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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Lu et al. (2008b) carried the project a step further, implementing and testing a 2×2 design of Bayesian models of learning causal strength: the two likelihood functions crossed with two priors (uninformative vs. a preference for sparse and strong causes). When compared to human data, model comparisons established that human causal learning is better explained by the assumptions underlying the power PC theory, rather than by those underlying

associative models. The sparse-and-strong prior accounted for subtle interactions involving generative and preventive causes that could not be explained by uninformative priors.

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

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

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

Gavin W. Jenkins, Larissa K. Samuelson, and John P. Spencer
Department of Psychology and Delta Center, University of Iowa, Iowa City, IA 52242-1407.
gavin-jenkins@uiowa.edu   larissa-samuelson@uiowa.edu
john-spencer@uiowa.edu
http://www.psychology.uiowa.edu/people/gavin_jenkins
http://www.psychology.uiowa.edu/people/larissa_samuelson
http://www.psychology.uiowa.edu/people/john_spencer

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

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

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

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

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

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

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

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

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

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

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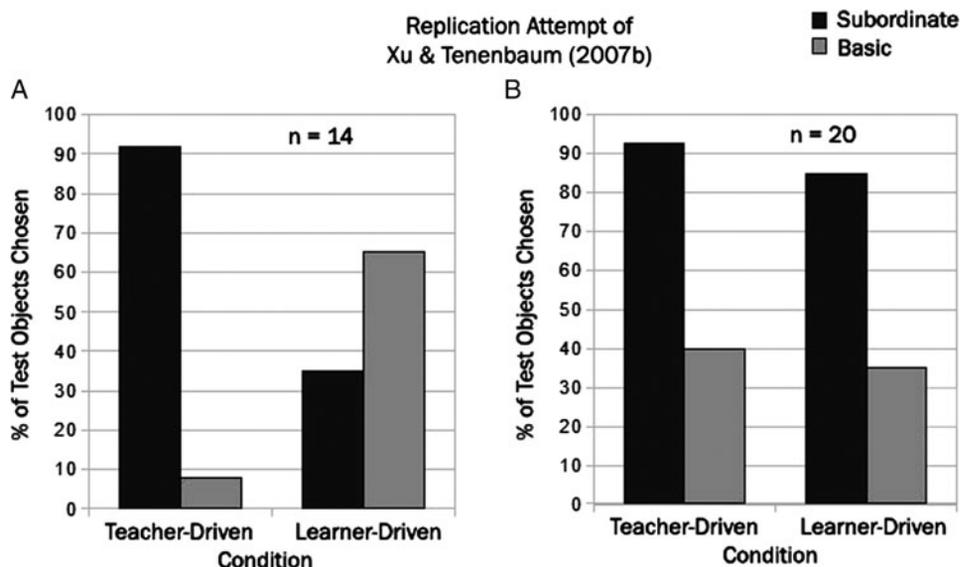

Figure 1 (Jenkins et al.). Replication attempt by Xu and Tenenbaum (2007a). A: Xu and Tenenbaum's results. B: Our exact replication attempt.

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

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

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

In praise of Ecumenical Bayes

Michael D. Lee

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Cognitive systems optimize energy rather than information

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

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

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

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

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

The idea that energy-minimization is an important constraint on cognitive processing is implicit in the focus on efficient computational procedures. We do not suppose that the metabolic cost of cognition is completely invariant of the type of thinking that people are engaged in, but marginal changes in metabolic

rates attributed to different types of cognition pale in comparison to the metabolic cost of simply keeping the brain running. Thus, the time taken by a process is a good proxy for energy conservation. On this view, for example, habits minimize energy, because they allow a complex behavior to be carried out quickly (e.g., Logan 1988; Schneider & Shiffrin 1977).

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

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

Enlightenment grows from fundamentals

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

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

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

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

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

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

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

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

Viewed more broadly, the principle of "enlightenment growing from fundamentals" is applicable beyond Bayesian modelling; our last example is therefore an inversion. We suggest that J&L understate the importance of computational considerations in good process modelling. For instance, one of their key examples comes from Sakamoto et al. (2008), who consider mechanistic models of category learning. That paper might be characterized as a "fundamentalist" work in process modelling, insofar as it gives no consideration to the computational level issues that pertain to their choice of learning problem. As consequence of this "process fundamentalism," the "rational" model that paper employs is not actually a rational model. It is highly mis-specified

for the problem of learning time-inhomogeneous categories. In recent work (Navarro & Perfors 2009), we discuss this concern and introduce extensions to the experimental framework aimed at highlighting the computational considerations involved; at present, we are working on model development to build on this. However, the goal in our work is *not* to deny the importance of process, but to learn which aspects of human behaviour are attributable to computational level issues and which aspects reflect process limitations. In this case, that goal is met by building on fundamental work on the process level (i.e., Sakamoto et al.'s 2008 paper) and adding computational considerations. In general, attaining the goal of "enlightened" research is possible only if fundamentals on both levels are taken seriously – if researchers deny neither psychological mechanism *nor* ideal computation.

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

The illusion of mechanism: Mechanistic fundamentalism or enlightenment?

Dennis Norris
MRC Cognition and Brain Sciences Unit, Cambridge CB2 7EF, United Kingdom.
dennis.norris@mrc-cbu.cam.ac.uk
http://www.mrc-cbu.cam.ac.uk/people/dennis.norris

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

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

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

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

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

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

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

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

fully aware of this deliberate limitation, and it is hard to see that any of the insights from connectionist modelling are undermined by this simplification.

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

Reverse engineering the structure of cognitive mechanisms

David Pietraszewski and Annie E. Wertz

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

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

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

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

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

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

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

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

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

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

Taking the rationality out of probabilistic models

Bob Rehder

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

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

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

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

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

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

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

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

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

Distinguishing literal from metaphorical applications of Bayesian approaches

Timothy T. Rogers and Mark S. Seidenberg

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

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

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

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

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

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

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

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

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

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

Bayesian computation and mechanism: Theoretical pluralism drives scientific emergence

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

[a]Department of Psychological Sciences, The University of Melbourne, Melbourne, VIC 3010, Australia; [b]School of Psychology, The University of Western Australia, Crawley, WA 6009, Australia.
dsewell@unimelb.edu.au   daniel.little@unimelb.edu.au
lewan@psy.uwa.edu.au
http://www.psych.unimelb.edu.au/people/staff/SewellD.html
http://www.psych.unimelb.edu.au/research/labs/knowlab/index.html
http://www.cogsciwa.com/

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

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

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

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

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

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

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

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

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

Is everyone Bayes? On the testable implications of Bayesian Fundamentalism

Maarten Speekenbrink and David R. Shanks
Division of Psychology and Language Sciences, University College London, London WC1E 6BT, United Kingdom.
m.speekenbrink@ucl.ac.uk   **d.shanks@ucl.ac.uk**
http://www.psychol.ucl.ac.uk/m.speekenbrink
http://www.psychol.ucl.ac.uk/david.shanks/Shanks.html

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

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

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

restrictions on the response sequence for a given reward sequence.

