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

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

The imaginary fundamentalists: The unshocking truth about Bayesian cognitive science

doi:10.1017/S0140525X11000239

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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