutes their beliefs (Lord et al. 1979). The cognitive-rationalist interpretation is that decision makers are simply making Bayesian inferences, taking into account subjective probabilities (e.g., their prior political beliefs) when evaluating new evidence. However, further findings contradict the claim that biased assimilation is merely the product of Bayesian inferences. For instance, individuals whose positive self-image is affirmed are less likely to exhibit biased assimilation (Cohen et al. 2000; see also Dunning et al. 1995; Sherman & Cohen 2002). This is consistent with the idea that biased information processing stems from a motivated desire to dismiss evidence that threatens valued beliefs and, by extension, the self (Sherman & Cohen 2006; Steele 1988). When a decision maker is feeling good about herself, there is less need to be biased. In addition, would-be parents who believe day care is bad for children, but plan to use day care themselves (and therefore desire to conclude that day care is just as good as home care), show biased assimilation in favor of day care (Bastardi et al. 2011). What decision makers *desire* to be true seems to trump what they *believe* to be factually true – the ostensive basis for any Bayesian inferences.

As J&L point out, one of the most problematic aspects of rational models is how little attention can be paid to whether the assumptions of the statistical model correspond to what is actually going on in people's heads as they engage in a task or make a decision. I once debated an economist who argued that micro-level psychological data on what goals people pursue in the dictator game are irrelevant: The material self-interest account *must* be true if people's offers correspond to the predictions of the statistical model. However, it is dangerous to assume that because a rational statistical model can mimic or reproduce a pattern of data, the underlying psychological process is a rational one. That a computer can mimic some of the outputs of human thought does not necessarily mean the mind functions in the same way as a computer.

The last defense of post hoc rationalism is to swap normative models of rationality entirely. In other words, researchers can speculate post-hoc as to what alternative goals decision-makers may have been pursuing, in order to preserve the view that participants were acting rationally. Never mind the goals to optimize material outcomes or achieve accuracy: Judgmental biases can be defined as "rational" because they preserve the decision maker's personal self-image, psychological well-adjustment, public reputation, cherished religious beliefs, desire to punish norm violators, existential goals, likelihood of survival in ancestral environments, or even the happiness of their marriage (Cosmides & Tooby 1994; Hamilton 1980; Krueger & Funder 2004; Lerner & Tetlock 1999; Tetlock 2002; Tetlock et al. 2000; 2007).

It has been argued that the heuristics-and-biases approach to cognition is itself biased, in the direction of attributions to irrationality (Krueger & Funder 2004). Despite its shortcomings, however, the heuristics-and-biases research program is at least based on a priori theoretical hypotheses. There are few cases of "post hoc irrationalism" in which robust empirical effects predicted a priori by Bayesian or otherwise rationalist models are redefined post hoc as due to motives such as the need for self-esteem or control.

Although Bayesian Enlightenment, as advocated by J&L, is a major improvement on Bayesian Fundamentalism, it is still subject to post hoc rationalism. An interface between Bayesian or otherwise rationalist models and data on psychological processes leaves plenty of room for the former to distort interpretations of the latter. A wealth of evidence indicates that human beings are subject to a powerful illusion of rationality and objectivity they are strongly motivated to maintain and which influences their perceptions of scientific data. Researchers are also human beings. It would be remarkable indeed if scientists were immune to the empirical phenomena we study.

Authors' Response

Pinning down the theoretical commitments of Bayesian cognitive models

doi:10.1017/S0140525X11001439

Matt Jones[a] and Bradley C. Love[b]

[a]*Department of Psychology and Neuroscience, University of Colorado, Boulder, CO 80309;* [b]*Department of Psychology, University of Texas, Austin, TX 78712*

mcj@colorado.edu **brad_love@mail.utexas.edu**

Abstract: Mathematical developments in probabilistic inference have led to optimism over the prospects for Bayesian models of cognition. Our target article calls for better differentiation of these technical developments from theoretical contributions. It distinguishes between Bayesian Fundamentalism, which is theoretically limited because of its neglect of psychological mechanism, and Bayesian Enlightenment, which integrates rational and mechanistic considerations and is thus better positioned to advance psychological theory. The commentaries almost uniformly agree that mechanistic grounding is critical to the success of the Bayesian program. Some commentaries raise additional challenges, which we address here. Other commentaries claim that all Bayesian models are mechanistically grounded, while at the same time holding that they should be evaluated only on a computational level. We argue this contradictory stance makes it difficult to evaluate a model's scientific contribution, and that the psychological commitments of Bayesian models need to be made more explicit.

R1. Introduction

The rapid growth of Bayesian cognitive modeling in recent years has outpaced careful consideration and discussion of what Bayesian models contribute to cognitive theory. Our target article aimed to initiate such a discussion. We argued there is a serious lack of constraint in models that explain behavior based solely on rational analysis of the environment, without consideration of psychological mechanisms, but that also fail to validate their assumptions about the environment or the learner's goals.

We referred to the approach of Bayesian modeling without consideration of mechanism as *Bayesian Fundamentalism*. We went on to advocate an approach we labeled *Bayesian Enlightenmen*t, in which elements of a Bayesian model are given a psychological interpretation, by addressing how the learner's hypotheses are represented, where they come from, what the learner's goals are, and how inference is carried out. Although several commentators argue for further challenges or shortcomings, no serious challenge was offered to the conclusion that, at the least, Bayesian models need this type of grounding. Primarily, the commentaries serve to reinforce, in various ways, the idea that it is critical to be clear on the psychological commitments and explanatory contributions of cognitive models.

Technical breakthroughs can often enable new theoretical progress, by allowing researchers to formalize and test hypotheses in ways that were not previously possible. Thus, development of new formal frameworks can be important to the progress of the field as a whole (**Chater, Goodman, Griffiths, Kemp, Oaksford, & Tenenbaum** [**Chater et al.**]; **Navarro & Perfors**). However, technical advances are not theories themselves, and there is a real danger in confusing the two. As cognitive scientists well know, it is critical for modelers to clarify which aspects of a model are meant as psychological commitments and which are implementation details. For example, sophisticated sampling methods for estimating posterior distributions enable derivation of predictions from complex Bayesian models that were previously intractable. However, if these approximation algorithms are not intended as psychological mechanisms, then any deviations they produce from optimality should not be taken as necessary predictions of the model. Likewise, probabilistic methods for specifying priors over structured hypotheses may enable computational analysis of new learning domains. Again, if the particular assumptions built into the hypothesis space are not meant as claims about the learner's knowledge or expectations (i.e., other choices would have been equally reasonable), then many predictions of the model should not be taken as necessary consequences of the underlying theory. Thus, when implementation decisions are not clearly separated from theoretical commitments, one cannot tell what the model's real predictions are, or, consequently, how it should be tested. For the same reasons, it can be unclear what new understanding the model provides, in terms of what was explained and what the explanation is (**Rehder**). In short, one cannot evaluate the model's scientific contribution.

In this reply, we argue there is still serious confusion and disagreement about the intended status of most Bayesian cognitive models. We then evaluate the potential theoretical contribution of Bayesian models under different possible interpretations. When Bayesian models are cast at a purely computational level, they are mostly empty. When Bayesian models are viewed as process models, they have potentially more to say, but the interesting predictions emerge not from Bayes' rule itself but from the specific assumptions about the learner's hypotheses, priors, and goals, as well from questions of how this information is represented and computed. Thus, we advocate shifting attention to these assumptions, viewed as psychological commitments rather than technical devices, and we illustrate how this stance shifts attention to important psychological questions that have been largely neglected within the Bayesian program to date. Finally, we consider several other challenges raised to the Bayesian program, and specifically to the proposed integration with mechanistic approaches that we labeled Bayesian Enlightenment. We conclude that the Bayesian framework has potential to add much to cognitive theory, provided modelers make genuine psychological commitments and are clear on what those commitments are.

R2. Theoretical status of Bayesian models

A primary confusion surrounding Bayesian cognitive models is whether they are intended as purely computational-level theories, or whether certain components of the model are to be taken as claims regarding psychological mechanism. Specifically: Are hypotheses and priors assumptions about the environment or the learner? That is, are they devices for the modeler to specify the

assumed statistical structure of the environment, or are they meant as psychological constructs? Are algorithms for approximating optimal inference to be viewed as tools for deriving model predictions or as psychological processes? More broadly, does the brain represent information in terms of probabilities, or does it just behave as though it does? Unfortunately, these questions are not always answered, and those answers that are given are often contradictory. This state of affairs seriously limits scientific evaluation of Bayesian models and makes it difficult to determine their explanatory contribution.

For all of our criticisms of J. R. Anderson's (1990) rational analysis in the target article, his viewpoint is clear and coherent. According to J. R. Anderson, rational models are distinguished from mechanistic models in that rational models do not make reference to mental representations or processes. Instead, these models specify relevant information structures in the environment and use optimal inference procedures that maximize performance for the assumed task goal. We labeled this view (in the context of probabilistic models) as Bayesian Fundamentalism in the target article and offered an unfavorable critique. On the positive side, the fundamentalist view is theoretically clear, whereas much of contemporary Bayesian modeling is not.

Indeed, we find many of the commentaries theoretically confusing and contradictory. Certainly, self-identified Bayesian advocates contradict one another. For example, **Gopnik** states that Bayesian models have psychological representations but not processes, whereas **Borsboom, Wagenmakers, & Romeijn** (**Borsboom et al.**) claim they are not representational but are process models. Borsboom et al.'s position is particularly curious because they assert that a Bayesian model is a process model but not a mechanistic model. This position contradicts their own definitions, as it is impossible to specify the state dynamics of a system (the process model, their terms) without specifying the system itself (the mechanism).

These different views on what constitutes a Bayesian model highlight that the theoretical underpinnings of models are not always as clear as one would hope. In mechanistic models, it is clear that key processing and representation claims involve postulated mental entities. In the fundamentalist rational view, it is clear that process and representation do not refer to mental entities. Unfortunately, many Bayesian models seem to waver among various intermediate positions. For example, positing one component of a model (e.g., process or representation) as a mental entity and the other as not, may evoke Cartesian dualism, in which ontologically different entities (e.g., non-physical and physical) interact. If one is not careful about the statuses of all model components, it is easy for them to slip from one to the other, making the model's position and contribution uncertain. Therefore, more care needs to be taken in spelling out exactly what kind of model one is specifying and its intended contribution (**Bowers & Davis**; **Fernbach & Sloman**).

Part of this confusion arises because terms like "representation" mean different things to different self-identified Bayesians and, more worrisome, can shift meaning within a single contribution. To be clear, mental representations (as opposed to mathematical representations of probability distributions in the world) are in the head and are acted on by mental processes. For example, in the Sternberg (1966) model of short-term memory, the mental representation of items in short-term memory consists of an ordered buffer that is operated over by an exhaustive search process. This is not a model of optimal inference based on environmental regularities but is, instead, an account of how information is represented and manipulated in the head. The specified mental processes and representations make predictions for response time and error patterns, and these predictions can be used to evaluate the model and explore implementational questions.

We find the slipperiness and apparent self-contradictions of some Bayesian proposals regarding their psychological status to be theoretically unhelpful. For example, **Chater et al.** state that, unlike Behaviorism, Bayesian cognitive science posits mental states, but then they contradict this position by stating that these theories are positioned at a computational level (in the sense of Marr 1982) and don't need to address other levels of explanation. We agree with Chater et al. that technical advances have led to a greater range of representations in Bayesian models, but if these models reside at the computational level then these are representations of probability distributions, not mental representations. That is, they reside in the head of the researcher, not the subject. Finally, Chater et al. emphasize the importance of descriptions of structured environments in the sense of J. R. Anderson's (1990) rational program (i.e., Bayesian Fundamentalism), which again contradicts claims that the Bayesian models they discuss have mental representations. There are many interesting ideas in this commentary, but it is impossible to integrate the points into a coherent and consistent theoretical picture.

We agree with **Fernbach & Sloman** that "modelers are not always as clear as they should be about whether these hypotheses represent psychological entities or merely a conceptual analysis of the task (or both), and the import of the model does depend critically on that." However, even these commentators confuse the status of Bayesian constructs. Fernbach & Sloman claim that Bayesian hypotheses constitute more than probability distributions over data; that, instead, they always correspond to psychological constructs or mental models relevant to the task in question – in direct contradiction to the previous quote from their commentary. If hypotheses are not psychological constructs, then indeed they are nothing but elements of the probabilistic calculus the modeler uses to derive predictions from the model. It should not be controversial that many Bayesian models used in ideal observer analyses do not contain mental representations, but are instead models of the task environment, just as it is uncontroversial that Bayesian models used in physics, chemistry, credit fraud detection, and so forth, do not contain mental representations.

Even within the cognitive sciences, Bayesian methods are often used as analysis tools (see discussion of "Agnostic Bayes" in the target article) that are not intended as psychological theories. Indeed, as **Lee** discusses, such methods provide a powerful means for evaluating all types of models. Lee notes that, oddly, many articles hold up Bayesian inference as the paragon of rationality and then test their models by using Frequentist statistics. This practice makes one wonder how strongly Bayesian

modelers truly believe in the rational principles of their theories. Lee's proposal to use Bayesian model selection to evaluate Bayesian cognitive models seems more self-consistent, and we agree that the Bayesian approach offers many useful tools for evaluating and comparing complex models (although some of the advantages he cites, such as parameter estimation and testing hierarchical models, are also compatible with maximum-likelihood techniques and Frequentist statistics).

As commentators **Glymour**, **Rehder**, and **Rogers & Seidenberg** have highlighted, it can be difficult to know what one is to take away from some Bayesian accounts. As these commentators discuss, hugely complex hypothesis spaces are often proposed but with no claim that people perform inference over these spaces in the manner the models do; and any connection with neuroscience is disavowed in favor of theory residing solely at the computational level. When models do make connections with broader efforts, the message can become confused. For example, **Borsboom et al.** assert that mechanisms for belief updating reside in the brain and can be studied to provide support for Bayesian models, but they then appeal to notions of optimality, stating that the substrate of computation is completely unimportant and only fitting behavioral data matters.

In conclusion, although we provide ample examples of Bayesian Fundamentalist contributions in the target article, we might have to agree with those commentators (**Chater et al.**; **Gopnik**; **Sewell, Little, & Lewandowsky** [**Sewell et al.**]) who argue there are no Bayesian Fundamentalists, because it is not always clear what position many Bayesians support. This lack of theoretical clarity is potentially a greater threat to theoretical progress than is the Bayesian Fundamentalist program itself. When the intended status of a Bayesian model is not made explicit, assumptions such as the choice of goals and hypothesis space can be referred to in vague language as constituting knowledge or representation, but when the assumptions are contradicted by data, the modeler can fall back on the computational position and say they were never intended to be psychologically real. The result is a model that appears to have rich representational structure and strong psychological implications, but which, when prodded, turns out to be quite empty.

R3. Bayesian models as computational-level theories

Setting aside how Bayesian models have been intended – which we have argued is often unclear – we now evaluate their potential theoretical contribution under a purely computational-level interpretation. By the "computational level" we mean the standard position taken by rational analysis (e.g., J. R. Anderson 1990) that one can explain aspects of behavior solely by consideration of what is optimal in a given environment, with no recourse to psychological constructs such as knowledge representation or decision processes. Our aim is to draw out the full implications of this position once a Bayesian model is truly held to it, rather than being afforded the sort of slipperiness identified earlier. As **Norris** points out, J. R. Anderson was aware of and cautioned against many of the limitations of his rational approach, but much of that message seems to have been lost amidst the expressive power of the Bayesian framework.