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

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

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

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

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

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

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

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

Post hoc rationalism in science

Eric Luis Uhlmann

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

eric.luis.uhlmann@gmail.com
www.socialjudgments.com

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

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

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

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

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

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

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

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

Authors' Response

Pinning down the theoretical commitments of Bayesian cognitive models

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

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

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

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

R1. Introduction

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

We referred to the approach of Bayesian modeling without consideration of mechanism as *Bayesian Fundamentalism.* We went on to advocate an approach we labeled *Bayesian Enlightenment*, in which elements of a Bayesian model are given a psychological interpretation, by addressing how the learner's hypotheses are represented, where they come from, what the learner's goals are, and how inference is carried out. Although several commentators argue for further challenges or shortcomings, no serious challenge was offered to the conclusion that, at the least, Bayesian models need this type of grounding. Primarily, the commentaries serve to reinforce, in various ways, the idea that it is critical to be clear on the psychological commitments and explanatory contributions of cognitive models.

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

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

R2. Theoretical status of Bayesian models

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

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

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

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

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

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

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

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

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

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

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

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

R3. Bayesian models as computational-level theories

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

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

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

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

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

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

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

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

Despite these dangers of a purely computational framing, the mathematical framework of probabilistic inference does have advantages that are not dependent on specification of psychological mechanism. One important principle is the idea that the brain somehow tracks uncertainty or variability in environmental parameters, rather than just point estimates. This insight has been influential in areas such as causal induction (**Holyoak & Lu**), but it is also not new (e.g., Fried & Holyoak 1984). Another strength of the Bayesian framework is that it offers natural accounts of how information can be combined from multiple sources, and in particular, how people can incorporate rich prior knowledge into any learning task (**Heit & Erickson**). However, this potential is unrealized if there is no independent assessment of what that prior knowledge is. Instead, the expressive flexibility of Bayesian models becomes a weakness, as it makes them unfalsifiable (**Bowers & Davis**; **Danks & Eberhardt**; **Glymour**; **Rogers & Seidenberg**). In some

cases, the assumptions of a Bayesian model are demonstrably false, as **Rehder** points out in the case of mutual exclusivity in categorization models, but even then the conclusion is unclear. Was the failed assumption theoretically central to the model, or just an implementation detail of a more general theory that might still hold? If so, what is that general theory that remains after the particular assumptions about the hypothesis space are set aside? Under a computational-level stance, all that is left is the claim of optimality with respect to an unspecified environment, which is no theory at all.

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

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

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

R4. Bayesian models as mechanistic theories

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

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

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

Borsboom et al. write that the counting rule "seems just about right," and perhaps it is neurologically correct in some cases (e.g., Gold & Shadlen 2001). However, even if this is true, the counting rule is not where the hard work of cognition is being done (**Anderson**). Likewise, although we fully agree with **Chater et al.** that interesting behavior can emerge from simple rules, it is not the counting rule that is responsible for this emergence; it is the structure of the hypothesis space. As **Gopnik** points out, "The central advance has not been Bayes' law itself, but the ability to formulate structured representations, such as causal graphical models, or 'Bayes nets' (Pearl 2000; Spirtes et al. 2000), or hierarchical causal models, category hierarchies or grammars." Thus, as argued above, the hypothesis space is where the interesting

psychology lies in most Bayesian models. If we consider it a core assumption of a model, then the model makes meaningful, testable predictions. Although most Bayesian models cast their hypothesis spaces as components of rational analysis and not psychological entities (or else are noncommittal), one can certainly postulate them as psychological representations (**Heit & Erickson**). This is one way in which Bayesian models can potentially make important contributions. Of course, the assumption of optimal inference with respect to the assumed representation could be, and probably often is, wrong (**Uhlmann**), but the important point for present purposes is that this claim becomes testable once the learner's representations and goals are pinned down as psychological commitments.

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

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

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

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

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

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

clarifies its contribution but also reveals important questions it fails to address. This is not an unavoidable weakness of the Bayesian approach, but it does suggest that applying more scrutiny to the assumptions of Bayesian models would start them on a path toward providing more complete psychological explanations.

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

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

R5. Prospects for integration

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

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

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

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

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

Second, one can study algorithms that approximate optimal inference (e.g., Daw & Courville 2007; Sanborn et al. 2010a). Under this approach, rational and mechanistic considerations enter at different levels of analysis, and the aim is to understand how they constrain each other.

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

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

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

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

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

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

R6. Conclusions

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

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

References/Jones & Love: Bayesian Fundamentalism or Enlightenment?

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Lee, M. D. & Sarnecka, B. W. (in press) Number knower-levels in young children: Insights from a Bayesian model. *Cognition*. [MDL]

Lerner, J. S. & Tetlock, P. E. (1999) Accounting for the effects of accountability. *Psychological Bulletin* 125:255–75. [ELU]

Lewandowsky, S., Griffiths, T. L. & Kalish, M. L. (2009) The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science* 33:969–98. [DKS]

Lewandowsky, S. & Heit, E. (2006) Some targets for memory models. *Journal of Memory and Language* 55:441–46. [EH]

Lewandowsky, S., Kalish, M. & Ngang, S. K. (2002) Simplified learning in complex situations: Knowledge partitioning in function learning. *Journal of Experimental Psychology: General* 131:163–93. [DKS]

Lewandowsky, S., Roberts, L. & Yang, L.-X. (2006) Knowledge partitioning in categorization: Boundary conditions. *Memory and Cognition* 34:1676–88. [DKS]

