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,

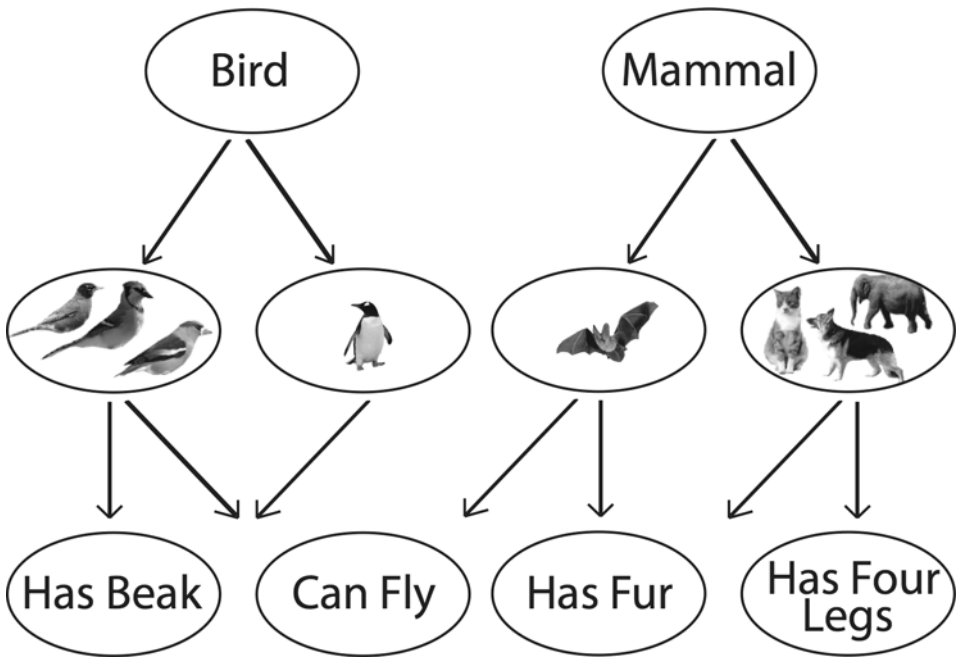

Figure 1. Simplified generative model based on Anderson's (1991b) rational model of categorization. Upper and lower nodes represent observable variables (category labels and features, respectively). Middle nodes represent clusters, which correspond to latent variables. Judgment of category membership 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