It is generally recognized that the specific representations of hypotheses and the algorithms for updating belief states are not meant as psychological commitments of a computational-level Bayesian model. However, the situation is more severe than this, because on a true computational-level stance the entire Bayesian calculus of latent variables, hypotheses, priors, likelihoods, and posteriors is just an analytic device for the modeler. Priors and likelihoods (as well as any hierarchical structure in the hypothesis space) are mathematically equivalent to a "flat" or unstructured model that directly specifies the joint distribution over all observations. Computing a posterior and using it to predict unobserved data is equivalent to calculating the probabilities of the unobserved data conditioned on observed data, with respect to this joint distribution. If process is irrelevant, then these conditional probabilities are the only content to a Bayesian model. That is, the model's only assertion is that people act in accordance with probabilities of future events conditioned on past events. In other words, people use past experience to decide what to do or expect in the future. The model says nothing whatsoever beyond this extremely general position, other than that decisions are optimal in a probabilistic sense, due to unspecified processes and with respect to (usually) unvalidated assumptions about the statistics of the environment.

Contrary to **Chater et al.**'s claim, this interpretation of a Bayesian model is very much like Behaviorism in its deliberate avoidance of psychological constructs. To argue, as Chater et al. do in point (iii) of their commentary's section 2, that "Behaviorists believe that no such computations exist, and further that there are no internal mental states over which such computations might be defined" is a misreading of Behaviorist philosophy. The actual Behaviorist position (e.g., Skinner 1938) was that psychological states are unobservable (not nonexistent) and hence should not be elements of scientific theories, and that behavior should be explained directly from the organism's experience. This position aligns very closely with the motivations offered for computational-level modeling based on rational analysis (e.g., J. R. Anderson 1990). Although Bayesian modeling generally involves significant computation, if the models are to be interpreted at the computational level, then by definition these computations have nothing to do with psychological states.

As noted in the target article, a strong case has been made that probabilistic inference is the best current framework for normative theories of cognition (Oaksford & Chater 2007). However, this observation does not say much about actual cognitive processes or the representations on which they operate. To state, as **Edelman & Shahbazi** do, that all viable approaches ultimately reduce to Bayesian methods does not imply that Bayesian inference encompasses their explanatory contribution. Such an argument is akin to concluding that, because the dynamics of all macroscopic physical systems can be modeled using Newton's calculus, or because all cognitive models can be programmed in Python, calculus or Python constitutes a complete and correct theory of cognition. This is not to say the rational principles are irrelevant, but they are not the whole story.

Furthermore, although ecological rationality can be a powerful explanatory principle (e.g., Gibson 1979; Gigerenzer & Brighton 2009), most Bayesian cognitive

models fail to realize this principle because they are not based on any actual measurement of the environment. This is a serious problem for a Bayesian model interpreted at the computational level, because, as just explained, statistical properties of the environment (specifically, probabilities of future events conditioned on past events), together with the learner's goals, constitute the entire content of the model. The fact that these properties are free to be chosen post hoc, via specification of hypotheses and priors, significantly compromises the theoretical contributions of Bayesian models (**Anderson**; **Bowers & Davis**; **Danks & Eberhardt**; **Glymour**; **Rehder**; **Rogers & Seidenberg**). The sketch proof by **Speekenbrink & Shanks** shows how nearly any pattern of behavior is consistent with Bayesian rationality, under the right choice of hypotheses, priors, and utility functions. **Rehder** goes as far as to suggest viewing the Bayesian framework as a programming language, in which Bayes' rule is universal but fairly trivial, and all of the explanatory power lies in the assumed goals and hypotheses. Thus, the basis of these assumptions requires far more scrutiny than is currently typical.

As with any underconstrained model, a Bayesian model developed without any verification of its assumptions is prone to overfit data, such that it is unlikely to extend to new situations. Hence, whereas **Borsboom et al.** argue that Bayesian models should not be constrained by mechanism as long as they can match existing data, we suggest such an approach is unlikely to predict new data correctly. The observations by **Jenkins, Samuelson, & Spencer** (**Jenkins et al.**) on the fragility of the *suspicious coincidence* effect in word learning illustrate this point.

The flexibility of rational explanation rears its head in other ways as well. At an empirical level, **Uhlmann** reviews evidence that people often change their goals to justify past decisions, a phenomenon that is difficult for any rational model to explain naturally. At a metatheoretical level, Uhlmann notes, "It would be remarkable indeed if scientists were immune to the empirical phenomena we study." Therefore, although rational principles are clearly an important ingredient in explaining cognition, cognitive scientists might be well advised to guard against a tendency to disregard all of the ways and mechanistic reasons that people are irrational.

Despite these dangers of a purely computational framing, the mathematical framework of probabilistic inference does have advantages that are not dependent on specification of psychological mechanism. One important principle is the idea that the brain somehow tracks uncertainty or variability in environmental parameters, rather than just point estimates. This insight has been influential in areas such as causal induction (**Holyoak & Lu**), but it is also not new (e.g., Fried & Holyoak 1984). Another strength of the Bayesian framework is that it offers natural accounts of how information can be combined from multiple sources, and in particular, how people can incorporate rich prior knowledge into any learning task (**Heit & Erickson**). However, this potential is unrealized if there is no independent assessment of what that prior knowledge is. Instead, the expressive flexibility of Bayesian models becomes a weakness, as it makes them unfalsifiable (**Bowers & Davis**; **Danks & Eberhardt**; **Glymour**; **Rogers & Seidenberg**). In some cases, the assumptions of a Bayesian model are demonstrably false, as **Rehder** points out in the case of mutual exclusivity in categorization models, but even then the conclusion is unclear. Was the failed assumption theoretically central to the model, or just an implementation detail of a more general theory that might still hold? If so, what is that general theory that remains after the particular assumptions about the hypothesis space are set aside? Under a computational-level stance, all that is left is the claim of optimality with respect to an unspecified environment, which is no theory at all.

Shifting from issues of representation to the decision process itself, **Danks & Eberhardt** and **Glymour** point out that even the empirical evidence used to support Bayesian models often seriously undermines the claim of Bayesian rationality. Specifically, arguments for Bayesian models often take the form that empirical choice probabilities align with probabilities in the model's posterior distribution. The reasoning seems to be that subjects are choosing in accordance with that posterior and are thus behaving consistently with Bayesian inference. However, a true rational account predicts no such behavior. Instead, subjects should be expected to maximize reward on every individual trial (i.e., to behave deterministically). The standard normative explanation for probability matching – which is endemic in psychology – is based on the need for exploration (e.g., Cohen et al. 2007), but this idea is not formalized in most Bayesian models. More importantly, feedback is independent of the subject's action in many laboratory tasks (e.g., those involving binary choice), which renders exploration irrelevant. Thus, normative ideas about exploration have been extended beyond their domain of applicability, partly because the connection between rational inference and actual choice behavior is not explicitly worked out in most Bayesian models.

Finally, **Al-Shawaf & Buss** and **Pietraszewski & Wertz** point out (echoing many of the points in the target article) that evolutionary psychology, the field that has most thoroughly explored optimality explanations for behavior, has come to a broad conclusion that one must consider mechanism in order for optimality theories to be successful. Explaining behavior from rational perspectives that eschew mechanism is problematic, because behavior is not directly selected but instead arises from selection operating on mechanisms and their interactions with the environment (see target article, sect. 5.3). Likewise, **Anderson** argues that measuring the environment is not always enough because there is still the problem of identifying the natural tasks that shaped evolution. Bayesian inference is a powerful tool for developing ideal observers once the evolutionarily relevant task has been identified, but it provides no help with the identification problem itself.

In summary, when Bayesian models are interpreted on a purely computational level and are held to that position, they turn out to be quite vacuous. Bayesian rationality reduces to the proposal that people act based on probabilities of future events conditioned on past events, with no further psychological implications. The derivation of those probabilities is based on assumptions that are generally unconstrained and untested. Lastly, even when a model is based on correct assumptions about the environment and the learner's goals, global optimality taken alone generally provides an inadequate explanation for behavior.

R4. Bayesian models as mechanistic theories

The alternative to a purely computational-level interpretation of a Bayesian model is to take one or more aspects of the model as corresponding to psychological constructs. In this section, we consider various such stances. We argue that Bayesian models can make useful theoretical contributions under these interpretations, but that those contributions come not from Bayesian inference itself but from other components of the models, which should be treated as more theoretically central than they currently are. This shift of emphasis can go a long way toward clarifying what a Bayesian model actually has to say and how it relates to previous proposals.

An obvious candidate within the Bayesian framework for treatment as a psychological mechanism, and the one most related to the idea of a unified Bayesian theory of cognition, is the belief updating embodied by Bayes' rule itself. As explained in the target article (sect. 3), exact Bayesian inference is equivalent to vote counting, whereby the evidence (technically, log prior probability and log likelihood) for each hypothesis is simply summed over successive independent observations. **Chater et al.** point out that many tasks addressed by Bayesian models require joint posterior distributions to be reduced to marginal distributions over single variables; but this introduces little additional complexity – just an exponential transform (from log posterior, the output of vote counting, to posterior) and then more summation. In most modern models, hypothesis spaces are continuous and hence summation is replaced in the model by integration, but this is an unimportant distinction, especially in a finite physical system. Therefore, the vote-counting interpretation is valid even for the more complex Bayesian models that have arisen in recent years.

Chater et al. go on to argue that much research with Bayesian models posits more complex algorithms than vote counting, for approximating posterior distributions when exact calculation is infeasible. However, most papers that use such algorithms explicitly disavow them as psychological assumptions (e.g., Griffiths et al. 2007). Instead, they are only meant as tools for the modeler to approximate the predictions of the model. More recent work that treats approximation algorithms as psychological processes, takes their deviations from optimal inference as real predictions, and compares alternative algorithms (e.g., Sanborn et al. 2010a) fits squarely into one of the approaches that we advocated as Bayesian Enlightenment in the target article (sect. 6.1).

Borsboom et al. write that the counting rule "seems just about right," and perhaps it is neurologically correct in some cases (e.g., Gold & Shadlen 2001). However, even if this is true, the counting rule is not where the hard work of cognition is being done (**Anderson**). Likewise, although we fully agree with **Chater et al.** that interesting behavior can emerge from simple rules, it is not the counting rule that is responsible for this emergence; it is the structure of the hypothesis space. As **Gopnik** points out, "The central advance has not been Bayes' law itself, but the ability to formulate structured representations, such as causal graphical models, or 'Bayes nets' (Pearl 2000; Spirtes et al. 2000), or hierarchical causal models, category hierarchies or grammars." Thus, as argued above, the hypothesis space is where the interesting psychology lies in most Bayesian models. If we consider it a core assumption of a model, then the model makes meaningful, testable predictions. Although most Bayesian models cast their hypothesis spaces as components of rational analysis and not psychological entities (or else are noncommittal), one can certainly postulate them as psychological representations (**Heit & Erickson**). This is one way in which Bayesian models can potentially make important contributions. Of course, the assumption of optimal inference with respect to the assumed representation could be, and probably often is, wrong (**Uhlmann**), but the important point for present purposes is that this claim becomes testable once the learner's representations and goals are pinned down as psychological commitments.

Therefore, casting assumptions about the hypothesis space, as well as about priors and goals, as psychological claims rather than merely elements of a rational analysis could significantly strengthen the theoretical import of Bayesian models. The problem, as argued above, is that too much Bayesian research is unclear on the intended psychological status of these assumptions (**Bowers & Davis**; **Fernbach & Sloman**). This ambiguity distorts the conclusions that can be drawn from such models. Often the message of a Bayesian model is taken to be that behavior in the domain in question can be explained as optimal probabilistic inference. Instead, the message should be that behavior can be explained as optimal inference, *if* the subject makes certain (often numerous and highly specific) assumptions about the task environment and is trying to optimize a particular function of behavioral outcomes. Logically, the latter is a weaker conclusion, but it is more nuanced and hence theoretically more substantive. The situation would be much less interesting if the correct theory of cognition were, "It's all optimal inference, end of story." Fortunately, that does not appear to be the case, in part because of empirical findings that contradict the predictions of specific rational models (**Baetu, Barberia, Murphy, & Baker** [**Baetu et al.**]; **Danks & Eberhart**; **Glymour**; **Hayes & Newell**; **Jenkins et al.**; **Uhlmann**), but also because optimal inference is not even a full-fledged theory until the learner's goals and background assumptions are specified.

Treating goals, hypotheses, and priors as part of the psychological theory should encourage more consideration of which assumptions of a Bayesian model are important to its predictions and which are implementation details. Recognizing this distinction is just good modeling practice, but it is as important in Bayesian modeling as in other frameworks (**Fernbach & Sloman**). Once this shift of perspective is in place, other questions arise, such as how the learner acquired the structural knowledge of the environment embodied by the proposed hypothesis space (or whether it is innate) and how it compares to knowledge assumed by other theories. Questions of this type are not often addressed in the context of Bayesian models, but taking them into consideration could help the models become much more psychologically complete.

To revisit our example from the target article of Kemp et al.'s (2007) model of second-order generalization in word learning, the model assumes there is potential regularity across named categories in terms of which object dimensions are relevant to defining each category. This

is a critical assumption of the model, in that it drives the model's most important predictions, and without it the model would not reproduce the core phenomenon – the shape bias in children's word learning – that it was developed to explain. Thus, the conclusion to be taken from the model is not that the shape bias is a direct consequence of optimal probabilistic inference, or even that the shape bias is a consequence of optimal inference allowing for overhypotheses across categories, but that the shape bias is consistent with optimal inference if the learner assumes potential regularity across categories in terms of dimension relevance. The question is, therefore, how to regard this last claim. From a strict rationalist perspective, it follows directly from the structure of the environment. This stance is problematic, as already noted, because the relevant property of the environment was not empirically verified in this case.

An alternative position is that the learner's expectation of dimensional regularity across categories is a psychological claim. This perspective takes the model out of the pure computational level and creates a starting point for mechanistic grounding. This move has three advantages: It clarifies what the model does and does not explain, identifies important questions remaining to be answered, and facilitates comparison to other models cast in different frameworks. Regarding the first two of these points, the model demonstrates that second-order generalization emerges from Bayesian inference together with the expectation of dimensional regularity, but many other questions remain, such as: How does the learner know to expect this particular regularity in the environment? How does he or she verify the pattern is present in the input data? (Like most Bayesian models, the model takes p[data | hypothesis] as a starting point, without reference to how this conditional probability is evaluated.) How does the learner produce new responses consistent with what he or she has inferred? These are all natural questions from a mechanistic perspective, and the model would be much stronger if it included answers to them.

As **Jenkins et al.** explain, the structure discovered by Bayesian models of development does not truly develop or emerge. It is built in a priori. All a Bayesian model does is determine which of the patterns or classes of patterns it was endowed with is most consistent with the data it is given. Thus, there is no explanation of where those patterns (i.e., hypotheses) come from. Once one realizes that the structure built into the (over)hypothesis space is at the core of the model's explanation, it is natural to compare those assumptions with the knowledge assumed within other theoretical frameworks (the third advantage listed in the previous paragraph). In the case of models of second-order generalization, such comparisons lead to recognition that the structural knowledge built into Kemp et al.'s (2007) overhypothesis space is essentially the same as that embodied by previous theories based on attention and association learning (Smith et al. 2002). One can then inquire about the source of this knowledge. Whereas the Bayesian model is silent on this question, subsequent work on the attentional model has suggested ways it could emerge from simpler learning processes (Colunga & Smith 2005). Although Colunga and Smith's model may not represent the final answer, it at least attempts to explain what Kemp et al.'s model merely assumes. Thus, taking a mechanistic stance toward Kemp et al.'s model clarifies its contribution but also reveals important questions it fails to address. This is not an unavoidable weakness of the Bayesian approach, but it does suggest that applying more scrutiny to the assumptions of Bayesian models would start them on a path toward providing more complete psychological explanations.