Little, D. R. & Lewandowsky, S. (2009) Beyond nonutilization: Irrelevant cues can gate learning in probabilistic categorization. *Journal of Experimental Psychology: Human Perception and Performance* 35:530–50. [DKS]

Logan, G. D. (1988) Toward an instance theory of automaticity. *Psychological Review* 95:492–527. [ABM]

Lord, C. G., Ross, L. & Lepper, M. R. (1979) Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology* 37:2098–109. [ELU]

Louwerse, M. M. & Connell, L. (2011) A taste of words: Linguistic context and perceptual simulation predict the modality of words. *Cognitive Science* 35:381–98. [LWB]

Love, B. C. (2002) Comparing supervised and unsupervised category learning. *Psychonomic Bulletin and Review* 9:829–35. [aMJ]

Love, B. C. (2005) Environment and goals jointly direct category acquisition. *Current Directions in Psychological Science* 14:195–99. [rMJ]

Lu, H., Rojas, R. R., Beckers, T. & Yuille, A. L. (2008a) Sequential causal learning in humans and rats. In: *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society*, ed. B. C. Love, K. McRae & V. M. Sloutsky, pp. 195–88. Cognitive Science Society. [KJH]

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W. & Holyoak, K. J. (2008b) Bayesian generic priors for causal learning. *Psychological Review* 115:955–82. [KJH]

Lucas, C., Gopnik, A. & Griffiths, T. (2010) Developmental differences in learning the form of causal relationships. In: *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, ed. S. Ohlsson & R. Catrambone, pp. 2852–57. Cognitive Science Society. [AG]

Lucas, C., Griffiths, T. L., Xu, F. & Fawcett, C. (2009) A rational model of preference learning and choice prediction by children. *Advances in Neural Information Processing Systems* 21:985–92. [aMJ]

Luce, R. D. (1963) Detection and recognition. In: *Handbook of mathematical psychology*, ed. R. D. Luce, R. R. Bush & E. Galanter, pp. 103–89. John Wiley. [aMJ]

Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. (2006) Bayesian inference with probabilistic population codes. *Nature Neuroscience* 9:1432–38. [DB]

Machamer, P., Darden, L. & Craver, C. F. (2000) Thinking about mechanisms. *Philosophy of Science* 67:1–25. [MH]

Machery, E. & Barrett, C. (2006) Debunking adapting minds. *Philosophy of Science* 73:232–46. [aMJ]

MacKay, D. (2002) *Information theory, inference, and learning algorithms*. Cambridge University Press. [NC]

Maloney, L. T. & Mamassian, P. (2009) Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Visual Neuroscience* 26:147–55. [BR]

Maloney, L. T. & Zhang, H. (2010) Decision-theoretic models of visual perception and action. *Vision Research* 50:2362–74. [BR]

Maloney, L. T. & Zhang, H. (2010) Decision-theoretic models of visual perception and action. *Vision Research* 50:2362–74. [BR]

Malt, B. C. & Smith, E. E. (1984) Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior* 23:250–69. [BKH]

Manning, C. & Schütze, H. (1999) *Foundations of statistical natural language processing*. MIT Press. [NC]

Marcus, G. F. (1998) Rethinking eliminative connectionism. *Cognitive Psychology* 37:243–82. [aMJ]

Marcus, G. F. (2008) *Kluge: The haphazard construction of the human mind*. Houghton Mifflin. [aMJ]

Markman, A. B. & Ross, B. H. (2003) Category use and category learning. *Psychological Bulletin* 129:592–615. [aMJ]

Marr, D. (1982) *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman. [BLA, LWB, NC, MH, arMJ, ABM, DN, DP]

Marr, D. (1982/2010) *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman/MIT Press. (Original work published in 1982; 2010 reprint edition by MIT Press). [DKS]

Marroquin, J., Mitter, S. & Poggio, T. (1987) Probabilistic solution of ill-posed problems in computational vision. *Journal of the American Statistical Association* 82:76–89. [SE]

Mayr, E. (1982) *The growth of biological thought: Diversity, evolution, and inheritance*. Harvard University Press. [aMJ]

McClelland, J. L. (1998) Connectionist models and Bayesian inference. In: *Rational models of cognition*, ed. M. Oaksford & N. Chater, pp. 21–53. Oxford University Press. [NC, EH]

McClelland, J. L. (2010) Emergence in cognitive science. *Topics in Cognitive Science* 2:751–70. [NC]

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S. & Smith, L. B. (2010) Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences* 14:348–56. [DKS]

McClelland, J. L. & Chappell, M. (1998) Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review* 105:724–60. [aMJ]

McClelland, J. L. & Patterson, K. (2002) Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences* 6(11):465–72. [TTR]

McClelland, J. L. & Rumelhart, D. E. (1988) *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. MIT Press. [IB]

McClelland, J. L., Rumelhart, D. E. & the PDP Research Group. (1986) *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2: Psychological and biological models*. MIT Press. [aMJ]

McCulloch, W. & Pitts, W. (1943) A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 7:115–33. [aMJ]

McKenzie, C. R. M. & Mikkelsen, L. A. (2007) A Bayesian view of covariation assessment. *Cognitive Psychology* 54:33–61. [aMJ]

McNamara, J. M. & Houston, A. I. (2009) Integrating function and mechanism. *Trends in Ecology and Evolution* 24:670–75. [aMJ]

Michaels, C. F. & Carello, C. (1981) *Direct perception*. Prentice-Hall. [aMJ]

Miller, E. K. & Cohen, J. D. (2001) An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience* 24:167–202. [aMJ]

Miller, G. A. (2003) The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences* 7:141–44. [aMJ]

Minsky, M. & Papert, S. A. (1969) *Perceptrons: An introduction to computational geometry*. MIT Press. [aMJ]

Mortimer, D., Feldner, J., Vaughan, T., Vetter, I., Pujic, Z., Rosoff, W. J., Burrage, K., Dayan, P., Richards, L. J. & Goodhill, G. J. (2009) Bayesian model predicts the response of axons to molecular gradients. *Proceedings of the National Academy of Sciences USA* 106:10296–301. [aMJ]

Mozer, M. C., Pashler, H. & Homaei, H. (2008) Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science* 32:1133–47. [aMJ]

Murphy, G. L. (1993) A rational theory of concepts. *Psychology of Learning and Motivation* 29:327–59. [aMJ]

