To appear in *Behavioral and Brain Sciences*

Pinning Down the Theoretical Commitments of Bayesian Cognitive Models

Matt Jones
University of Colorado
mcj@colorado.edu

Bradley C. Love
University of Texas
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 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 only be evaluated 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.

Pinning Down the Theoretical Commitments of Bayesian Cognitive Models

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 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.

Theoretical status of Bayesian models

A primary confusion surrounding Bayesian cognitive models is whether they are intended as purely computation-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 they 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 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 that 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. 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, etc. 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 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 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 have highlighted (**Glymour**, **Rehder**, **Rogers & Seidenberg**), 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 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 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 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.

Bayesian models as computation-level theories

Setting aside how Bayesian models have been intended—which we argue above is often unclear—we now evaluate their potential theoretical contribution under a purely computation-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 above. 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 computation-level Bayesian model. However, the situation is more severe than this, because on a true computation-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 that "Behaviorists believe 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 nonexistant) 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 computation-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, 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, that 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. Thus, 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 correctly predict new data. **Jenkins et al.**'s observations 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." Thus, 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 computation-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 & Eberhart** 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, section 4.3). 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 they 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.

Bayesian models as mechanistic theories

The alternative to a purely computation-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 (section 2), 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 el 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 articles that utilize 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. 2010) fits squarely into one of the approaches that we advocated as Bayesian Enlightenment in the target article (section 5.1).

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. 1993) 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 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 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 thus 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 noted above, 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 the advantages of clarifying what the model does and does not explain, identifying important questions remaining to be answered, and facilitating comparison to other models cast in different frameworks. Regarding the first two questions, the model demonstrates that second-order generalization emerges from Bayesian inference together with the expectation of

dimensional regularity, but many other questions remain, such as: How does the learner know to expect this particular regularity in the environment? How does he or she verify the pattern is present in the input data? (Like most Bayesian models, the model takes p[data|hypothesis] as a starting point, without reference to how this conditional probability is evaluated.) How does the learner produce new responses consistent with what he or she has inferred? These are all natural questions from a mechanistic perspective, and the model would be much stronger if it included answers to them.

As Jenkins et al. explain, the structure discovered by Bayesian models of development does not truly develop or emerge. It is built in a priori. All a Bayesian model does is determine which of the patterns or classes of patterns it was endowed with is most consistent with the data it is given. Thus there is no explanation of where those patterns (i.e., hypotheses) come from. Once one realizes that the structure built into the (over-) hypothesis space is at the core of the model's explanation, it is natural to compare those assumptions with the knowledge assumed within other theoretical frameworks (the third advantage listed above). 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 (1991) 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 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.

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, several challenges were raised in the commentaries, which we address here.

Regarding the general proposal of incorporating rational or computational principles into mechanistic modeling, **Anderson** argues that computation-level modeling is incoherent because the computational level does not exist, because 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 computation-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 computation-level modeling are often identified, **Fernbach & Sloman** point out 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, section 4.3).

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 above for treating Bayesian hypotheses as assumptions about the learner's knowledge, rather than products of rational analysis. It also fits with our proposal in the target article (section 5.2) for bringing rational analysis inside mechanistic models, 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. 2010). 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, section 4.4). 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, to explain why one algorithm or implementation is more appropriate than another for a given task. Nevertheless, there remain questions 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, sections 1.2 & 5.2). Thus incorporating rational principles is potentially fruitful within essentially 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 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 both the computational goals of the learning system (i.e., the statistics to be estimated) and 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 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 computation-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 the purely rational model considered by Sakamoto et al. is misspecified 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 J.R. 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 it is 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, **Sewall 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.

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, **Sewall 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

Anderson, J. R. (1990) *The adaptive character of thought*. Erlbaum.

Anderson, J. R. (1991) The adaptive nature of human categorization. *Psychological Review* 98:409–29.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

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 B* 362:933-942.

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

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

Elliott, S. W., & Anderson, J. R. (1995). Effect of memory decay on predictions from changing categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 21:815-836.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 10:234-257.

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

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

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

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

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

Love, B. C. (2005). Environment and goals jointly direct category acquisition. *Current Directions in Psychological Science*, 14:195-199.

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. 205–225. Cambridge Press.

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

Pearl, J. (2000). *Causality*. Cambridge University Press.

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 Research and Theory*, Ed. A.H. Black & W.F. Prokasy, pp. 64–99. Appleton-Century-Crofts.

Sakamoto, Y., Jones, M., & Love, B. C. (2008). Putting the psychology back into psychological models: Mechanistic vs. rational approaches. *Memory & Cognition, 36*, 1057-1065.

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.

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.

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

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.

Spirtes, P., Glymour, C., & Scheines, R. (2000) *Causation, prediction, and search*. MIT Press.

Sternberg, S. (1966). High-speed scanning in human memory. *Science* 153:652-654.

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.