Hayes & Newell offer a similar analysis of J. R. Anderson's (1991b) rational model of categorization. Beyond the several lines of empirical evidence they offer against the rational model, the important point for the present argument is that these issues are not even considered until one pins down the psychological commitments of the model. That the model generates predictions by averaging over hypotheses (instead of using the most likely possibility; cf. Murphy & Ross 2007), that it does not allow for within-cluster feature correlations, and that what it learns is independent of the prediction task it is given (cf. Love 2005), are all strong assumptions. The crucial role of these assumptions can easily be overlooked when they are viewed as merely part of the rational analysis, but if viewed as psychological claims they open up the model to more careful evaluation and further development.

In conclusion, Bayesian models may have significant potential if cast as mechanistic theories. Framing hypothesis spaces as psychological commitments regarding the background knowledge and expectations of the learner seems particularly promising, as it mitigates many of the weaknesses of Bayesian Fundamentalism and opens up the models to the same sort of scientific evaluation used for other approaches. This stance also raises other questions, perhaps most importantly as to where the background expectations (i.e., the environmental structure embodied by the hypothesis space) come from, as well as how that knowledge is represented and how it compares to assumptions of previous theories. These questions have received little attention but could make Bayesian theories much more powerful and complete if answered. In general, Bayesian models have not yet delivered much on the mechanistic level, but we suspect this is due more to their not having been pushed in this direction than to any inherent limitation of the approach.

R5. Prospects for integration

The preceding sections argue that Bayesian models can potentially contribute much to cognitive theory, but they must be tied down to explicit psychological commitments for this potential to be realized. The target article proposed several specific avenues for integration of rational and mechanistic approaches to cognitive modeling, and we are encouraged by the general consensus among commentators that these approaches, which we referred to as Bayesian Enlightenment, embody the proper psychological role of Bayesian models in cognitive science (**Chater et al.**; **Danks & Eberhardt**; **Edelman & Shahbazi**; **Gopnik**; **Herschbach & Bechtel**; **Holyoak & Lu**; **Navarro & Perfors**; **Rehder**). Some research in this vein is already underway, and we hope the present dialogue helps to focus the issues and hasten this transition. Nevertheless, the commentaries raised several challenges, which we address here.

Regarding the general proposal of incorporating rational or computational principles into mechanistic

modeling, **Anderson** argues that computational-level modeling is incoherent, and in fact he questions the very existence of a computational level of analysis on grounds that the brain was not designed top-down. Unlike computer programs, brain function emerged through self-organization. Anderson suggests that the brain does not perform calculations any more than other objects compute their dynamics. We believe this position mischaracterizes computational-level modeling. Just as physical laws of motion are useful for understanding object dynamics, computational theories can be informative about cognitive behavior even if they do not capture the internal workings of the brain (notwithstanding our various other criticisms). The question of whether a level of explanation "exists" in the system being modeled is an ontological red herring in our view, and it has little bearing on whether the explanations are scientifically useful. If certain rational principles can help to explain a wide range of behaviors (e.g., see **Chater et al.**'s example of explaining away), then those principles have contributed to scientific understanding. However, we certainly agree with Anderson that the rational principles must be suitably grounded and constrained, and the additional assumptions needed to explain the data (e.g., regarding goals and hypotheses) must be recognized and scrutinized as well.

Although rational analysis and computational-level modeling are often identified, **Fernbach & Sloman** point out that they are not the same. Rational models explain behavior by appeal to optimality, whereas computational models describe the function of behavior regardless of whether it is optimal. In practice, most rational models are computational because they only consider optimality of behavior, rather than of the behavior together with the system that produces it. However, **Markman & Otto** observe that restricting to behavior alone produces an incomplete definition of rationality, because it ignores factors like time and metabolic cost. Thus, a complete rational account of cognition should take mechanism into account (see target article, sect. 5.3).

Nevertheless, rationality is generally viewed as a property of the cognitive system as a whole (and its interaction with the environment), whereas mechanistic modeling involves iteratively decomposing phenomena into components and showing how the components interact to produce the whole (**Herschbach & Bechtel**). This contrast raises the question of how rational and mechanistic explanations can be integrated. The solutions Herschbach & Bechtel offer align well with our proposals and generally fall into two categories. First, one can consider optimality of one aspect of the cognitive system with respect to knowledge or constraints provided by other components. This approach aligns well with our call for treating Bayesian hypotheses as assumptions about the learner's knowledge, rather than as products of rational analysis. It also fits with our proposal in the target article (sect. 6.2) for bringing rational analysis inside mechanistic models, in order to derive optimal behavior of one process in the context of the rest of the model (e.g., Shiffrin & Steyvers 1998; Wilder et al. 2009).

Second, one can study algorithms that approximate optimal inference (e.g., Daw & Courville 2007; Sanborn et al. 2010a). Under this approach, rational and mechanistic considerations enter at different levels of analysis, and the aim is to understand how they constrain each other. **Bowers & Davis** and **Herschbach & Bechtel** question this approach, arguing that it is no more effective than mechanistic modeling alone (see also the discussion of bounded rationality in the target article, sect. 5.4). In the end, a mechanistic model is evaluated only by how well it matches the data, not by how well it approximates some rational model of the data. However, rational considerations can still play an important role by constraining the search for mechanistic explanations. Understanding the function a mechanism serves should help guide hypotheses about how it works. When phenomena in different domains can be linked by a common rational explanation, this can suggest a common underlying mechanism. Also, understanding the relationship between a mechanistic model and a rational analysis, in terms of both how the model implements and how it deviates from the optimal solution, can help to identify which aspects of the model are necessary for its predictions. This approach can mitigate the tendency **Norris** warns of for modelers to ascribe psychological reality to superfluous mechanisms not entailed by the data. In these ways, rational considerations can provide principled constraints on development of mechanistic models. As **Danks & Eberhardt** argue, integration of rational and mechanistic models should not amount to reduction of the former to the latter, because such an approach would relinquish the explanatory benefits of the computational level. Instead, rational explanations should "pull up" mechanistic ones, in order to explain why one algorithm or implementation is more appropriate than another for a given task. Nevertheless, questions remain of how somewhat subjective notions of appropriateness should be incorporated into model selection.

Because the rationality metaphor is based on a mathematical ideal and has no physical target, it is compatible with essentially any mechanism (target article, sects. 2.2 and 6.2). Thus, incorporating rational principles is potentially fruitful within any mechanistic modeling framework. For example, **Barsalou** suggests connecting the Bayesian framework to the perceptuomotor simulation mechanisms proposed in theories of grounded cognition. Such an approach could fulfill our call for grounding Bayesian hypotheses in the learner's knowledge in an especially concrete way. Although we believe there is much work to do before theories of grounded cognition can be given a rigorous Bayesian interpretation, it is encouraging to see people thinking in this direction. Based on the previous considerations, one important goal in this line of research would be to understand not just how Bayesian inference can be implemented by simulation mechanisms, but what the implications are of this rational interpretation for the details of how these simulation mechanisms should operate.

Concerning the opposite connection, of mechanistic implications for rational analysis, **Chater et al.** claim that studying cognition at the algorithmic level cannot provide insight into the computational level (e.g., into the purpose the algorithm). On the contrary, investigating how cognitive mechanisms deviate from rational predictions can inform both what the function of the system is and how it is carried out. For example, the experimental results and accompanying modeling of Sakamoto et al. (2008) indicate that categories in their task are psychologically represented in terms of central tendency and variability (implemented in their model as mean and

variance), and that the goal of learning is to estimate these statistics for use in classifying new items. The novel sequential effect predicted by the model and confirmed in the experiments arises due to cue competition effects from learning these two statistics from joint prediction error (Rescorla & Wagner 1972). Thus, the explanation of the experimental results requires inference of the computational goals of the learning system (i.e., the statistics to be estimated) as well as of how those goals are implemented.

Clarity on the status of model assumptions is as important for mechanistic models as we have argued it is for rational models (**Norris**). Norris uses the mechanistic model of Sakamoto et al. (2008) to question whether we advocate going too far in reifying mechanism for its own sake. However, he acknowledges the Sakamoto et al. model does not suffer this problem and praises its intermediate level of abstractness. Indeed, our position is that it would be pointless to commit to excess detail that does not contribute to a model's predictions. The model in that study proposes that category means and variances are learned through joint error correction, because this mechanism is responsible for the model's primary prediction. The model makes no commitments about how the computations behind the update rule are carried out, because those details have no bearing on that prediction (although they could be relevant for explaining other data). **Navarro & Perfors** also criticize this model, suggesting it gives no consideration to computational-level issues. However, a primary principle of the model concerns what environmental (i.e., category) statistics people track, and the update rule used to learn them has well-understood computational connections to least-squares estimation. Navarro & Perfors go on to claim that the purely rational model considered by Sakamoto et al. is mis-specified for the task, but this comment leads back to one of the core weaknesses of rational analysis, that it depends on the learner's assumptions about the environment. The rational model in question is indeed optimal for a certain class of environments, and it is closely related to a rational model of a similar task proposed by Elliott and Anderson (1995). There is certainly a Bayesian model that will reproduce Sakamoto et al.'s findings, based on the right choice of generative model for the task, but this is not informative without a theory of where those assumptions come from, or else of the mechanisms from which they implicitly emerge. Such a theory is not forthcoming from a fundamentalist approach, but is it possible from enlightened approaches that consider mechanism and rational principles jointly.

Finally, several commentators argue that integrative research is not possible before technical frameworks have been developed. **Navarro & Perfors** and **Edelman & Shahbazi** argue that much previous fundamentalist research has paved the way for work that gives real consideration to the processes and representations underlying Bayesian models. Likewise, **Sewell et al.** suggest that individual work focusing on one framework or level of analysis is useful because the field as a whole implements a division of labor that leads to integration. We generally agree with this assessment, provided the integrative work gets done. The important point is that fundamentalist research cannot be the end goal, because it offers little theoretical contribution on its own. Nearly all scientific methods undergo initial technical development before they can be used to advance theory, but the two should not be confused. Thus, once again, the conclusion is that it is critical to carefully consider the contribution and commitments of any model, so that one can discriminate advances in theoretical understanding from prerequisite technical advances.

R6. Conclusions

Bayesian methods have advanced rapidly in recent years, offering the hope that they may help answer some of the more difficult problems in cognitive science. As **Lee** eloquently states (see also **Edelman & Shahbazi**), Bayesian inference offers a "coherent solution to the problem of drawing inferences over structured models from sparse and noisy data. That seems like a central challenge faced by the mind, and so it is not surprising the metaphor has led to insightful models of human cognition." However, in most cases, more clarity is needed on just what those insights are.

Much of the current confusion arises from ambiguity in the levels of analysis at which Bayesian models are intended. The standard position from rational analysis (J. R. Anderson 1990) is that a rational model is based purely on the environment and makes no reference to psychological constructs. Many Bayesian writings, including some of the present commentaries (**Borsboom et al.**; **Chater et al.**; **Fernbach & Sloman**; **Gopnik**), endorse this position while simultaneously arguing that processes or representations within Bayesian models should be regarded as psychological entities. The danger with this sort of inconsistency is that Bayesian models might appear to say much more than they actually do, because researchers can attribute rich psychological assumptions to their models but be free to disavow them as merely computational when they are contradicted by data.

Pinning down the theoretical status of Bayesian models would help clarify their core assumptions and predictions, thus making it easier to evaluate their scientific contribution. As we have argued, when Bayesian models are held to the computational level, they are largely vacuous. This position, which we have labeled Bayesian Fundamentalism, amounts to the claim that people act according to probabilities of future events based on past events, usually without any validation of what those probabilities actually are. More promising is the approach we have labeled Bayesian Enlightenment, which involves treating some or all of a model's components as psychological constructs. This approach fits well with **Rehder**'s proposal to drop the "rational" label and adopt the term "probabilistic model." Probabilistic models still naturally incorporate rational principles, but emphasizing the psychological realization of these principles shifts attention to other important issues, such as the source of and justification for the prior knowledge built into the hypothesis space, which assumptions are critical to model predictions, and how they compare to other proposals. Pinning down the psychological commitments of Bayesian models in this way clarifies what they do and do not explain and enables them to be developed into more complete psychological theories.

Rogers & Seidenberg note that connectionism had problems of underconstraint similar to those noted here for

Bayesian models, but that connectionism has since become far more productive by grounding in neuroscience. Likewise, **Sewell et al.** argue that the setbacks for connectionism, Behaviorism, and evolutionary psychology discussed in our target article all led to eventual important progress as a result of addressing noted shortcomings. We believe the present critique has the potential to have a similar positive effect, and like these commentators, we predict Bayesian modeling will follow a similar path of maturation and integration into the rest of cognitive science.

References

[The letters "a" and "r" before author's initials stand for target article and response references, respectively]

Albert, M. (2001) Bayesian learning and expectations formation: Anything goes. In: *Foundations of Bayesianism*, ed. D. Corfield & J. Williamson, pp. 341–62. Kluwer. [DB]

Ali, N., Chater, N. & Oaksford, M. (2011) The mental representation of causal conditional reasoning: Mental models or causal models. *Cognition* 119:403–18. [NC]

Ali, N., Schlottmann, A., Shaw, C., Chater, N. & Oaksford, M. (2010) Conditionals and causal discounting in children. In: *Cognition and conditionals: Probability and logic in human thinking*, ed. M. Oaksford & N. Chater, pp. 117–34. Oxford University Press. [NC]

Anderson, J. R. (1978) Arguments concerning representations for mental imagery. *Psychological Review* 85:249–77. [EH]

Anderson, J. R. (1990) *The adaptive character of thought*. Erlbaum. [NC, PMF, arMJ, ABM, DN]

Anderson, J. R. (1991a) Is human cognition adaptive? *Behavioral and Brain Sciences* 14:471–517. [NC]

Anderson, J. R. (1991b) The adaptive nature of human categorization. *Psychological Review* 98:409–29. [BKH, arMJ, DKS]

Anderson, J. R., Bothell, D., Lebiere, C. & Matessa, M. (1998) An integrated theory of list memory. *Journal of Memory and Language* 38:341–80. [aMJ]

Anderson, J. R. & Schooler, L. J. (1991) Reflections of the environment in memory. *Psychological Science* 2:396–408. [aMJ]

Andrews, M., Vigliocco, G. & Vinson, D. (2009) Integrating experiential and distributional data to learn semantic representations. *Psychological Review* 116:463–98. [LWB]

Armor, D. A. (1999) *The illusion of objectivity: Bias in the belief in freedom from bias*. Doctoral dissertation, University of California, Los Angeles. [ELU]

Austerweil, J. & Griffiths, T. L. (2008) Analyzing human feature learning as nonparametric Bayesian inference. *Advances in Neural Information Processing Systems* 21:97–104. [aMJ]

Baetu, I. & Baker, A. G. (2009) Human judgments of positive and negative causal chains. *Journal of Experimental Psychology: Animal Behavior Processes* 35:153–68. [IB]

Baker, A. G., Murphy, R. A. & Vallee-Tourangeau, F. (1996) Associative and normative models of causal induction: Reacting to versus understanding cause. In: *The psychology of learning and motivation, vol. 34*, ed. D. R. Shanks, K. J. Holyoak & D. L. Medin., pp. 1–45. Academic Press. [IB]