Murphy, G. L. & Medin, D. L. (1985) The role of theories in conceptual coherence. *Psychological Review* 92:289–316. [EH]

Murphy, G. L. & Ross, B. H. (2007) Use of single or multiple categories in category-based induction. In: *Inductive reasoning: Experimental, developmental, and computational approaches*, ed. A. Feeney & E. Heit, pp. 205–25. Cambridge Press. [BKH, rMJ]

Murphy, G. L. & Ross, B. H. (2010) Category vs. object knowledge in category-based induction. *Journal of Memory and Language* 63:1–17. [BKH]

Mussa-Ivaldi, F. A. & Giszter, S. F. (1992) Vector field approximation: A computational paradigm for motor control and learning. *Biological Cybernetics* 67:491–500. [SE]

Navarro, D. J. (2010) Learning the context of a category. In: *Advances in neural information processing systems, vol. 23*, ed. J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel & A. Culotta, pp. 1795–803. MIT Press. [DKS]

Navarro, D. J., Dry, M. J. & Lee, M. D. (in press) Sampling assumptions in inductive generalization. *Cognitive Science*. [DJN]

Navarro, D. J. & Perfors, A. F. (2009) Learning time-varying categories. In: *Proceedings of the 31st Annual Conference of the Cognitive Science Society, Austin, TX*, ed. N. Taatgen, H. van Rijn, L. Schomaker & J. Nerbonne, pp. 419–24. Cognitive Science Society. [DJN]

Navarro, D. J. & Perfors, A. F. (2011) Hypothesis generation, sparse categories and the positive test strategy. *Psychological Review* 118:120–34. [DJN]

Neal, R. M. (1992) Connectionist learning of belief networks. *Artificial Intelligence* 56:71–113. [NC]

Neisser, U. (1967) *Cognitive psychology*. Appleton-Century-Crofts. [DKS]

Nersessian, N. J. (1986) A cognitive-historical approach to meaning in scientific theories. In: *The process of science: Contemporary philosophical approaches to understanding scientific practice*, ed. N. J. Nersessian. Martinus Nijhoff. [aMJ]

Neuhoff, J. G. (2001) An adaptive bias in the perception of looming auditory motion. *Ecological Psychology* 13:87–110. [LA-S]

Newell, A. & Simon, H. A. (1972) *Human problem solving*. Prentice-Hall. [aMJ]

Newell, B. R., Paton, H., Hayes, B. K. & Griffiths, O. (2010) Speeded induction under uncertainty: The influence of multiple categories and feature conjunctions. *Psychonomic Bulletin and Review* 17:869–74. [BKH]

Newport, E. L. (1990) Maturational constraints on language learning. *Cognitive Science* 14:11–28. [aMJ]

Norman, K. A. & O'Reilly, R. C. (2003) Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. *Psychological Review* 110(4):611–46. [EH, TTR]

Norton, M. I., Vandello, J. A. & Darley, J. M. (2004) Casuistry and social category bias. *Journal of Personality and Social Psychology* 87:817–31. [ELU]

Nosofsky, R. M. (1986) Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115:39–57. [DKS]

Nosofsky, R. M., Palmeri, T. J. & Mckinley, S. C. (1994) Rule-plus-exception model of classification learning. *Psychological Review* 104:266–300. [aMJ]

O'Reilly, R. C. & Norman, K. A. (2002) Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends in Cognitive Sciences* 6(12):505–10. [TTR]

Oaksford, M. & Chater, N. (1994) A rational analysis of the selection task as optimal data selection. *Psychological Review* 101:608–31. [NC, aMJ, DJN]

Oaksford, M. & Chater, N. (1998a) An introduction to rational models of cognition. In: *Rational models of cognition*, ed. M. Oaksford & N. Chater, pp. 1–18. Oxford University Press. [aMJ]

Oaksford, M. & Chater, N., ed. (1998b) *Rational models of cognition*. Oxford University Press. [NC]

Oaksford, M. & Chater, N. (2003) Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin and Review* 10:289–318. [NC]

Oaksford, M. & Chater, N. (2007) *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press. [DB, DD, arMJ]

Oaksford, M. & Chater, N. (2010) Conditionals and constraint satisfaction: Reconciling mental models and the probabilistic approach? In: *Cognition and conditionals: Probability and logic in human thinking*, ed. M. Oaksford & N. Chater, pp. 309–34. Oxford University Press. [NC, aMJ]

Öhman, A., Flykt, A. & Esteves, F. (2001) Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General* 130:466–78. [LA-S]

Olshausen, B. A. & Field, D. J. (1996) Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. *Nature* 381:607–9. [TTR]

Oppenheim, R. W. (1981) Ontogenetic adaptations and retrogressive processes in the development of the nervous system and behavior. In: *Maturation and development: Biological and psychological perspectives*, ed. K. J. Connolly & H. F. R. Prechtl, pp. 73–108. International Medical. [aMJ]

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A. & Shafir, E. (1990) Category-based induction. *Psychological Review* 97:185–200. [BKH]

Papadopoulos, C., Hayes, B. K. & Newell, B. R. (2011) Non-categorical approaches to feature prediction with uncertain categories. *Memory and Cognition* 39:304–18. [BKH]

Pavlov, I. P. (1927) *Conditioned reflexes*. Oxford University Press. [IB]

Payne, J. W., Bettman, J. R. & Johnson, E. J. (1993) *The adaptive decision maker*. Cambridge University Press. [ABM]

Pearl, J. (1988) *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann. [NC, KJH]

Pearl, J. (2000) *Causality: Models, reasoning, and inference*. Cambridge University Press. [DB, AG, rMJ]

Perales, J. C. & Shanks, D. R. (2007) Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin and Review* 14:577–96. [KJH]

Perfors, A. (2011) Memory limitations alone do not lead to over-regularization: An experimental and computational investigation. In: *Proceedings of the 33rd Annual Conference of the Cognitive Science Society, Boston, MA*, ed. L. Carlson, C. Hoelscher & T. F. Shipley, pp. 3274–79. Cognitive Science Society. [DJN]

Perfors, A. & Tenenbaum, J. (2009) Learning to learn categories. In: *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, ed. N. Taatgen, H. van Rijn, L. Schomaker & J. Nerbonne, pp. 136–41. Cognitive Science Society. [DJN]

Perfors, A., Tenenbaum, J. B. & Regier, T. (2011) The learnability of abstract syntactic principles. *Cognition* 118:306–38. [NC]

Perfors, A., Tenenbaum, J. & Wonnacott, E. (2010) Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language* 37:607–42. [NC, DJN]

Perrigo, G., Belvin, L. & Vom Saal, F. S. (1991) Individual variation in the neural timing of infanticide and parental behavior in male house mice. *Physiology and Behavior* 50:287–96. [DP]

Perrigo, G., Belvin, L. & Vom Saal, F. S. (1992) Time and sex in the male mouse: Temporal regulation of infanticide and parental behavior. *Chronobiology International* 9:421–33. [DP]

Perry, L. K., Cook, S. W. & Samuelson L. K. (in preparation) An exploration of context, task, and stimuli effects on similarity perception. [GWJ]

Pezzulo, G., Barsalou, L. W., Cangelosi, A., Fischer, M. A., McRae, K. & Spivey, M. (2011) The mechanics of embodiment: A dialogue on embodiment and computational modeling. *Frontiers in Cognition* 2(5):1–21. [LWB]

Pinker, S. (1995) *The language instinct: How the mind creates language*. Perennial. [aMJ]

Pinker, S. (2002) *The blank slate: The modern denial of human nature*. Viking. [aMJ]

Pinker, S. & Prince, A. (1988) On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28:73–193. [aMJ]

Pitt, M. A., Myung, I. J. & Zhang, S. (2002) Toward a method of selecting among computational models of cognition. *Psychological Review* 109:472–91. [aMJ]

Plaut, D. C. (1995) Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology* 17:291–321. [IB]

Plaut, D. C., McClelland, J. L., Seidenberg, M. S. & Patterson, K. (1996) Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review* 103:56–115. [aMJ]

Poggio, T. (1990) A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology* 55:899–910. [SE]

Pollack, J. B. (1990) Recursive distributed representations. *Artificial Intelligence* 46:77–105. [aMJ]

Pothos, E. M. & Chater, N. (2002) A simplicity principle in unsupervised human categorization. *Cognitive Science* 26:303–43. [aMJ]

Pronin, E., Gilovich, T. & Ross, L. (2004) Objectivity in the eye of the beholder: Perceptions of bias in self versus others. *Psychological Review* 111:781–99. [ELU]

Pronin, E., Lin, D. Y. & Ross, L. (2002) The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin* 28:369–81. [ELU]

Pyszczynski, T. & Greenberg, J. (1987) Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. In: *Advances in experimental social psychology, vol. 20*, ed. L. Berkowitz, pp. 297–340. Academic Press. [ELU]

Rachman, S. (1997) The evolution of cognitive behaviour therapy. In: *Science and practice of cognitive behaviour therapy*, ed. D. Clark, C. G. Fairburn & M. G. Gelder, pp. 1–26. Oxford University Press. [aMJ]

Raiffa, H. & Schlaifer, R (1961) *Applied statistical decision theory*. Harvard University Press. [aMJ]

Ramscar, M., Yarlett, D., Dye, M., Denny, K. & Thorpe, K. (2010) The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science* 34:1–49. [aMJ]

Ravi, S. & Knight, K. (2009) Minimized models for unsupervised part-of-speech tagging. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, ed. K.-Y. Su, pp. 504–12. Association for Computational Linguistics. Available at: http://www.aclweb.org/anthology-new/P/P09/P09-1057.pdf [aMJ]

Rehder, B. (2009) Causal-based property generalization. *Cognitive Science* 33:301–43. [BR]

Rehder, B. & Burnett, R. (2005) Feature inference and the causal structure of categories. *Cognitive Psychology* 50:264–314. [NC, BR]

Rehder, B. & Kim, S. (2010) Causal status and coherence in causal-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 36:1171–206. [BR]

Rescorla, R. A. & Wagner, A. R. (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: *Classical conditioning II: Current theory and research*, ed. A. H. Black & W. F. Prokasy, pp. 64–99. Appleton-Century-Crofts. [IB, NC, KJH, rMJ]

Ricci, G. & Levi-Civita, T. (1900) Méthodes de calcul différentiel absolu et leurs applications [Methods of absolute differential calculus and their applications]. *Mathematische Annalen* 54(1–2):125–201. [aMJ]

Rips, L. J. (1990) Reasoning. *Annual Review of Psychology* 41:321–53. [EH]

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges J. R. & Patterson, K. (2004) The structure and deterioration of semantic memory: A computational and neuropsychological investigation. *Psychological Review* 111(1):205–35. [TTR]

Rogers, T. T. & McClelland, J. L. (2004) *Semantic cognition: A parallel distributed processing approach*. MIT Press. [TTR]

Rogers, T. T. & Plaut, D. C. (2002) Connectionist perspectives on category-specific deficits. In: *Category-specificity in brain and mind*, ed. E. Forde & G. W. Humphreys, pp. 251–89. Psychology Press. [aMJ]

Rosch, E. (1978) Principles of categorization. In: *Cognition and categorization*, ed. E. Rosch & B. B. Lloyd, pp. 27–48. Erlbaum. [aMJ]

Rosch, E. & Mervis, C. B. (1975) Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7:573–605. [BKH]

Rosen, R. (1991) *Life itself*. Columbia University Press. [BLA]

Rosenblatt, F. (1962) *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Spartan Books. [aMJ]

Ross, B. H. & Murphy, G. L. (1996) Category-based predictions: Influence of uncertainty and feature associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22:736–53. [BKH]

Ross, L. & Ward, A. (1996) Naive realism in everyday life: Implications for social conflict and misunderstanding. In: *Values and knowledge*, ed. E. S. Reed & E. Turiel, pp. 103–35. Erlbaum. [ELU]

Rougier, N. P., Noelle, D., Braver, T. S., Cohen, J. D. & O'Reilly, R. C. (2005) Prefrontal cortex and the flexibility of cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences USA* 102:7338–43. [aMJ]

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986) Learning representations by back-propagating errors. *Nature* 323:533–36. [aMJ]

Rumelhart, D. E. & McClelland, J. L. (1985) Levels indeed! A response to Broadbent. *Journal of Experimental Psychology: General* 114:193–97. [DKS]