Baker, C. L., Saxe, R. & Tenenbaum, J. B. (2009) Action understanding as inverse planning. *Cognition* 113:329–49. [aMJ]

Baldwin, J. D. & Baldwin, J. I. (1977) The role of learning phenomena in the ontogeny of exploration and play. In: *Primate bio-social development: Biological, social and ecological determinants*, ed. S. Chevalier-Skolnikoff & F. E. Poirer, pp. 343–406. Garland. [aMJ]

Barrett, H. C. & Kurzban, R. (2006) Modularity in cognition: Framing the debate. *Psychological Review* 113:628–47. [DKS]

Barsalou, L. W. (1985) Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 11:629–54. [LWB]

Barsalou, L. W. (1987) The instability of graded structure: Implications for the nature of concepts. In: *Concepts and conceptual development: Ecological and intellectual factors in categorization*, ed. U. Neisser, pp. 101–40. Cambridge University Press. [LWB]

Barsalou, L. W. (1990) On the indistinguishability of exemplar memory and abstraction in memory representation. In: *Advances in social cognition*, ed. T. K. Srull & R. S. Wyer, pp. 61–88. Erlbaum. [EH]

Barsalou, L. W. (1999) Perceptual symbol systems. *Behavioral and Brain Sciences* 22:577–660. [LWB]

Barsalou, L. W. (2003a) Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London: Biological Sciences* 358:1177–87. [LWB]

Barsalou, L. W. (2003b) Situated simulation in the human conceptual system. *Language and Cognitive Processes* 18:513–62. [LWB]

Barsalou, L. W. (2005) Continuity of the conceptual system across species. *Trends in Cognitive Sciences* 9:309–11. [LWB]

Barsalou, L. W. (2008a) Grounded cognition. *Annual Review of Psychology* 59:617–45. [LWB]

Barsalou, L. W. (2008b) Grounding symbolic operations in the brain's modal systems. In: *Embodied grounding: Social, cognitive, affective, and neuroscientific approaches*, ed. G. R. Semin & E. R. Smith, pp. 9–42. Cambridge University Press. [LWB]

Barsalou, L. W. (2008c) Situating concepts. In: *Cambridge handbook of situated cognition*, ed. P. Robbins & M. Aydede, pp. 236–63. Cambridge University Press. [LWB]

Barsalou, L. W. (2009) Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society of London: Biological Sciences* 364:1281–89. [LWB]

Barsalou, L. W., Barbey, A. K., Simmons, W. K. & Santos, A. (2005) Embodiment in religious knowledge. *Journal of Cognition and Culture* 5:14–57. [LWB]

Barsalou, L. W., Breazeal, C. & Smith, L. B. (2007) Cognition as coordinated non-cognition. *Cognitive Processing* 8:79–91. [LWB]

Barsalou, L. W., Niedenthal, P. M., Barbey, A. & Ruppert, J. (2003) Social embodiment. In: *The psychology of learning and motivation, vol. 43*, ed. B. Ross, pp. 43–92. Academic Press. [LWB]

Barsalou, L. W., Santos, A., Simmons, W. K. & Wilson, C. D. (2008) Language and simulation in conceptual processing. In: *Symbols, embodiment, and meaning*, ed. M. De Vega, A. M. Glenberg & A. C. Graesser, pp. 245–83. Oxford University Press. [LWB]

Bar-Tal, D. (2002) From intractable conflict through conflict resolution to reconciliation: Psychological analysis. *Political Psychology* 21(2):351–65. [PMF]

Bartlett, F. C. (1932) *Remembering: A study in experimental and social psychology*. Cambridge University Press. [EH]

Bastardi, A., Uhlmann, E. L. & Ross, L. (2011) Wishful thinking: Belief, desire, and the motivated evaluation of scientific evidence. *Psychological Science* 22:731–32. [ELU]

Bechtel, W. (2008) *Mental mechanisms*. Routledge. [MH]

Bechtel, W. & Abrahamsen, A. (2005) Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36:421–41. [MH]

Bechtel, W. & Abrahamsen, A. (2010) Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science. *Studies in History and Philosophy of Science Part A* 41:321–33. [MH]

Bechtel, W. & Richardson, R. C. (1993/2010) *Discovering complexity: Decomposition and localization as strategies in scientific research*. 1993 edition, Princeton University Press; 2010 edition, MIT Press. [MH]

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E. & Pouget, A. (2008) Probabilistic population codes for Bayesian decision making. *Neuron* 60:1142–52. [MH, aMJ]

Binmore, K. (2009) *Rational decisions*. Princeton University Press. [aMJ]

Bishop, C. M. (1996) *Neural networks for pattern recognition*. Oxford University Press. [NC]

Bishop, C. M. (2006) *Pattern recognition and machine learning*. Springer. [SE]

Bjorklund, D. F. & Pellegrini, A. D. (2000) Child development and evolutionary psychology. *Child Development* 71:1607–708. [aMJ]

Bonawitz, E. B., Ferranti, D., Saxe, R., Gopnik, A., Meltzoff, A. N., Woodward, J. & Schulz, L. E. (2010) Just do it? Investigating the gap between prediction and action in toddlers' causal inferences. *Cognition* 115(1):104–17. [AG]

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S. & Cohen, J. D. (2001) Conflict monitoring and cognitive control. *Psychological Review* 108(3):624–52. [TTR]

Botvinick, M. M. & Plaut, D. C. (2004) Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review* 111(2):395–429. [TTR]

Boucher, L., Palmeri, T. J., Logan, G. D. & Schall, J. D. (2007) Inhibitory control in mind and brain: An interactive race model of countermanding saccades. *Psychological Review* 114:376–97. [aMJ]

Bowers, J. S. & Davis, C. J. (submitted) Bayesian just-so stories in cognitive psychology and neuroscience. [JSB]

Bowlby, J. (1969) *Attachment and loss, vol. 1: Attachment*. Basic Books. [aMJ]

Bracha, H. S. (2004) Freeze, flight, fight, fright, faint: Adaptationist perspectives on the acute stress response spectrum. *CNS Spectrums* 9:679–85. [LA-S]

Brighton, H. & Gigerenzer, G. (2008) Bayesian brains and cognitive mechanisms: Harmony or dissonance? In: *Bayesian rationality: The probabilistic approach*

to human reasoning, ed. M. Oaksford & N. Chater, pp. 189–208. Oxford University Press. [aMJ]

Brown, S. D. & Steyvers, M. (2009) Detecting and predicting changes. *Cognitive Psychology* 58:49–67. [aMJ]

Brown, S. D., Wagenmakers, E.-J. & Steyvers, M. (2009) Observing evidence accumulation during multi-alternative decisions. *Journal of Mathematical Psychology* 53:453–62. [DB]

Buchsbaum, D., Griffiths, T. L. & Gopnik, A. (in press) Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*. [AG]

Buller, D. J. (2005) *Adapting minds: Evolutionary psychology and the persistent quest for human nature*. MIT Press. [aMJ]

Burgess, N. & Hitch, G. J. (1999) Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review* 106:551–81. [aMJ]

Busemeyer, J. R. & Johnson, J. G. (2008) Microprocess models of decision making. In: *Cambridge handbook of computational psychology*, ed. R. Sun, pp. 302–21. Cambridge University Press. [aMJ]

Buss, D. M. (1994) *The evolution of desire: Strategies of human mating*. Basic Books. [aMJ]

Buss, D. M. (2011) *Evolutionary psychology: The new science of the mind*, 4th edition. Allyn & Bacon. [LA-S]

Buss, D. M., Haselton, M. G., Shackelford, T. K., Bleske, A. L. & Wakefield, J. C. (1998) Adaptations, exaptations, and spandrels. *American Psychologist* 53:533–48. [LA-S, aMJ]

Caramazza, A. & Shelton, J. R. (1998) Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience* 10:1–34. [aMJ]

Card, S., Moran, T., & Newell, A. (1983) *The psychology of human-computer interaction*. Erlbaum. [BR]

Chater, N. & Manning, C. D. (2006) Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences* 10:335–44. [aMJ]

Chater, N. & Oaksford, M. (1990) Autonomy, implementation and cognitive architecture: A reply to Fodor and Pylyshyn. *Cognition* 34:93–107. [NC]

Chater, N. & Oaksford, M. (1999) The probability heuristics model of syllogistic reasoning. *Cognitive Psychology* 38:191–258. [aMJ]

Chater, N. & Oaksford, M. (2008) The probabilistic mind: Prospects for a Bayesian cognitive science. In: *The probabilistic mind: Prospects for rational models of cognition*, ed. M. Oaksford & N. Chater, pp. 3–31. Oxford University Press. [aMJ]

Chater, N., Oaksford, M., Nakisa, R. & Redington, M. (2003) Fast, frugal, and rational: How rational norms explain behavior. *Organizational Behavior and Human Decision Processes* 90:63–86. [NC, aMJ]

Chater, N. & Vitányi, P. (2007) "Ideal learning" of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology* 51:135–63. [NC]

Chater, N., Reali, F. & Christiansen, M. H. (2009) Restrictions on biological adaptation in language evolution. *Proceedings of the National Academy of Sciences USA* 106:1015–20. [aMJ]

Chater, N., Tenenbaum, J. & Yuille, A. (2006) Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences* 10(7):287–91. [SE, aMJ]

Cheng, P. W. (1997) From covariation to causation: A causal power theory. *Psychological Review* 104:367–405. [KJH]

Cheng, P. W. & Holyoak, K. J. (1995) Complex adaptive systems as intuitive statisticians: Causality, contingency, and prediction. In: *Comparative approaches to cognitive science*, ed. H. L. Roitblat & J.-A. Meyer, pp. 271–302. MIT Press. [KJH]

Cherniak, C. (1986) *Minimal rationality*. MIT Press. [ABM]

Chomsky, N. (1959) A review of B. F. Skinner's *Verbal Behavior. Language* 35:26–58. [aMJ]

Chomsky, N. (1965) *Aspects of the theory of syntax*. MIT Press. [TTR]

Clark, D. D. & Sokoloff, L. (1999) Circulation and energy metabolism in the brain. In: *Basic neurochemistry: Molecular, cellular and medical aspects*, ed. G. J. Siegel, B. W. Agranoff, R. W. Albers, S. K. Fisher & M. D. Uhler, pp. 637–70. Lippincott-Raven. [aMJ]

Clearfield, M. W., Dineva, E., Smith, L. B., Diedrich, F. J. & Thelen, E. (2009) Cue salience and infant perseverative reaching: Tests of the dynamic field theory. *Developmental Science* 12:26–40. [aMJ]

Cohen, G. L., Aronson, J. & Steele, C. M. (2000) When beliefs yield to evidence: Reducing biased evaluation by affirming the self. *Personality and Social Psychology Bulletin* 26:1151–64. [ELU]

Cohen, J. D., McClure, S. M. & Yu, A. J. (2007) Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* 362:933–42. [rMJ]

Colunga, E. & Smith, L. (2005) From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review* 112(2):347–82. [arMJ, DJN]

Conati, C., Gertner, A., VanLehn, K. & Druzdzel, M. (1997) On-line student modeling for coached problem solving using Bayesian networks. In: *User modeling: Proceedings of the Sixth International Conference, UM97, Berlin, 1997*, pp. 231–42, ed. A. Jameson, C. Paris & C. Tasso. Springer. [aMJ]

Confer, J. C., Easton, J. A., Fleischman, D. S., Goetz, C. D., Lewis, D. M. G., Perilloux, C. & Buss, D. M. (2010) Evolutionary psychology: Controversies, questions, prospects, and limitations. *American Psychologist* 65:110–26. [LA-S]

Copernicus, N. (1995) *On the Revolutions of the Heavenly Spheres*, p. 3. Prometheus. [CG]

Cosmides, L. & Tooby, J. (1987) From evolution to behavior: Evolutionary psychology as the missing link. In: *The latest on the best: Essays on evolution and optimality*, ed. J. Dupre, pp. 277–306. MIT Press. [DP]

Cosmides, L. & Tooby, J. (1992) Cognitive adaptations for social exchange. In: *The adapted mind: Evolutionary psychology and the generation of culture*, ed. J. Barkow, L. Cosmides & J. Tooby. pp. 163–228. Oxford University Press. [aMJ]

Cosmides, L. & Tooby, J. (1994) Origins of domain specificity: The evolution of functional organization. In: *Mapping the mind: Domain specificity in cognition and culture*, ed. L. A. Hirschfeld & S. Gelman, pp. 85–116. Cambridge University Press. [ELU]

Courville, A. C., Daw, N. D. & Touretzky, D. S. (2006) Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences* 10:294–300. [NC]

Cowell, R. A., Bussey, T. J. & Saksida, L. M. (2006) Why does brain damage impair memory? A connectionist model of object recognition memory in perirhinal cortex. *Journal of Neuroscience* 26:12186–97. [IB]

Craver, C. F. (2007) *Explaining the brain: What a science of the mind-brain could be*. Oxford University Press. [MH]

Cree, G. S. & McRae, K. (2003) Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General* 132:163–201. [aMJ]

Crick, F. (1989) The recent excitement about neural networks. *Nature* 337:129–32. [aMJ]

Czerlinski, J., Gigerenzer, G. & Goldstein, D. G. (1999) How good are simple heuristics? In: *Simple heuristics that make us smart*, ed. G. Gigerenzer & P. M. Todd, pp. 97–118. Oxford University Press. [aMJ]

Danks, D. (2003) Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology* 47(2):109–21. [AG]

Danks, D. (2008) Rational analyses, instrumentalism, and implementations. In: *The probabilistic mind: Prospects for Bayesian cognitive science*, ed. M. Oaksford & N. Chater, pp. 59–75. Oxford University Press. [aMJ]

Daugman, J. G. (2001) Brain metaphor and brain theory. In: *Philosophy and the neurosciences: A reader*, ed. W. Bechtel, P. Mandik, J. Mundale & R. S. Stufflebeam, pp. 23–36. Blackwell. [aMJ]

Davis, T. & Love, B. C. (2010) Memory for category information is idealized through contrast with competing options. *Psychological Science* 21:234–42. [aMJ]

Daw, N. & Courville, A. (2007) The pigeon as particle filter. *Advances in Neural Information Processing Systems* 20:1528–35. [arMJ]

Daw, N. D., Courville, A. C. & Dayan, P. (2008) Semi-rational models: The case of trial order. In: *The probabilistic mind: Prospects for rational models of cognition*, ed. M. Oaksford & N. Chater, pp. 431–52. Oxford University Press. [KJH, aMJ]

Daw, N. D., Niv, Y. & Dayan, P. (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience* 8:1704–11. [aMJ]

Dawes, R. M. & Corrigan, B. (1974) Linear models in decision making. *Psychological Bulletin* 81:95–106. [aMJ]

Dawkins, R. (1982) *The extended phenotype*. W. H. Freeman. [LA-S]

Dawkins, R. (1987) *The blind watchmaker*. W. W. Norton. [aMJ]

Denève, S. (2008) Bayesian spiking neurons. I: Inference. *Neural Computation* 20:91–117. [aMJ]

Denève, S., Latham, P. E. & Pouget, A. (1999) Reading population codes: A neural implementation of ideal observers. *Nature Neuroscience* 2:740–45. [aMJ]

Dennett, D. C. (1987) *The intentional stance*. MIT Press. [DKS]

Dennis, S. & Humphreys, M. S. (1998) Cuing for context: An alternative to global matching models of recognition memory. In: *Rational models of cognition*, ed. M. Oaksford & N. Chater, pp. 109–27. Oxford University Press. [aMJ]

Devlin, J. T., Gonnerman, L. M., Andersen, E. S. & Seidenberg, M. S. (1998) Category-specific semantic deficits in focal and widespread brain damage: A computational account. *Journal of Cognitive Neuroscience* 10:77–94. [aMJ]

Dickinson, A. M. (2000) The historical roots of organizational behavior management in the private sector: The 1950s–1980s. *Journal of Organizational Behavior Management* 20(3/4): 9–58. [aMJ]

Doll, B. B., Jacobs, W. J., Sanfey, A. G. & Frank, M. J. (2009) Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research* 1299:74–94. [aMJ]

Doucet, A., Godsill, S. & Andrieu, C. (2000) On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing* 10:197–208. [aMJ]

Dube, C., Rotello, C. & Heit, E. (2010) Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review* 117:831–63. [EH]

Dunbar, K. (1995) How scientists really reason: Scientific reasoning in real-world laboratories. In: *Mechanisms of insight*, ed. R. J. Sternberg & J. Davidson, pp. 365–95. MIT Press. [aMJ]

Dunning, D. & Cohen, G. L. (1992) Egocentric definitions of traits and abilities in social judgment. *Journal of Personality and Social Psychology* 63:341–55. [ELU]

Dunning, D., Leuenberger, A. & Sherman, D. A. (1995) A new look at motivated inference: Are self-serving theories of success a product of motivational forces? *Journal of Personality and Social Psychology* 69:58–68. [ELU]

Dupre, J. (1981) Natural kinds and biological taxa. *Philosophical Review* 90:66–90. [BR]

Dyson, F. W., Eddington, A. S. & Davidson, C. (1920) A determination of the deflection of light by the sun's gravitational field, from observations made at the total eclipse of May 29, 1919. *Philosophical Transactions of the Royal Society of London, Series A: Mathematical, Physical and Engineering Sciences* 220:291–333. [aMJ]

Eberhardt, F. & Danks, D. (2011) Confirmation in the cognitive sciences: The problematic case of Bayesian models. *Minds and Machines* 21(3):389–410. [DD, CG]

Edelman, S. (2008a) A swan, and pike, and a crawfish walk into a bar. *Journal of Experimental and Theoretical Artificial Intelligence* 20:261–68. [SE]

Edelman, S. (2008b) *Computing the mind: How the mind really works*. Oxford University Press. [SE]

Edelman, S. (2008c) On the nature of minds, or: Truth and consequences. *Journal of Experimental and Theoretical Artificial Intelligence* 20:181–96. [SE]

Ein-Dor, T., Mikulincer, M., Doron, G. & Shaver, P. R. (2010) The attachment paradox: How can so many of us (the insecure ones) have no adaptive advantages? *Perspectives on Psychological Science* 5(2):123–41. [aMJ]

Einstein, A. (1916) Die Grundlage der allgemeinen Relativitätstheorie [The foundation of the generalized theory of relativity]. *Annalen der Physik* 354(7):769–822. [aMJ]

Elliott, S. W. & Anderson, J. R. (1995) Effect of memory decay on predictions from changing categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21:815–36. [rMJ]

Elman, J. L. (1990) Finding structure in time. *Cognitive Science* 14:179–211. [aMJ]

Elman, J. L. (1993) Learning and development in neural networks: The importance of starting small. *Cognition* 48:71–99. [aMJ]

Engelfriet, J. & Rozenberg, G. (1997) Node replacement graph grammars. In: *Handbook of graph grammars and computing by graph transformation, vol. 1*, ed. G. Rozenberg, pp. 1–94. World Scientific. [aMJ]

Epstein, S., Lipson, A., Holstein, C. & Huh, E. (1992) Irrational reactions to negative outcomes: Evidence for two conceptual systems. *Journal of Personality and Social Psychology* 62:328–39. [ELU]

Estes, W. K. (1957) Theory of learning with constant, variable, or contingent probabilities of reinforcement. *Psychometrika* 22:113–32. [aMJ]

Feldman, J. A. (2010) Cognitive science should be unified: Comment on Griffiths et al. and McClelland et al. *Trends in Cognitive Sciences* 14:341. [DKS]

Feldman, J. A. & Ballard, D. H. (1982) Connectionist models and their properties. *Cognitive Science* 6:205–54. [NC]

Fitelson, B. (1999) The plurality of Bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science* 66:362–78. [aMJ]

Fodor, J. A. (1983) *The modularity of mind*. MIT Press. [DKS]

Fodor, J. A. & Pylyshyn, Z. (1988) Connectionism and cognitive architecture: A critical analysis. *Cognition* 28: 3–71. [aMJ]

Foraker, S., Regier, T., Khetarpal, N., Perfors, A. & Tenenbaum, J. B. (2009) Indirect evidence and the poverty of the stimulus: The case of anaphoric one. *Cognitive Science* 33:287–300. [NC]

Frank, M. J., Seeberger, L. & O'Reilly, R. C. (2004) By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. *Science* 306:1940–43. [aMJ]

Fried, L. S. & Holyoak, K. J. (1984) Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10:234–57. [rMJ]

Gabbay, D., Hogger, C. & Robinson, J., eds. (1994) *Handbook of logic in artificial intelligence and logic programming, vol. 3: Nonmonotonic reasoning and uncertain reasoning*. Oxford University Press. [aMJ]

Garey, M. R. & Johnson, D. S. (1979) *Computers and intractability: A guide to the theory of NP-completeness*. W. H. Freeman. [NC]

Geisler, W. S. (1989) Sequential ideal-observer analysis of visual discriminations. *Psychological Review* 96:267–314. [ABM]

Geisler, W. S. & Diehl, R. L. (2003) A Bayesian approach to the evolution of perceptual and cognitive systems. *Cognitive Science* 27:379–402. [aMJ]

Geisler, W. S., Perry, J. S., Super, B. J. & Gallogly, D. P. (2001) Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research* 41:711–24. [aMJ]

Geisler, W. S. & Ringach, D. (2009) Natural systems analysis. *Visual Neuroscience* 26:1–3. [BLA]

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2003) *Bayesian data analysis*, 2nd edition. Chapman and Hall. [NC]

Geman, S., Bienenstock, E. & Doursat, R. (1992) Neural networks and the bias/variance dilemma. *Neural Computation* 4:1–58. [aMJ]

Geman, S. & Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6:721–41. [aMJ]

Genesereth, M. R. & Nilsson, N. J. (1987) *Logical foundations of artificial intelligence*. Morgan Kaufman. [BR]

Gentner, D. (1983) Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7:155–70. [aMJ]

Gentner, D., Brem, S., Ferguson, R. W., Markman, A. B., Levidow, B. B., Wolff, P. & Forbus, K. D. (1997) Analogical reasoning and conceptual change: A case study of Johannes Kepler. *Journal of the Learning Sciences* 6(1):3–40. [aMJ]

Gibson, J. J. (1957) Survival in a world of probable objects. *Contemporary Psychology* 2:33–35. [SE]

Gibson, J. J. (1979) *The ecological approach to visual perception*. Houghton Mifflin. [arMJ]

Gigerenzer, G. & Brighton, H. (2009) Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science* 1:107–43. [arMJ]

Gigerenzer, G. & Todd, P. M. (1999) *Simple heuristics that make us smart*. Oxford University Press. [MH, aMJ]

Glimcher, P. W., Camerer, C., Poldrack, R. A. & Fehr, E. (2008) *Neuroeconomics: Decision making and the brain*. Academic Press. [SE]

Glymour, C. (2007) Bayesian Ptolemaic psychology. In: *Probability and inference: Essays in Honor of Henry E. Kyburg, Jr.*, ed. W. Harper & G. Wheeler, pp. 123–41. Kings College Publishers. [CG]

Goel, V. (2007) Anatomy of deductive reasoning. *Trends in Cognitive Sciences* 11:435–441. [EH]

Gold, J. I. & Shadlen, M. N. (2001) Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences* 5:10–16. [arMJ]

Goldstone, R. (1994) An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, and Computers* 26:381–86. [GWJ]

Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., Schulz, L. E. & Tenenbaum, J. B. (2006) Intuitive theories of mind: A rational approach to false belief. In: *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society, Vancouver, Canada*, ed. R. Sun, pp. 1382–87. Cognitive Science Society. [aMJ]

Goodman, N. D., Mansinghka, V. K. & Tenenbaum, J. B. (2007) Learning grounded causal models. In: *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, ed. D. S. McNamara & G. Trafton, pp. 305–10. Erlbaum. [NC]

Goodman, N. D., Mansinghka, V. K., Roy, D., Bonawitz, K. & Tenenbaum, J. B. (2008a) Church: A language for generative models. In: *Proceedings of the Twenty-Fourth Annual Conference on Uncertainty in Artificial Intelligence (UAI), July 9–12, 2008, Helsinki, Finland*, ed. D. McAllester & P. Myllymaki, pp. 220–29. AUAI Press. [NC]

Goodman, N. D., Tenenbaum, J. B., Feldman, J. & Griffiths, T. L. (2008b) A rational analysis of rule-based concept learning. *Cognitive Science* 32(1):108–54. [NC, aMJ]

Goodman, N. D., Ullman, T. D. & Tenenbaum, J. B. (2011) Learning a theory of causality. *Psychological Review* 118:110–19. [NC]

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T. & Danks, D. (2004) A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review* 111(1):3–32. [NC, AG]

Gopnik, A. & Schulz, L., eds. (2007) *Causal learning: Psychology, philosophy, and computation*. Oxford University Press. [AG]

Gottlieb, G. (1992) *Individual development and evolution: The genesis of novel behavior*. Oxford University Press. [aMJ]

Gould, S. J. & Lewontin, R. (1979) The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London Series B: Biological Sciences* 205:581–98. [LA-S, aMJ]

Green, D. M. & Swets, J. A. (1966) *Signal detection theory and psychophysics*. John Wiley. [aMJ]

Griffiths, O., Hayes, B. K., Newell, B. & Papadopoulos, C. (in press) Where to look first for an explanation of induction with uncertain categories. *Psychonomic Bulletin and Review*. [BKH]

Griffiths, T. L. & Ghahramani, Z. (2006) Infinite latent feature models and the Indian buffet process. In: *Advances in neural information processing systems, vol. 18*, ed. J. Weiss, B. Schölkopf & J. Platt, pp. 475–82. MIT Press. [NC, aMJ]

Griffiths, T. L. & Kalish, M. L. (2007) Language evolution by iterated learning with Bayesian agents. *Cognitive Science* 31:441–80. [DKS]

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A. & Tenenbaum, J. (2010) Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences* 14(8):357–64. [NC, AG, DKS]

Griffiths, T. L., Kemp, C. & Tenenbaum, J. B. (2008a) Bayesian models of cognition. In: *Cambridge handbook of computational psychology*, ed. R. Sun, pp. 59–100. Cambridge University Press. [NC]

Griffiths, T. L., Sanborn, A. N., Canini, K. R. & Navarro, D. J. (2008b) Categorization as nonparametric Bayesian density estimation. In: *The probabilistic mind: Prospects for rational models of cognition*, ed. M. Oaksford & N. Chater. Oxford University Press. [aMJ]

Griffiths, T. L., Steyvers, M. & Tenenbaum, J. B. (2007) Topics in semantic representation. *Psychological Review* 114:211–44. [arMJ]

Griffiths, T. L. & Tenenbaum, J. B. (2005) Structure and strength in causal induction. *Cognitive Psychology* 51:354–84. [KJH]

Griffiths, T. L. & Tenenbaum, J. B. (2006) Optimal predictions in everyday cognition. *Psychological Science* 17(9):767–73. [JSB, aMJ]

Griffiths, T. L. & Tenenbaum, J. B. (2009) Theory-based causal induction. *Psychological Review* 116:661–716. [NC, KJH, aMJ]

Guttman, N. & Kalish, H. I. (1956) Discriminability and stimulus generalization. *Journal of Experimental Psychology* 51:79–88. [aMJ]

Hagmayer, Y., Sloman, S. A., Lagnado, D. A. & Waldmann, M. R. (2007) Causal reasoning through intervention. In: *Causal learning: Psychology, philosophy, and computation*, ed. A. Gopnik & L. Schultz, pp. 86–100. Oxford University Press. [IB]

Hamilton, V. L. (1980) Intuitive psychologist or intuitive lawyer: Alternative models of the attribution process. *Journal of Personality and Social Psychology* 39:767–73. [ELU]

Hamilton, W. D. (1964) The genetical theory of social behavior. *Journal of Theoretical Biology* 7:1–52. [aMJ]

Harm, M. W. & Seidenberg, M. (2004) Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review* 111(3):662–720. [TTR]

Harm, M. W. & Seidenberg, M. S. (1999) Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review* 106:491–528. [IB]

Hastie, T., Tibshirani, R. & Friedman, J. H. (2009) *The elements of statistical learning*, 2nd edition. Springer. [SE]

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109. [aMJ]

Hattori, M. & Oaksford, M. (2007) Adaptive non-interventional heuristics for covariation detection in causal induction: Model comparison and rational analysis. *Cognitive Science* 31:765–814. [KJH]

Hayes, B. K., Heit, E. & Swendsen, H. (2010) Inductive reasoning. *Wiley Interdisciplinary Reviews: Cognitive Science* 1:278–92. [EH]

Hayes, B. K., Kurniawan, H. & Newell, B. R. (2011) Rich in vitamin C or just a convenient snack? Multiple-category reasoning with cross-classified foods. *Memory and Cognition* 39:92–106. [BKH]

Hayes, B. K. & Newell, B. R. (2009) Induction with uncertain categories: When do people consider the alternative categories? *Memory and Cognition* 37:730–43. [BKH]

Hebb, D. O. (1949) *The organization of behavior: A neuropsychological theory*. John Wiley. [aMJ]

Heit, E. (1995) Belief revision in models of category learning. In: *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society, Pittsburgh, PA, July 22–25, 1995*, ed. J. D. Moore & J. F. Lehman, pp. 176–81. Erlbaum. [EH]

Heit, E. (1997) Knowledge and concept learning. In: *Knowledge, concepts, and categories*, ed. K. Lamberts & D. Shanks, pp. 7–41. Psychology Press. [EH]

Heit, E. (1998) A Bayesian analysis of some forms of inductive reasoning. In: *Rational models of cognition*, ed. M. Oaksford & N. Chater, pp. 248–74. Oxford University Press. [EH]

Heit, E. (2000) Properties of inductive reasoning. *Psychonomic Bulletin and Review* 7:569–92. [SE, EH]

Heit, E. (2001) Background knowledge and models of categorization. In: *Similarity and categorization*, ed. U. Hahn & M. Ramscar, pp. 155–78. Oxford University Press. [EH]

Heit, E. & Bott, L. (2000) Knowledge selection in category learning. In: *Psychology of learning and motivation, vol. 39*, ed. D. L. Medin, pp. 163–99. Academic Press. [EH]

Heit, E., Briggs, J. & Bott, L. (2004) Modeling the effects of prior knowledge on learning incongruent features of category members. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30:1065–81. [EH]

Heller, K., Sanborn, A. N. & Chater, N. (2009) Hierarchical learning of dimensional biases in human categorization. In: *Advances in neural information processing systems, vol. 22*, ed. J. Lafferty & C. Williams. pp. 727–35. MIT Press. [NC]

Hogarth, R. M. & Karelaia, N. (2005) Ignoring information in binary choice with continuous variables: When is less "more"? *Journal of Mathematical Psychology* 49:115–24. [aMJ]

Holyoak, K. J. & Cheng, P. W. (2011) Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology* 62:135–63. [KJH]