Rumelhart, D. E., McClelland, J. L. & the PDP research group. (1986) *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations*. MIT Press. [aMJ]

Russell, S. & Norvig, P. (2011) *Artificial intelligence: A modern approach*, 3rd edition. Prentice-Hall. [NC]

Sakamoto, Y., Jones, M. & Love, B. C. (2008) Putting the psychology back into psychological models: Mechanistic versus rational approaches. *Memory and Cognition* 36(6):1057–65. [arMJ, DJN, DN]

Salmon, W. C. (1984) *Scientific explanation and the causal structure of the world*. Princeton University Press. [BR]

Samuelson, L. K., Schutte, A. R. & Horst, J. S. (2009) The dynamic nature of knowledge: Insights from a dynamic field model of children's novel noun generalization. *Cognition* 110:322–45. [GWJ]

Sanborn, A. N., Griffiths, T. L. & Navarro, D. J. (2010a) Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review* 117:1144–67. [NC, BKH, arMJ, DJN, BR, DKS]

Sanborn, A. N., Griffiths, T. L. & Shiffrin, R. M. (2010b) Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology* 60:63–106. [aMJ]

Sargent, T. J. (1993) *Bounded rationality in macroeconomics*. Oxford University Press. [aMJ]

Savage, L. J. (1954) *The foundations of statistics*. John Wiley/Dover. [aMJ, MS]

Schall, J. D. (2004) On building a bridge between brain and behavior. *Annual Review of Psychology* 55:23–50. [DKS]

Schervish, M. J. (1995) *Theory of statistics*. Springer Series in Statistics. Springer. [SE]

Schneider, W. & Shiffrin, R. M. (1977) Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review* 84(1):1–66. [ABM]

Schulz, L. E. & Bonawitz, E. B. (2007) Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology* 43:1045–50. [AG]

Schulz, L. E., Bonawitz, E. B. & Griffiths, T. L. (2007a) Can being scared cause tummy aches? Naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology* 43(5):1124–39. [AG]

Schulz, L. E, Gopnik, A. & Glymour, C. (2007b) Preschool children learn about causal structure from conditional interventions. *Developmental Science* 10(3):322–32. [AG]

Schulz, L. E., Goodman, N. D., Tenenbaum, J. B. & Jenkins, A. C. (2008) Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data. *Cognition* 109(2):211–23. [AG]

Schustack, M. W. & Sternberg, R. J. (1981) Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General* 110:101–20. [KJH]

Schwarz, G. E. (1978) Estimating the dimension of a model. *Annals of Statistics* 6(2):461–64. [aMJ]

Seidenberg, M. & Plaut, D. C. (in press) Idealization and its discontents: The legacy of the past tense debate. *Cognitive Science*. [TTR]

Sewell, D. K. & Lewandowsky, S. (2011) Restructuring partitioned knowledge: The role of recoordination in category learning. *Cognitive Psychology* 62:81–122. [DKS]

Shafto, P., Kemp, C., Mansinghka, V. M. & Tenenbaum, J. B. (2011) A probablistic model of cross-categorization. *Cognition*. 120:1–25. [aMJ]

Shanks, D. R. & Dickinson, A. (1987) Associative accounts of causality judgment. In: *The psychology of learning and motivation, vol. 21*, ed. G. H. Bower, pp. 229–61. Academic Press. [KJH]

Shannon, C. E. (1949) *The mathematical theory of communication*. University of Illinois Press. [ABM]

Shepard, R. N. (1987) Toward a universal law of generalization for psychological science. *Science* 237:1317–23. [DJN]

Sherman, D. K. & Cohen, G. L. (2002) Accepting threatening information: Self-affirmation and the reduction of defensive biases. *Current Directions in Psychological Science* 11:119–23.   [ELU]

Sherman, D. K. & Cohen, G. L. (2006) The psychology of self-defense: Self-affirmation theory. In: *Advances in Experimental Social Psychology, vol. 38*, ed. M. P. Zanna, pp. 183–242. Academic Press.   [ELU]

Shi, L., Griffiths, T. L., Feldman, N. H. & Sanborn, A. N. (2010) Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin and Review* 17:443–64.   [NC, EH, DKS]

Shiffrin, R. M. & Steyvers, M. (1997) A model for recognition memory: REM – Retrieving effectively from memory. *Psychonomic Bulletin & Review* 4:145–66.   [EH]

Shiffrin, R. M. & Steyvers, M. (1998) The effectiveness of retrieval from memory. In: *Rational models of cognition*, ed. M. Oaksford & N. Chater, pp. 73–95. Oxford University Press.   [arMJ]

Shiffrin, R. M., Lee, M. D., Kim, W. & Wagenmakers, E. J. (2008) A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science* 32:1248–84.   [DKS]

Simon, H. A. (1957a) A behavioral model of rational choice. In: *Models of man, social and rational: Mathematical essays on rational human behavior in a social setting*, ed. H. A. Simon, pp. 241–60. John Wiley.   [aMJ]

Simon, H. A. (1957b) *Models of man: Social and rational*. John Wiley.   [ABM]

Simoncelli, E. P. & Olshausen, B. A. (2001) Natural image statistics and neural representation. *Annual Review of Neuroscience* 24:1193–216.   [TTR]

Skinner, B. F. (1938) *The behavior of organisms: An experimental analysis*. Appleton-Century.   [arMJ]

Skinner, B. F. (1957) *Verbal behavior*. Appleton-Century-Crofts.   [aMJ]

Skinner, B. F. (1958) Reinforcement today. *American Psychologist* 13:94–99.   [aMJ]

Sloman, S. A. (1993) Feature-based induction. *Cognitive Psychology* 25:231–80.   [BKH]

Sloman, S. A. & Fernbach, P. M. (2008) The value of rational analysis: An assessment of causal reasoning and learning. In: *The probabilistic mind: Prospects for rational models of cognition*, ed. Chater, N. & Oaksford, M, pp. 485–500. Oxford University Press.   [PMF, aMJ]

Smith, D. L. (2007) Beyond Westemarck: Can shared mothering or maternal phenotype matching account for incest avoidance? *Evolutionary Psychology* 5:202–22.   [aMJ]

Smith, L. B. (2000) Learning how to learn words: An associative crane. In: *Becoming a word learner: A debate on lexical acquisition*, ed. R. M. Golinkoff, K. Hirsh-Pasek, K., L. Bloom, L.B. Smith, A. L. Woodward, N. Akhtar, M. Tomasello & G. Hollich, pp. 51–80. Oxford University Press.   [GWJ]