Holyoak, K. J., Lee, H. S. & Lu, H. (2010) Analogical and category-based inference: A theoretical integration with Bayesian causal models. *Journal of Experimental Psychology: General* 139:702–27. [KJH]

Horgan, J. (1999) The undiscovered mind: How the human brain defies replication, medication, and explanation. *Psychological Science* 10:470–74. [aMJ]

Hornik, K., Stinchcombe, M. & White, H. (1989) Multilayer feedforward networks are universal approximators. *Neural Networks* 2:359–66. [aMJ]

Howson, C. & Urbach, P. (1991) Bayesian reasoning in science. *Nature* 350:371–74. [SE]

Hsu, A. & Chater, N. (2010) The logical problem of language acquisition: A probabilistic perspective. *Cognitive Science* 34:972–1016. [NC]

Hsu, A., Chater, N. & Vitányi, P. M. B. (in press) The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis. *Cognition*. [NC]

Huber, D. E., Shiffrin, R. M., Lyle, K. B. & Ruys, K. I. (2001) Perception and preference in short-term word priming. *Psychological Review* 108:149–82. [aMJ]

Hume, D. (1740) *A treatise of human nature*. [Available online through a variety of sources, including Project Gutenberg at: http://www.gutenberg.org/ebooks/4705] [SE]

Hummel, J. E. & Biederman, I. (1992) Dynamic binding in a neural network for shape recognition. *Psychological Review* 99:480–517. [aMJ]

Huszar, F., Noppeney, U. & Lengyel, M. (2010) Mind reading by machine learning: A doubly Bayesian method for inferring mental representations. In: *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, ed. R. Catrambone & S. Ohlsson, pp. 2810–15. Cognitive Science Society. [MDL]

Jacobs, R. A. (1997) Nature, nurture, and the development of functional specializations: A computational approach. *Psychonomic Bulletin and Review* 4:299–309. [EH]

Jaynes, E. T. (1968) Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics* 4:227–41. [aMJ]

Jeffreys, H. (1946) An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London Series A: Mathematical, Physical and Engineering Sciences* 186:453–61. [aMJ]

Jenkins, G. J., Smith, J. R., Spencer, J. P. & Samuelson, L. K. (in press) When more evidence makes word learning less suspicious. In: *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, ed. L. Carlson, C. Hölscher, & T. Shipley. Cognitive Science Society. [GWJ]

Joanisse, M. F. & Seidenberg, M. S. (1999) Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Sciences USA* 96:7592–97. [aMJ]

Joanisse, M. F. & Seidenberg, M. S. (2003) Phonology and syntax in specific language impairment: Evidence from a connectionist model. *Brain and Language* 86:40–56. [aMJ]

Johnson, M., Griffiths, T. L. & Goldwater, S. (2007) Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In: *Advances in neural information processing systems, vol. 19*, ed. B. Schölkopf, J. Platt & T. Hofmann, pp. 641–48. MIT Press. [NC]

Johnson, M. H. (1998) The neural basis of cognitive development. In: *Handbook of child psychology, vol. 2: Cognition, perception, and language*, ed. D. Kuhm & R. S. Siegler, pp. 1–49. Wiley. [aMJ]

Johnson-Laird, P. N. (1994) Mental models, deductive reasoning, and the brain. In: *The cognitive neurosciences*, ed. M. S. Gazzaniga, pp. 999–1008. MIT Press. [EH]

Jones, M. & Sieck, W. R. (2003) Learning myopia: An adaptive recency effect in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29:626–40. [aMJ]

Jones, M. & Zhang, J. (2003) Which is to blame: Instrumental rationality, or common knowledge? *Behavioral and Brain Sciences* 26:166–67. [aMJ]

Kalish, M. L., Griffiths, T. L. & Lewandowsky, S. (2007) Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin and Review* 14:288–94. [DKS]

Kalish, M. L., Lewandowsky, S. & Kruschke, J. K. (2004) Population of linear experts: Knowledge partitioning and function learning. *Psychological Review* 111:1072–99. [DKS]

Kamin, L. J. (1969) Predictability, surprise, attention, and conditioning. In: *Punishment and aversive behavior*, ed. B. A. Campbell & R. M. Church, pp. 279–96. Appleton-Century-Crofts. [NC]

Kant, I. (1787/1961) *Critique of pure reason*, trans. N. K. Smith. St. Martin's Press. (Original work published in 1787). [aMJ]

Kawato, M. (1999) Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology* 9:718–27. [SE]

Kelso, J. A. S. (1995) *Dynamic patterns: The self-organization of brain and behavior*. MIT Press. [MH]

Kemp, C. & Tenenbaum, J. B. (2008) The discovery of structural form. *Proceedings of the National Academy of Sciences USA* 105:10687–692. [NC, aMJ]

Kemp, C. & Tenenbaum, J. B. (2009) Structured statistical models of inductive reasoning. *Psychological Review* 116:20–58. [NC, EH, BR]

Kemp, C., Goodman, N. D. & Tenenbaum, J. B. (2010a) Learning to learn causal relations. *Cognitive Science* 34:1185–1243. [NC]

Kemp, C., Perfors, A. & Tenenbaum, J. B. (2007) Learning overhypotheses with hierarchical Bayesian models. *Developmental Science* 10:307–21. [GWJ, arMJ, DJN]

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T. & Ueda, N. (2006) Learning systems of concepts with an infinite relational model. In: *Proceedings of the 21st National Conference on Artificial Intelligence, vol. 1*, ed. A. Cohn, pp. 381–88. AAAI Press. [NC]

Kemp, C., Tenenbaum, J. B., Niyogi, S. & Griffiths, T. L. (2010b) A probabilistic model of theory formation. *Cognition* 114(2):165–96. [NC]

Kim, J. J., Krupa, D. J. & Thompson, R. F. (1998) Inhibitory cerebello-olivary projections and blocking effect in classical conditioning. *Science* 279:570–73. [IB]

Klayman, J. & Ha, Y.-W. (1987) Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review* 94:211–28. [DJN]

Knill, D. & Richards, W., eds. (1996) *Perception as Bayesian inference*. Cambridge University Press. [SE]

Körding, K. P. & Wolpert, D. M. (2006) Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences* 10:319–26. [SE]

Köver, H., Bao, S. (2010) Cortical plasticity as a mechanism for storing Bayesian priors in sensory perception. *PLoS ONE* 5(5):e10497. [aMJ]

Krueger, J. I. & Funder, D. C. (2004) Towards a balanced social psychology: Causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences* 27:313–27. [ELU]

Krugman, P. (2009) How did economists get it so wrong? *New York Times*, MM36, September 2. [aMJ]

Kruschke, J. K. (2006) Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review* 113:677–99. [KJH, DKS]

Kruschke, J. K. (2008) Bayesian approaches to associative learning: From passive to active learning. *Learning and Behavior* 36:210–26. [DKS]

Kruschke, J. K. (2010) Bridging levels of analysis: Comment on McClelland et al. and Griffiths et al. *Trends in Cognitive Sciences* 14:344–45. [DKS]

Kruschke, J. K. (2010) What to believe: Bayesian methods for data analysis. *Trends in Cognitive Science* 14:293–300. [MDL]

Kuhn, T. S. (1970) *The structure of scientific revolutions*, 2nd edition. University of Chicago Press. [DKS]

Kunda, Z. (1987) Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology* 53:37–54. [ELU]

Kurz, E. M. & Tweney, R. D. (1998) The practice of mathematics and science: From calculus to the clothesline problem. In: *Rational models of cognition*, ed. M. Oaksford & N. Chater, pp. 415–38. Oxford University Press. [aMJ]

Kushnir, T. & Gopnik, A. (2005) Young children infer causal strength from probabilities and interventions. *Psychological Science* 16(9):678–83. [AG]

Kushnir, T. & Gopnik, A. (2007) Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental Psychology* 43(1):186–96. [AG]

Kushnir, T., Xu, F. & Wellman, H. (2010) Young children use statistical sampling to infer the preferences of others. *Psychological Science* 21:1134–40. [AG]

Lagnado, D. A. & Sloman, S. A. (2006) Time as guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32(3):451–60. [IB]

Lakatos, I. (1970) Falsification and the methodology of scientific research programmes. In: *Criticism and the growth of knowledge*, ed. I. Lakatos & A. Musgrave, pp. 91–196. Cambridge University Press. [BR]

Lee, M. D. (2008) Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin and Review* 15:1–15. [MDL]

Lee, M. D. (2010) Emergent and structured cognition in Bayesian models: Comment on Griffiths et al. and McClelland et al. *Trends in Cognitive Sciences* 14:345–46. [MDL]

Lee, M. D. (2011) How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology* 55:1–7. [MDL]

Lee, M. D. & Sarnecka, B. W. (2010) A model of knower-level behavior in number-concept development. *Cognitive Science* 34:51–67. [aMJ]

Lee, M. D. & Sarnecka, B. W. (2010) A model of knower-level behavior in number-concept development. *Cognitive Science* 34:51–67. [MDL]

Lee, M. D. & Sarnecka, B. W. (in press) Number knower-levels in young children: Insights from a Bayesian model. *Cognition*. [MDL]

Lerner, J. S. & Tetlock, P. E. (1999) Accounting for the effects of accountability. *Psychological Bulletin* 125:255–75. [ELU]

Lewandowsky, S., Griffiths, T. L. & Kalish, M. L. (2009) The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science* 33:969–98. [DKS]

Lewandowsky, S. & Heit, E. (2006) Some targets for memory models. *Journal of Memory and Language* 55:441–46. [EH]

Lewandowsky, S., Kalish, M. & Ngang, S. K. (2002) Simplified learning in complex situations: Knowledge partitioning in function learning. *Journal of Experimental Psychology: General* 131:163–93. [DKS]

Lewandowsky, S., Roberts, L. & Yang, L.-X. (2006) Knowledge partitioning in categorization: Boundary conditions. *Memory and Cognition* 34:1676–88. [DKS]

Little, D. R. & Lewandowsky, S. (2009) Beyond nonutilization: Irrelevant cues can gate learning in probabilistic categorization. *Journal of Experimental Psychology: Human Perception and Performance* 35:530–50. [DKS]

Logan, G. D. (1988) Toward an instance theory of automaticity. *Psychological Review* 95:492–527. [ABM]

Lord, C. G., Ross, L. & Lepper, M. R. (1979) Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology* 37:2098–109. [ELU]

Louwerse, M. M. & Connell, L. (2011) A taste of words: Linguistic context and perceptual simulation predict the modality of words. *Cognitive Science* 35:381–98. [LWB]

Love, B. C. (2002) Comparing supervised and unsupervised category learning. *Psychonomic Bulletin and Review* 9:829–35. [aMJ]

Love, B. C. (2005) Environment and goals jointly direct category acquisition. *Current Directions in Psychological Science* 14:195–99. [rMJ]

Lu, H., Rojas, R. R., Beckers, T. & Yuille, A. L. (2008a) Sequential causal learning in humans and rats. In: *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society*, ed. B. C. Love, K. McRae & V. M. Sloutsky, pp. 195–88. Cognitive Science Society. [KJH]

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W. & Holyoak, K. J. (2008b) Bayesian generic priors for causal learning. *Psychological Review* 115:955–82. [KJH]

Lucas, C., Gopnik, A. & Griffiths, T. (2010) Developmental differences in learning the form of causal relationships. In: *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, ed. S. Ohlsson & R. Catrambone, pp. 2852–57. Cognitive Science Society. [AG]

Lucas, C., Griffiths, T. L., Xu, F. & Fawcett, C. (2009) A rational model of preference learning and choice prediction by children. *Advances in Neural Information Processing Systems* 21:985–92. [aMJ]

Luce, R. D. (1963) Detection and recognition. In: *Handbook of mathematical psychology*, ed. R. D. Luce, R. R. Bush & E. Galanter, pp. 103–89. John Wiley. [aMJ]

Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. (2006) Bayesian inference with probabilistic population codes. *Nature Neuroscience* 9:1432–38. [DB]

Machamer, P., Darden, L. & Craver, C. F. (2000) Thinking about mechanisms. *Philosophy of Science* 67:1–25. [MH]

Machery, E. & Barrett, C. (2006) Debunking adapting minds. *Philosophy of Science* 73:232–46. [aMJ]

MacKay, D. (2002) *Information theory, inference, and learning algorithms*. Cambridge University Press. [NC]

Maloney, L. T. & Mamassian, P. (2009) Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Visual Neuroscience* 26:147–55. [BR]

Maloney, L. T. & Zhang, H. (2010) Decision-theoretic models of visual perception and action. *Vision Research* 50:2362–74. [BR]

Maloney, L. T. & Zhang, H. (2010) Decision-theoretic models of visual perception and action. *Vision Research* 50:2362–74. [BR]

Malt, B. C. & Smith, E. E. (1984) Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior* 23:250–69. [BKH]

Manning, C. & Schütze, H. (1999) *Foundations of statistical natural language processing*. MIT Press. [NC]

Marcus, G. F. (1998) Rethinking eliminative connectionism. *Cognitive Psychology* 37:243–82. [aMJ]

Marcus, G. F. (2008) *Kluge: The haphazard construction of the human mind*. Houghton Mifflin. [aMJ]

Markman, A. B. & Ross, B. H. (2003) Category use and category learning. *Psychological Bulletin* 129:592–615. [aMJ]

Marr, D. (1982) *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman. [BLA, LWB, NC, MH, arMJ, ABM, DN, DP]

Marr, D. (1982/2010) *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman/MIT Press. (Original work published in 1982; 2010 reprint edition by MIT Press). [DKS]

Marroquin, J., Mitter, S. & Poggio, T. (1987) Probabilistic solution of ill-posed problems in computational vision. *Journal of the American Statistical Association* 82:76–89. [SE]

Mayr, E. (1982) *The growth of biological thought: Diversity, evolution, and inheritance*. Harvard University Press. [aMJ]

McClelland, J. L. (1998) Connectionist models and Bayesian inference. In: *Rational models of cognition*, ed. M. Oaksford & N. Chater, pp. 21–53. Oxford University Press. [NC, EH]

McClelland, J. L. (2010) Emergence in cognitive science. *Topics in Cognitive Science* 2:751–70. [NC]

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S. & Smith, L. B. (2010) Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences* 14:348–56. [DKS]

McClelland, J. L. & Chappell, M. (1998) Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review* 105:724–60. [EH]

McClelland, J. L. & Patterson, K. (2002) Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences* 6(11):465–72. [TTR]

McClelland, J. L. & Rumelhart, D. E. (1988) *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. MIT Press. [IB]

McClelland, J. L., Rumelhart, D. E. & the PDP Research Group. (1986) *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2: Psychological and biological models*. MIT Press. [aMJ]

McCulloch, W. & Pitts, W. (1943) A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 7:115–33. [aMJ]

McKenzie, C. R. M. & Mikkelsen, L. A. (2007) A Bayesian view of covariation assessment. *Cognitive Psychology* 54:33–61. [aMJ]

McNamara, J. M. & Houston, A. I. (2009) Integrating function and mechanism. *Trends in Ecology and Evolution* 24:670–75. [aMJ]

Michaels, C. F. & Carello, C. (1981) *Direct perception*. Prentice-Hall. [aMJ]

Miller, E. K. & Cohen, J. D. (2001) An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience* 24:167–202. [aMJ]

Miller, G. A. (2003) The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences* 7:141–44. [aMJ]

Minsky, M. & Papert, S. A. (1969) *Perceptrons: An introduction to computational geometry*. MIT Press. [aMJ]

Mortimer, D., Feldner, J., Vaughan, T., Vetter, I., Pujic, Z., Rosoff, W. J., Burrage, K., Dayan, P., Richards, L. J. & Goodhill, G. J. (2009) Bayesian model predicts the response of axons to molecular gradients. *Proceedings of the National Academy of Sciences USA* 106:10296–301. [aMJ]

Mozer, M. C., Pashler, H. & Homaei, H. (2008) Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science* 32:1133–47. [aMJ]

Murphy, G. L. (1993) A rational theory of concepts. *Psychology of Learning and Motivation* 29:327–59. [aMJ]

Murphy, G. L. & Medin, D. L. (1985) The role of theories in conceptual coherence. *Psychological Review* 92:289–316. [EH]

Murphy, G. L. & Ross, B. H. (2007) Use of single or multiple categories in category-based induction. In: *Inductive reasoning: Experimental, developmental, and computational approaches*, ed. A. Feeney & E. Heit, pp. 205–25. Cambridge Press. [BKH, rMJ]

Murphy, G. L. & Ross, B. H. (2010) Category vs. object knowledge in category-based induction. *Journal of Memory and Language* 63:1–17. [BKH]

Mussa-Ivaldi, F. A. & Giszter, S. F. (1992) Vector field approximation: A computational paradigm for motor control and learning. *Biological Cybernetics* 67:491–500. [SE]

Navarro, D. J. (2010) Learning the context of a category. In: *Advances in neural information processing systems, vol. 23*, ed. J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel & A. Culotta, pp. 1795–803. MIT Press. [DKS]

Navarro, D. J., Dry, M. J. & Lee, M. D. (in press) Sampling assumptions in inductive generalization. *Cognitive Science*. [DJN]

Navarro, D. J. & Perfors, A. F. (2009) Learning time-varying categories. In: *Proceedings of the 31st Annual Conference of the Cognitive Science Society, Austin, TX*, ed. N. Taatgen, H. van Rijn, L. Schomaker & J. Nerbonne, pp. 419–24. Cognitive Science Society. [DJN]

Navarro, D. J. & Perfors, A. F. (2011) Hypothesis generation, sparse categories and the positive test strategy. *Psychological Review* 118:120–34. [DJN]

Neal, R. M. (1992) Connectionist learning of belief networks. *Artificial Intelligence* 56:71–113. [NC]

Neisser, U. (1967) *Cognitive psychology*. Appleton-Century-Crofts. [DKS]

Nersessian, N. J. (1986) A cognitive-historical approach to meaning in scientific theories. In: *The process of science: Contemporary philosophical approaches to understanding scientific practice*, ed. N. J. Nersessian. Martinus Nijhoff. [aMJ]

Neuhoff, J. G. (2001) An adaptive bias in the perception of looming auditory motion. *Ecological Psychology* 13:87–110. [LA-S]

Newell, A. & Simon, H. A. (1972) *Human problem solving*. Prentice-Hall. [aMJ]

Newell, B. R., Paton, H., Hayes, B. K. & Griffiths, O. (2010) Speeded induction under uncertainty: The influence of multiple categories and feature conjunctions. *Psychonomic Bulletin and Review* 17:869–74. [BKH]

Newport, E. L. (1990) Maturational constraints on language learning. *Cognitive Science* 14:11–28. [aMJ]

Norman, K. A. & O'Reilly, R. C. (2003) Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. *Psychological Review* 110(4):611–46. [EH, TTR]

Norton, M. I., Vandello, J. A. & Darley, J. M. (2004) Casuistry and social category bias. *Journal of Personality and Social Psychology* 87:817–31. [ELU]

Nosofsky, R. M. (1986) Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115:39–57. [DKS]

Nosofsky, R. M., Palmeri, T. J. & Mckinley, S. C. (1994) Rule-plus-exception model of classification learning. *Psychological Review* 104:266–300. [aMJ]

O'Reilly, R. C. & Norman, K. A. (2002) Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends in Cognitive Sciences* 6(12):505–10. [TTR]

Oaksford, M. & Chater, N. (1994) A rational analysis of the selection task as optimal data selection. *Psychological Review* 101:608–31. [NC, aMJ, DJN]

Oaksford, M. & Chater, N. (1998a) An introduction to rational models of cognition. In: *Rational models of cognition*, ed. M. Oaksford & N. Chater, pp. 1–18. Oxford University Press. [aMJ]

Oaksford, M. & Chater, N., ed. (1998b) *Rational models of cognition*. Oxford University Press. [NC]

Oaksford, M. & Chater, N. (2003) Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin and Review* 10:289–318. [NC]

Oaksford, M. & Chater, N. (2007) *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press. [DB, DD, arMJ]

Oaksford, M. & Chater, N. (2010) Conditionals and constraint satisfaction: Reconciling mental models and the probabilistic approach? In: *Cognition and conditionals: Probability and logic in human thinking*, ed. M. Oaksford & N. Chater, pp. 309–34. Oxford University Press. [NC, aMJ]

Öhman, A., Flykt, A. & Esteves, F. (2001) Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General* 130:466–78. [LA-S]

Olshausen, B. A. & Field, D. J. (1996) Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. *Nature* 381:607–9. [TTR]

Oppenheim, R. W. (1981) Ontogenetic adaptations and retrogressive processes in the development of the nervous system and behavior. In: *Maturation and development: Biological and psychological perspectives*, ed. K. J. Connolly & H. F. R. Prechtl, pp. 73–108. International Medical. [aMJ]

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A. & Shafir, E. (1990) Category-based induction. *Psychological Review* 97:185–200. [BKH]

Papadopoulos, C., Hayes, B. K. & Newell, B. R. (2011) Non-categorical approaches to feature prediction with uncertain categories. *Memory and Cognition* 39:304–18. [BKH]

Pavlov, I. P. (1927) *Conditioned reflexes*. Oxford University Press. [IB]

Payne, J. W., Bettman, J. R. & Johnson, E. J. (1993) *The adaptive decision maker*. Cambridge University Press. [ABM]

Pearl, J. (1988) *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann. [NC, KJH]

Pearl, J. (2000) *Causality: Models, reasoning, and inference*. Cambridge University Press. [DB, AG, rMJ]

Perales, J. C. & Shanks, D. R. (2007) Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin and Review* 14:577–96. [KJH]

Perfors, A. (2011) Memory limitations alone do not lead to over-regularization: An experimental and computational investigation. In: *Proceedings of the 33rd Annual Conference of the Cognitive Science Society, Boston, MA*, ed. L. Carlson, C. Hoelscher & T. F. Shipley, pp. 3274–79. Cognitive Science Society. [DJN]

Perfors, A. & Tenenbaum, J. (2009) Learning to learn categories. In: *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, ed. N. Taatgen, H. van Rijn, L. Schomaker & J. Nerbonne, pp. 136–41. Cognitive Science Society. [DJN]

Perfors, A., Tenenbaum, J. B. & Regier, T. (2011) The learnability of abstract syntactic principles. *Cognition* 118:306–38. [NC]

Perfors, A., Tenenbaum, J. & Wonnacott, E. (2010) Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language* 37:607–42. [NC, DJN]

Perrigo, G., Belvin, L. & Vom Saal, F. S. (1991) Individual variation in the neural timing of infanticide and parental behavior in male house mice. *Physiology and Behavior* 50:287–96. [DP]

Perrigo, G., Belvin, L. & Vom Saal, F. S. (1992) Time and sex in the male mouse: Temporal regulation of infanticide and parental behavior. *Chronobiology International* 9:421–33. [DP]

Perry, L. K., Cook, S. W. & Samuelson L. K. (in preparation) An exploration of context, task, and stimuli effects on similarity perception. [GWJ]

Pezzulo, G., Barsalou, L. W., Cangelosi, A., Fischer, M. A., McRae, K. & Spivey, M. (2011) The mechanics of embodiment: A dialogue on embodiment and computational modeling. *Frontiers in Cognition* 2(5):1–21. [LWB]

Pinker, S. (1995) *The language instinct: How the mind creates language*. Perennial. [aMJ]

Pinker, S. (2002) *The blank slate: The modern denial of human nature*. Viking. [aMJ]

Pinker, S. & Prince, A. (1988) On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28:73–193. [aMJ]

Pitt, M. A., Myung, I. J. & Zhang, S. (2002) Toward a method of selecting among computational models of cognition. *Psychological Review* 109:472–91. [aMJ]

Plaut, D. C. (1995) Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology* 17:291–321. [IB]

Plaut, D. C., McClelland, J. L., Seidenberg, M. S. & Patterson, K. (1996) Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review* 103:56–115. [aMJ]

Poggio, T. (1990) A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology* 55:899–910. [SE]

Pollack, J. B. (1990) Recursive distributed representations. *Artificial Intelligence* 46:77–105. [aMJ]

Pothos, E. M. & Chater, N. (2002) A simplicity principle in unsupervised human categorization. *Cognitive Science* 26:303–43. [aMJ]

Pronin, E., Gilovich, T. & Ross, L. (2004) Objectivity in the eye of the beholder: Perceptions of bias in self versus others. *Psychological Review* 111:781–99. [ELU]

Pronin, E., Lin, D. Y. & Ross, L. (2002) The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin* 28:369–81. [ELU]

Pyszczynski, T. & Greenberg, J. (1987) Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. In: *Advances in experimental social psychology, vol. 20*, ed. L. Berkowitz, pp. 297–340. Academic Press. [ELU]

Rachman, S. (1997) The evolution of cognitive behaviour therapy. In: *Science and practice of cognitive behaviour therapy*, ed. D. Clark, C. G. Fairburn & M. G. Gelder, pp. 1–26. Oxford University Press. [aMJ]

Raiffa, H. & Schlaifer, R (1961) *Applied statistical decision theory*. Harvard University Press. [aMJ]

Ramscar, M., Yarlett, D., Dye, M., Denny, K. & Thorpe, K. (2010) The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science* 34:1–49. [aMJ]

Ravi, S. & Knight, K. (2009) Minimized models for unsupervised part-of-speech tagging. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, ed. K.-Y. Su, pp. 504–12. Association for Computational Linguistics. Available at: http://www.aclweb.org/anthology-new/P/P09/P09-1057.pdf [aMJ]

Rehder, B. (2009) Causal-based property generalization. *Cognitive Science* 33:301–43. [BR]

Rehder, B. & Burnett, R. (2005) Feature inference and the causal structure of categories. *Cognitive Psychology* 50:264–314. [NC, BR]

Rehder, B. & Kim, S. (2010) Causal status and coherence in causal-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 36:1171–206. [BR]

Rescorla, R. A. & Wagner, A. R. (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: *Classical conditioning II: Current theory and research*, ed. A. H. Black & W. F. Prokasy, pp. 64–99. Appleton-Century-Crofts. [IB, NC, KJH, rMJ]

Ricci, G. & Levi-Civita, T. (1900) Méthodes de calcul différentiel absolu et leurs applications [Methods of absolute differential calculus and their applications]. *Mathematische Annalen* 54(1–2):125–201. [aMJ]

Rips, L. J. (1990) Reasoning. *Annual Review of Psychology* 41:321–53. [EH]

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges J. R. & Patterson, K. (2004) The structure and deterioration of semantic memory: A computational and neuropsychological investigation. *Psychological Review* 111(1):205–35. [TTR]

Rogers, T. T. & McClelland, J. L. (2004) *Semantic cognition: A parallel distributed processing approach*. MIT Press. [TTR]

Rogers, T. T. & Plaut, D. C. (2002) Connectionist perspectives on category-specific deficits. In: *Category-specificity in brain and mind*, ed. E. Forde & G. W. Humphreys, pp. 251–89. Psychology Press. [aMJ]

Rosch, E. (1978) Principles of categorization. In: *Cognition and categorization*, ed. E. Rosch & B. B. Lloyd, pp. 27–48. Erlbaum. [aMJ]

Rosch, E. & Mervis, C. B. (1975) Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7:573–605. [BKH]

Rosen, R. (1991) *Life itself*. Columbia University Press. [BLA]

Rosenblatt, F. (1962) *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Spartan Books. [aMJ]

Ross, B. H. & Murphy, G. L. (1996) Category-based predictions: Influence of uncertainty and feature associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22:736–53. [BKH]

Ross, L. & Ward, A. (1996) Naive realism in everyday life: Implications for social conflict and misunderstanding. In: *Values and knowledge*, ed. E. S. Reed & E. Turiel, pp. 103–35. Erlbaum. [ELU]

Rougier, N. P., Noelle, D., Braver, T. S., Cohen, J. D. & O'Reilly, R. C. (2005) Prefrontal cortex and the flexibility of cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences USA* 102:7338–43. [aMJ]

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986) Learning representations by back-propagating errors. *Nature* 323:533–36. [aMJ]

Rumelhart, D. E. & McClelland, J. L. (1985) Levels indeed! A response to Broadbent. *Journal of Experimental Psychology: General* 114:193–97. [DKS]

Rumelhart, D. E., McClelland, J. L. & the PDP research group. (1986) *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations*. MIT Press. [aMJ]

Russell, S. & Norvig, P. (2011) *Artificial intelligence: A modern approach*, 3rd edition. Prentice-Hall. [NC]

Sakamoto, Y., Jones, M. & Love, B. C. (2008) Putting the psychology back into psychological models: Mechanistic versus rational approaches. *Memory and Cognition* 36(6):1057–65. [arMJ, DJN, DN]

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press. [BR]

Samuelson, L. K., Schutte, A. R. & Horst, J. S. (2009) The dynamic nature of knowledge: Insights from a dynamic field model of children's novel noun generalization. *Cognition* 110:322–45. [GWJ]

Sanborn, A. N., Griffiths, T. L. & Navarro, D. J. (2010a) Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review* 117:1144–67. [NC, BKH, arMJ, DJN, BR, DKS]

Sanborn, A. N., Griffiths, T. L. & Shiffrin, R. M. (2010b) Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology* 60:63–106. [aMJ]

Sargent, T. J. (1993) *Bounded rationality in macroeconomics*. Oxford University Press. [aMJ]

Savage, L. J. (1954) *The foundations of statistics*. John Wiley/Dover. [aMJ, MS]

Schall, J. D. (2004) On building a bridge between brain and behavior. *Annual Review of Psychology* 55:23–50. [DKS]

Schervish, M. J. (1995) *Theory of statistics*. Springer Series in Statistics. Springer. [SE]

Schneider, W. & Shiffrin, R. M. (1977) Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review* 84(1):1–66. [ABM]

Schulz, L. E. & Bonawitz, E. B. (2007) Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology* 43:1045–50. [AG]

Schulz, L. E., Bonawitz, E. B. & Griffiths, T. L. (2007a) Can being scared cause tummy aches? Naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology* 43(5):1124–39. [AG]

Schulz, L. E, Gopnik, A. & Glymour, C. (2007b) Preschool children learn about causal structure from conditional interventions. *Developmental Science* 10(3):322–32. [AG]

Schulz, L. E., Goodman, N. D., Tenenbaum, J. B. & Jenkins, A. C. (2008) Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data. *Cognition* 109(2):211–23. [AG]

Schustack, M. W. & Sternberg, R. J. (1981) Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General* 110:101–20. [KJH]

Schwarz, G. E. (1978) Estimating the dimension of a model. *Annals of Statistics* 6(2):461–64. [aMJ]

Seidenberg, M. & Plaut, D. C. (in press) Idealization and its discontents: The legacy of the past tense debate. *Cognitive Science*. [TTR]

Sewell, D. K. & Lewandowsky, S. (2011) Restructuring partitioned knowledge: The role of recoordination in category learning. *Cognitive Psychology* 62:81–122. [DKS]