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L. & Samuelson, L. (2002) Object name learning provides on-the-job training for attention. *Psychological Science* 13:13–19.   [GWJ, arMJ]

Smith, L. B. & Thelen, E. (2003) Development as a dynamic system. *Trends in Cognitive Science* 7(8):343–48.   [GWJ]

Smith, P. K. (1982) Does play matter? Functional and evolutionary aspects of animal and human play. *Behavioral and Brain Sciences* 5:139–84.   [aMJ]

Smolensky, P. (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11:1–23.   [aMJ]

Smolin, L. (2006) *The trouble with physics: The rise of string theory, the fall of a science, and what comes next*. Houghton Mifflin Harcourt.   [aMJ]

Sobel, D. M. & Kirkham, N. Z. (2006) Blickets and babies: The development of causal reasoning in toddlers and infants. *Developmental Psychology* 42(6):1103–15.   [AG]

Sobel, D. M., Tenenbaum, J. B. & Gopnik, A. (2004) Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in pre-schoolers. *Cognitive Science* 28(3):303–33.   [AG, aMJ]

Soltani, A. & Wang, X.-J. (2010) Synaptic computation underlying probabilistic inference. *Nature Neuroscience* 13(1):112–19.   [aMJ]

Spencer, J. P. & Perone, S. (2008) Defending qualitative change: The view from dynamical systems theory. *Child Development* 79:1639–47.   [GWJ]

Spencer, J. P., Perone, S. & Johnson, J. S. (2009) The dynamic field theory and embodied cognitive dynamics. In: *Toward a unified theory of development: Connectionism and dynamic systems theory reconsidered*, ed. J. P. Spencer, M. S. Thomas & J. L. McClelland, pp. 86–118. Oxford University Press.   [aMJ]

Spencer, J. P., Perone, S., Smith, L. B. & Samuelson, L. K. (2011) The process behind the "suspicious coincidence": Using space and time to learn words. *Psychological Science*. doi: 10.1177/0956797611413934   [GWJ]

Spencer, J. P., Thomas, M. S. & McClelland, J. L. (2009) *Toward a unified theory of development: Connectionism and dynamic systems theory re-considered*. Oxford University Press.   [GWJ]

Sperber, D. & Hirschfeld, L. A. (2003) The cognitive foundations of cultural stability and diversity. *Trends in Cognitive Sciences* 8:40–46.   [aMJ]

Spirtes, P., Glymour, C. & Scheines, R. (2000) *Causation, prediction, and search*, 2nd edition. (original edition published in 1993) MIT Press.   [AG, rMJ]

Stankiewicz, B. J., Legge, G. E., Mansfield, J. S. & Schlicht, E. J. (2006) Lost in virtual space: Studies in human and ideal spatial navigation. *Journal of Experimental Psychology: Human Perception and Performance* 32:688–704.   [aMJ]

Stanovich, K. E. & West, R. F. (1998) Individual differences in rational thought. /precode> *Journal of Experimental Psychology: General* 127(2):161–88.   [ABM]

Steele, C. M. (1988) The psychology of self-affirmation: Sustaining the integrity of the self. In: *Advances in Experimental Social Psychology*, vol. 21, ed. L. Berkowitz, pp. 261–302. Academic Press.   [ELU]

Sternberg, S. (1966) High-speed scanning in human memory. *Science* 153:652–54.   [rMJ]

Steyvers, M., Lee, M. D. & Wagenmakers, E.-J. (2009) A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology* 53:168–79.   [aMJ]

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J. & Blum, B. (2003) Inferring causal networks from observations and interventions. *Cognitive Science* 27:453–89.   [IB, PMF, aMJ]

Stigler, S. M. (1961) The economics of information. *Journal of Political Economy* 69:213–25.   [aMJ]

Stocker, A. A. & Simoncelli, E. P. (2006) Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience* 9(4):578–85.   [JSB]

Teller, D. Y. (1984) Linking propositions. *Vision Research* 10:1233–46.   [DKS]

Tenenbaum, J. B. & Griffiths, T. L. (2001) Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences* 24(4):629–40.   [PMF, aMJ, ABM, DJN]

Tenenbaum, J. B. & Griffiths, T. L. (2001) Structure learning in human causal induction. In: *Advances in neural information processing systems, vol. 13*, ed. T. K. Leen, T. G. Dietterich & V. Tresp, pp. 59–65. MIT Press.   [KJH]

Tenenbaum, J. B., Griffiths, T. L. & Kemp, C. (2006) Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences* 10:309–18.   [aMJ]

Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. (2011) How to grow a mind: Statistics, structure, and abstraction. *Science* 331(6022):1279–85.   [BLA, NC, GWJ]

Tetlock, P. E. (2002) Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. *Psychological Review* 109:451–71.   [ELU]

Tetlock, P. E., Kristel, O., Elson, B., Green, M. & Lerner, J (2000) The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology* 78:853–70.   [ELU]

Tetlock, P. E., Visser, P., Singh, R., Polifroni, M., Elson, B., Mazzocco, P. & Rescober, P. (2007) People as intuitive prosecutors: The impact of social control motives on attributions of responsibility. *Journal of Experimental Social Psychology* 43:195–209.   [ELU]

Thagard, P. (1989) Explanatory coherence. *Behavioral and Brain Sciences* 12:435–502.   [aMJ]

Thaler, R. H. & Sunstein, C. R. (2008) *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.   [aMJ]

Thibaux, R. & Jordan, M. I. (2007) Hierarchical beta processes and the Indian buffet process. In: *Proceedings of the Tenth Conference on Artificial Intelligence and Statistics (AISTATS)*, ed. M. Meila & X. Shen. Society for Artificial Intelligence and Statistics. (Online Publication). Available at: http://www.stat.umn.edu/%7Eaistat/proceedings/start.htm   [aMJ]

Thomas, M. S. C. & McClelland, J. L. (2008) Connectionist models of cognition. In: *The Cambridge handbook of computational psychology*, ed. R. Sun, pp. 23–58. Cambridge University Press.   [DKS]