Shafto, P., Kemp, C., Mansinghka, V. M. & Tenenbaum, J. B. (2011) A probabilistic model of cross-categorization. *Cognition*. 120:1–25. [aMJ]

Shanks, D. R. & Dickinson, A. (1987) Associative accounts of causality judgment. In: *The psychology of learning and motivation, vol. 21*, ed. G. H. Bower, pp. 229–61. Academic Press. [KJH]

Shannon, C. E. (1949) *The mathematical theory of communication*. University of Illinois Press. [ABM]

Shepard, R. N. (1987) Toward a universal law of generalization for psychological science. *Science* 237:1317–23. [DJN]

Sherman, D. K. & Cohen, G. L. (2002) Accepting threatening information: Self-affirmation and the reduction of defensive biases. *Current Directions in Psychological Science* 11:119–23. [ELU]

Sherman, D. K. & Cohen, G. L. (2006) The psychology of self-defense: Self-affirmation theory. In: *Advances in Experimental Social Psychology, vol. 38*, ed. M. P. Zanna, pp. 183–242. Academic Press. [ELU]

Shi, L., Griffiths, T. L., Feldman, N. H. & Sanborn, A. N. (2010) Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin and Review* 17:443–64. [NC, EH, DKS]

Shiffrin, R. M. & Steyvers, M. (1997) A model for recognition memory: REM – Retrieving effectively from memory. *Psychonomic Bulletin & Review* 4:145–66. [EH]

Shiffrin, R. M. & Steyvers, M. (1998) The effectiveness of retrieval from memory. In: *Rational models of cognition*, ed. M. Oaksford & N. Chater, pp. 73–95. Oxford University Press. [arMJ]

Shiffrin, R. M., Lee, M. D., Kim, W. & Wagenmakers, E. J. (2008) A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science* 32:1248–84. [DKS]

Simon, H. A. (1957a) A behavioral model of rational choice. In: *Models of man, social and rational: Mathematical essays on rational human behavior in a social setting*, ed. H. A. Simon, pp. 241–60. John Wiley. [aMJ]

Simon, H. A. (1957b) *Models of man: Social and rational*. John Wiley. [ABM]

Simoncelli, E. P. & Olshausen, B. A. (2001) Natural image statistics and neural representation. *Annual Review of Neuroscience* 24:1193–216. [TTR]

Skinner, B. F. (1938) *The behavior of organisms: An experimental analysis*. Appleton-Century. [arMJ]

Skinner, B. F. (1957) *Verbal behavior*. Appleton-Century-Crofts. [aMJ]

Skinner, B. F. (1958) Reinforcement today. *American Psychologist* 13:94–99. [aMJ]

Sloman, S. A. (1993) Feature-based induction. *Cognitive Psychology* 25:231–80. [BKH]

Sloman, S. A. & Fernbach, P. M. (2008) The value of rational analysis: An assessment of causal reasoning and learning. In: *The probabilistic mind: Prospects for rational models of cognition*, ed. Chater, N. & Oaksford, M, pp. 485–500. Oxford University Press. [PMF, aMJ]

Smith, D. L. (2007) Beyond Westemarck: Can shared mothering or maternal phenotype matching account for incest avoidance? *Evolutionary Psychology* 5:202–22. [aMJ]

Smith, L. B. (2000) Learning how to learn words: An associative crane. In: *Becoming a word learner: A debate on lexical acquisition*, ed. R. M. Golinkoff, K. Hirsh-Pasek, K., L. Bloom, L.B. Smith, A. L. Woodward, N. Akhtar, M. Tomasello & G. Hollich, pp. 51–80. Oxford University Press. [GWJ]

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L. & Samuelson, L. (2002) Object name learning provides on-the-job training for attention. *Psychological Science* 13:13–19. [GWJ, arMJ]

Smith, L. B. & Thelen, E. (2003) Development as a dynamic system. *Trends in Cognitive Science* 7(8):343–48. [GWJ]

Smith, P. K. (1982) Does play matter? Functional and evolutionary aspects of animal and human play. *Behavioral and Brain Sciences* 5:139–84. [aMJ]

Smolensky, P. (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11:1–23. [aMJ]

Smolin, L. (2006) *The trouble with physics: The rise of string theory, the fall of a science, and what comes next*. Houghton Mifflin Harcourt. [aMJ]

Sobel, D. M. & Kirkham, N. Z. (2006) Blickets and babies: The development of causal reasoning in toddlers and infants. *Developmental Psychology* 42(6):1103–15. [AG]

Sobel, D. M., Tenenbaum, J. B. & Gopnik, A. (2004) Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science* 28(3):303–33. [AG, aMJ]

Soltani, A. & Wang, X.-J. (2010) Synaptic computation underlying probabilistic inference. *Nature Neuroscience* 13(1):112–19. [aMJ]

Spencer, J. P. & Perone, S. (2008) Defending qualitative change: The view from dynamical systems theory. *Child Development* 79:1639–47. [GWJ]

Spencer, J. P., Perone, S. & Johnson, J. S. (2009) The dynamic field theory and embodied cognitive dynamics. In: *Toward a unified theory of development: Connectionism and dynamic systems theory reconsidered*, ed. J. P. Spencer, M. S. Thomas & J. L. McClelland, pp. 86–118. Oxford University Press. [aMJ]

Spencer, J. P., Perone, S., Smith, L. B. & Samuelson, L. K. (2011) The process behind the "suspicious coincidence": Using space and time to learn words. *Psychological Science*. doi: 10.1177/0956797611413934 [GWJ]

Spencer, J. P., Thomas, M. S. & McClelland, J. L. (2009) *Toward a unified theory of development: Connectionism and dynamic systems theory re-considered*. Oxford University Press. [GWJ]

Sperber, D. & Hirschfeld, L. A. (2003) The cognitive foundations of cultural stability and diversity. *Trends in Cognitive Sciences* 8:40–46. [aMJ]

Spirtes, P., Glymour, C. & Scheines, R. (2000) *Causation, prediction, and search*, 2nd edition. (original edition published in 1993) MIT Press. [AG, rMJ]

Stankiewicz, B. J., Legge, G. E., Mansfield, J. S. & Schlicht, E. J. (2006) Lost in virtual space: Studies in human and ideal spatial navigation. *Journal of Experimental Psychology: Human Perception and Performance* 32:688–704. [aMJ]

Stanovich, K. E. & West, R. F. (1998) Individual differences in rational thought. / precode> *Journal of Experimental Psychology: General* 127(2):161–88. [ABM]

Steele, C. M. (1988) The psychology of self-affirmation: Sustaining the integrity of the self. In: *Advances in Experimental Social Psychology*, vol. 21, ed. L. Berkowitz, pp. 261–302. Academic Press. [ELU]

Sternberg, S. (1966) High-speed scanning in human memory. *Science* 153:652–54. [rMJ]

Steyvers, M., Lee, M. D. & Wagenmakers, E.-J. (2009) A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology* 53:168–79. [aMJ]

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J. & Blum, B. (2003) Inferring causal networks from observations and interventions. *Cognitive Science* 27:453–89. [IB, PMF, aMJ]

Stigler, S. M. (1961) The economics of information. *Journal of Political Economy* 69:213–25. [aMJ]

Stocker, A. A. & Simoncelli, E. P. (2006) Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience* 9(4):578–85. [JSB]

Teller, D. Y. (1984) Linking propositions. *Vision Research* 10:1233–46. [DKS]

Tenenbaum, J. B. & Griffiths, T. L. (2001) Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences* 24(4):629–40. [PMF, aMJ, ABM, DJN]

Tenenbaum, J. B. & Griffiths, T. L. (2001) Structure learning in human causal induction. In: *Advances in neural information processing systems, vol. 13*, ed. T. K. Leen, T. G. Dietterich & V. Tresp, pp. 59–65. MIT Press. [KJH]

Tenenbaum, J. B., Griffiths, T. L. & Kemp, C. (2006) Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences* 10:309–18. [aMJ]

Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. (2011) How to grow a mind: Statistics, structure, and abstraction. *Science* 331(6022):1279–85. [BLA, NC, GWJ]

Tetlock, P. E. (2002) Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. *Psychological Review* 109:451–71. [ELU]

Tetlock, P. E., Kristel, O., Elson, B., Green, M. & Lerner, J (2000) The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology* 78:853–70. [ELU]

Tetlock, P. E., Visser, P., Singh, R., Polifroni, M., Elson, B., Mazzocco, P. & Rescober, P. (2007) People as intuitive prosecutors: The impact of social control motives on attributions of responsibility. *Journal of Experimental Social Psychology* 43:195–209. [ELU]

Thagard, P. (1989) Explanatory coherence. *Behavioral and Brain Sciences* 12:435–502. [aMJ]

Thaler, R. H. & Sunstein, C. R. (2008) *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press. [aMJ]

Thibaux, R. & Jordan, M. I. (2007) Hierarchical beta processes and the Indian buffet process. In: *Proceedings of the Tenth Conference on Artificial Intelligence and Statistics (AISTATS)*, ed. M. Meila & X. Shen. Society for Artificial Intelligence and Statistics. (Online Publication). Available at: http://www.stat.umn.edu/%7Eaistat/proceedings/start.htm [aMJ]

Thomas, M. S. C. & McClelland, J. L. (2008) Connectionist models of cognition. In: *The Cambridge handbook of computational psychology*, ed. R. Sun, pp. 23–58. Cambridge University Press. [DKS]

Thompson, P., Brooks, K. & Hammett, S. T. (2006) Speed can go up as well as down at low contrast: Implications for models of motion perception. *Vision Research* 46(6–7):782–86. [JSB]

Thompson-Schill, S., Ramscar, M. & Chrysikou, M. (2009) Cognition without control: When a little frontal lobe goes a long way. *Current Directions in Psychological Science* 8:259–63. [aMJ]

Tikhonov, A. N. & Arsenin, V. Y. (1977) *Solutions of ill-posed problems*. W. H. Winston. [SE]

Tinbergen, N. (1963) On aims and methods of ethology. *Zeitschrift für Tierpsychologie* 20:410–33. [LA-S]

Tooby, J. & Cosmides, L. (2005) Conceptual foundations of evolutionary psychology. In: *The handbook of evolutionary psychology*, ed. D. M. Buss, pp. 5–67. Wiley. [aMJ]

Tooby, J., & Cosmides, L. (1992) The psychological foundations of culture. In: *The adapted mind: Evolutionary psychology and the generation of culture,* J. H. Barkow, L. Cosmides, & J. Tooby, pp. 19–136. Oxford University Press. [LA-S]

Tooby, J., Cosmides, L., Sell, A., Lieberman, D. & Sznycer, D. (2008) Internal regulatory variables and the design of human motivation: A computational and

evolutionary approach. In: *Handbook of approach and avoidance motivation*, ed. A. Elliot, pp. 251–71. Erlbaum. [DP]

Turing, A. (1950) Computing, machinery and intelligence. *Mind* 49:433–60. [ABM]

Tversky, A. & Kahneman, D. (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185:1124–31. [aMJ]

Uhlmann, E. L. & Cohen, G. L. (2005) Constructed criteria: Redefining merit to justify discrimination. *Psychological Science* 16:474–80. [ELU]

Vul, E., Frank, M. C., Alvarez, G. A. & Tenenbaum, J. B. (2009) Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. *Advances in Neural Information Processing Systems* 22:1955–63. [aMJ]

Vul, E., Goodman, N. D., Griffiths, T. L. & Tenenbaum, J. B. (2009a) One and done: Optimal decisions from very few samples. In: *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, ed. N. Taatgen & H. van Rijn, pp. 148–53. Erlbaum. [NC]

Vul, E., Hanus, D. & Kanwisher, N. (2009b) Attention as inference: Selection is probabilistic, responses are all-or-none samples. *Journal of Experimental Psychology: General* 138:546–60. [NC]

Waelti, P., Dickinson, A. & Schultz, W. (2001) Dopamine responses comply with basic assumptions. *Nature* 412:43–48. [IB]

Waldmann, M. R. & Holyoak, K. J. (1992) Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General* 121:222–36. [KJH]

Waldmann, M. R. & Martignon, L. (1998) A Bayesian network model of causal learning. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, ed. M. A. Gernsbacher & S. J. Derry, pp. 1102–107. Erlbaum. [KJH]

Wallsten, T. S. (1971) Subjectively expected utility theory and subjects' probability estimates: Use of measurement-free techniques. *Journal of Experimental Psychology* 88:31–40. [MS]

Walsh, C. R. & Sloman, S. A. (2008) Updating beliefs with causal models: Violations of screening off. In: *Memory and Mind: A festschrift for Gordon H. Bower*, ed. M. A. Gluck, J. R. Anderson & S. M. Kosslyn, pp. 345–57. Erlbaum. [NC]

Wason, P. (1960) On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology* 12:129–40. [DJN]

Wasserman, L. (2003) *All of statistics: A concise course in statistical inference*. Springer Texts in Statistics. Springer. [SE]

Watson, J. B. (1913) Psychology as the behaviorist views it. *Psychological Review* 20:158–77. [aMJ]

Weiss, Y., Simoncelli, E. P. & Adelson, E. H. (2002) Motion illusions as optimal percepts. *Nature Neuroscience* 5(6):598–604. [JSB]

Wertheimer, M. (1923/1938) Laws of organization in perceptual forms. In: *A source book of Gestalt psychology*, ed. & trans. W. Ellis, pp. 71–88. Routledge & Kegan Paul. (Original work published in 1923). [aMJ]

Wilder, M. H., Jones, M. & Mozer, M. C. (2009) Sequential effects reflect parallel learning of multiple environmental regularities. *Advances in Neural Information Processing Systems* 22:2053–61. [MH, arMJ]

Wilson-Mendenhall, C. D., Barrett, L. F., Simmons, W. K. & Barsalou, L. W. (2011) Grounding emotion in situated conceptualization. *Neuropsychologia* 49:1105–27. [LWB]

Woit, P. (2006) *Not even wrong: The failure of string theory and the search for unity in physical law*. Basic Books. [aMJ]

Wolpert, D. (1996) The lack of a priori distinctions between learning algorithms. *Neural Computation* 8:1341–90. [aMJ]

Wood, J. N. & Grafman, J. (2003) Human prefrontal cortex: Processing and representational perspectives. *Nature Reviews: Neuroscience* 4:129–47. [aMJ]

Xu, F. & Tenenbaum, J. B. (2007a) Sensitivity to sampling in Bayesian word learning. *Developmental Science* 10(3):288–97. [GWJ, DJN]

Xu, F. & Tenenbaum, J. B. (2007b) Word learning as Bayesian inference. *Psychological Review* 114(2):245–72. [GWJ, aMJ, TTR]

Yamauchi, T. & Markman, A. B. (1998) Category learning by inference and classification. *Journal of Memory and Language* 39:124–48. [aMJ]

Yang, L.-X. & Lewandowsky, S. (2003) Context-gated knowledge partitioning in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29:663–79. [DKS]

Yang, L.-X. & Lewandowsky, S. (2004) Knowledge partitioning in categorization: Constraints on exemplar models. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30:1045–64. [DKS]

Yeh, W. & Barsalou, L. W. (2006) The situated nature of concepts. *American Journal of Psychology* 119:349–84. [LWB]

Yu, A. & Cohen, J. (2008) Sequential effects: Superstition or rational behavior? *Advances in Neural Information Processing Systems* 21:1873–80. [aMJ]

Yuille, A. & Kersten, D. (2006) Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences* 10:301–308. [NC]

Zambrano, E. (2005) Testable implications of subjective expected utility theory. *Games and Economic Behavior* 53:262–68. [MS]