Thompson, P., Brooks, K. & Hammett, S. T. (2006) Speed can go up as well as down at low contrast: Implications for models of motion perception. *Vision Research* 46(6–7):782–86.   [JSB]

Thompson-Schill, S., Ramscar, M. & Chrysikou, M. (2009) Cognition without control: When a little frontal lobe goes a long way. *Current Directions in Psychological Science* 8:259–63.   [aMJ]

Tikhonov, A. N. & Arsenin, V. Y. (1977) *Solutions of ill-posed problems*. W. H. Winston.   [SE]

Tinbergen, N. (1963) On aims and methods of ethology. *Zeitschrift für Tierpsychologie* 20:410–33.   [LA-S]

Tooby, J. & Cosmides, L. (2005) Conceptual foundations of evolutionary psychology. In: *The handbook of evolutionary psychology*, ed. D. M. Buss, pp. 5–67. Wiley.   [aMJ]

Tooby, J., & Cosmides, L. (1992) The psychological foundations of culture. In: *The adapted mind: Evolutionary psychology and the generation of culture,* J. H. Barkow, L. Cosmides, & J. Tooby, pp. 19–136. Oxford University Press.   [LA-S]

Tooby, J., Cosmides, L., Sell, A., Lieberman, D. & Sznycer, D. (2008) Internal regulatory variables and the design of human motivation: A computational and

evolutionary approach. In: *Handbook of approach and avoidance motivation*, ed. A. Elliot, pp. 251–71. Erlbaum. [DP]

Turing, A. (1950) Computing, machinery and intelligence. *Mind* 49:433–60. [ABM]

Tversky, A. & Kahneman, D. (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185:1124–31. [aMJ]

Uhlmann, E. L. & Cohen, G. L. (2005) Constructed criteria: Redefining merit to justify discrimination. *Psychological Science* 16:474–80. [ELU]

Vul, E., Frank, M. C., Alvarez, G. A. & Tenenbaum, J. B. (2009) Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. *Advances in Neural Information Processing Systems* 22:1955–63. [aMJ]

Vul, E., Goodman, N. D., Griffiths, T. L. & Tenenbaum, J. B. (2009a) One and done: Optimal decisions from very few samples. In: *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, ed. N. Taatgen & H. van Rijn, pp. 148–53. Erlbaum. [NC]

Vul, E., Hanus, D. & Kanwisher, N. (2009b) Attention as inference: Selection is probabilistic, responses are all-or-none samples. *Journal of Experimental Psychology: General* 138:546–60. [NC]

Waelti, P., Dickinson, A. & Schultz, W. (2001) Dopamine responses comply with basic assumptions. *Nature* 412:43–48. [IB]

Waldmann, M. R. & Holyoak, K. J. (1992) Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General* 121:222–36. [KJH]

Waldmann, M. R. & Martignon, L. (1998) A Bayesian network model of causal learning. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, ed. M. A. Gernsbacher & S. J. Derry, pp. 1102–107. Erlbaum. [KJH]

Wallsten, T. S. (1971) Subjectively expected utility theory and subjects' probability estimates: Use of measurement-free techniques. *Journal of Experimental Psychology* 88:31–40. [MS]

Walsh, C. R. & Sloman, S. A. (2008) Updating beliefs with causal models: Violations of screening off. In: *Memory and Mind: A festschrift for Gordon H. Bower*, ed. M. A. Gluck, J. R. Anderson & S. M. Kosslyn, pp. 345–57. Erlbaum. [NC]

Wason, P. (1960) On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology* 12:129–40. [DJN]

Wasserman, L. (2003) *All of statistics: A concise course in statistical inference*. Springer Texts in Statistics. Springer. [SE]

Watson, J. B. (1913) Psychology as the behaviorist views it. *Psychological Review* 20:158–77. [aMJ]

Weiss, Y., Simoncelli, E. P. & Adelson, E. H. (2002) Motion illusions as optimal percepts. *Nature Neuroscience* 5(6):598–604. [JSB]

Wertheimer, M. (1923/1938) Laws of organization in perceptual forms. In: *A source book of Gestalt psychology*, ed. & trans. W. Ellis, pp. 71–88. Routledge & Kegan Paul. (Original work published in 1923). [aMJ]

Wilder, M. H., Jones, M. & Mozer, M. C. (2009) Sequential effects reflect parallel learning of multiple environmental regularities. *Advances in Neural Information Processing Systems* 22:2053–61. [MH, arMJ]

Wilson-Mendenhall, C. D., Barrett, L. F., Simmons, W. K. & Barsalou, L. W. (2011) Grounding emotion in situated conceptualization. *Neuropsychologia* 49:1105–27. [LWB]

Woit, P. (2006) *Not even wrong: The failure of string theory and the search for unity in physical law*. Basic Books. [aMJ]

Wolpert, D. (1996) The lack of a priori distinctions between learning algorithms. *Neural Computation* 8:1341–90. [aMJ]

Wood, J. N. & Grafman, J. (2003) Human prefrontal cortex: Processing and representational perspectives. *Nature Reviews: Neuroscience* 4:129–47. [aMJ]

Xu, F. & Tenenbaum, J. B. (2007a) Sensitivity to sampling in Bayesian word learning. *Developmental Science* 10(3):288–97. [GWJ, DJN]

Xu, F. & Tenenbaum, J. B. (2007b) Word learning as Bayesian inference. *Psychological Review* 114(2):245–72. [GWJ, aMJ, TTR]

Yamauchi, T. & Markman, A. B. (1998) Category learning by inference and classification. *Journal of Memory and Language* 39:124–48. [aMJ]

Yang, L.-X. & Lewandowsky, S. (2003) Context-gated knowledge partitioning in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29:663–79. [DKS]

Yang, L.-X. & Lewandowsky, S. (2004) Knowledge partitioning in categorization: Constraints on exemplar models. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30:1045–64. [DKS]

Yeh, W. & Barsalou, L. W. (2006) The situated nature of concepts. *American Journal of Psychology* 119:349–84. [LWB]

Yu, A. & Cohen, J. (2008) Sequential effects: Superstition or rational behavior? *Advances in Neural Information Processing Systems* 21:1873–80. [aMJ]

Yuille, A. & Kersten, D. (2006) Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences* 10:301–308. [NC]

Zambrano, E. (2005) Testable implications of subjective expected utility theory. *Games and Economic Behavior* 53:262–68. [MS]