Cognitive Science 32 (2008) 155–161
Copyright © 2008 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1080/03640210701802113

Is a Single-Bladed Knife Enough to Dissect Human Cognition? Commentary on Griffiths et al.

Wai-Tat Fu

Human Factors Division and Beckman Institute, University of Illinois at Urbana-Champaign

Abstract

Griffiths, Christian, and Kalish (this issue) present an iterative-learning paradigm applying a Bayesian model to understand inductive biases in categorization. The authors argue that the paradigm is useful as an exploratory tool to understand inductive biases in situations where little is known about the task. It is argued that a theory developed *only* at the computational level is much like a single-bladed knife that is only useful in highly idealized situations. To be useful as a general tool that cuts through the complex fabric of cognition, we need at least two-bladed scissors that combine both computational and psychological constraints to characterize human behavior. To temper its sometimes expansive claims, it is time to show what a Bayesian model *cannot* explain. Insight as to how human reality may differ from the Bayesian predictions may shed more light on human cognition than the simpler focus on what the Bayesian approach *can* explain. There remains much to be done in terms of integrating Bayesian approaches and other approaches in modeling human cognition.

Keywords: Bounded rationality; Cognitive modeling; Bayesian modeling; Psychological representations

1. Introduction

Griffiths, Christian, and Kalish (this issue) present results from applying an iterated learning paradigm to study human inductive biases. Human responses were analyzed based on a Bayesian model, which assumed that people optimize their performance by conforming to Bayes's rule that specifies how posterior conditional probabilities (of a hypothesis being true, given an example data pattern) are computed from the product of the prior probability of the hypothesis and the likelihood that the example data pattern comes from the given hypothesis. The common procedure for many Bayesian approaches is to assume human responses correspond directly with the posterior probabilities and design the priors and likelihoods to fit observed posteriors. If there is a good fit, the Bayesian model is said to provide a computational theory of the corresponding cognitive function.

Correspondence should be addressed to Wai-Tat Fu, Human Factors Division and Beckman Institute, University of Illinois at Urbana-Champaign, 405 North Mathews Ave., Urbana, IL 61801. E-mail: wfu@uiuc.edu

Griffiths et al. (this issue) argue that combining the Bayesian model with the iterative learning paradigm provides an exploratory tool to study human inductive biases, especially in contexts where "defining a computation model is difficult" (p. 29). The authors consider their approach as "nonparametric," as "it will yield samples from the prior regardless of the form of that prior, and those samples can be used to construct estimates of the corresponding distribution that make only weak assumptions about the nature of human inductive biases" (p. 30). Indeed, part of the appeal of the Bayesian approach is that it provides a principled way of computing both the statistical properties and structures that describe how knowledge can be modified by new evidence without specific assumptions about psychological representations or processes. In Marr's (1982) term, the model aims at prescribing a computational theory of cognition that can be implemented by different representations and processes in different situations. In other words, the computational theory helps to frame the information-processing problem imposed by the task environment, and different process models are possible solutions to the information-processing problem.

I argue that developing theory *only* at the computational level is not enough to understand cognition. Borrowing Simon's (1956) term, computational theory together with psychological representations and processes form two blades of the scissors that must be used to dissect cognition.[1] I start with commenting on the list of general assumptions in the paradigm by Griffiths et al. (this issue), and argue that the explanatory power from a pure computational theory is quite limited. In addition, I argue that in many situations, the information-processing problem is difficult to define without assumptions on the underlying psychological representations and processes. Finally, I discuss in general how Bayesian approaches may benefit from integrating with theories at the representational and implementational levels.

2. General assumptions in a Bayesian paradigm

As in most Bayesian approaches, Griffiths et al. (this issue) matched human responses to the posterior distributions in their model and constructed priors and likelihoods to find the best-fit parameters (maximum likelihood estimators) that produce the posteriors. The combination of priors, likelihoods, and Bayes's rule is said to reflect the nature of the computations performed by the respondents in the task. Predictions of the model therefore depend critically on the assumptions about the priors, likelihoods, and how well the Bayes's rule can be used to characterize learning. In this section, I focus on the assumptions of the priors and likelihood. In the remaining sections, I discuss the general Bayesian assumptions.

2.1. The hypothesis space

One major assumption of the Bayesian approach is a hypothesis space that defines some form of abstract knowledge of the particular domain of interest. In the model by Griffith et al. (this issue), the hypothesis space consisted of all possible categories in their task. The learner was assumed to be biased to choose some of these categories more often than others, and these biases were represented by a set of prior probabilities. Although a natural and more challenging question for most Bayesian models is where these priors come from, Griffiths et al. capitalized

on previous research on this task and assumed that they could be classified into six unique types and represented by six prior probabilities. Their model was then developed at an abstract level such that the behavior of the model is independent of the specific stimuli used in their task. In other words, at this level of abstraction the model predictions would be exactly the same whether the dimensions were distinguished by colors, shape, brightness, or loudness. Like most Bayesian models, the model by Griffiths et al. was developed at the computational level, and thus was concerned with predicting the functional characteristics of the system rather than predicting how those stimuli were processed and responses generated. In fact, as stated in Griffiths et al. the lack of assumptions of representations and processes is considered a valuable property, as the model, combined with the iterated learning paradigm, can potentially be used as an exploratory tool to reveal the underlying computational characteristics of inductive reasoning.

How general is Griffiths et al.'s (this issue) assumption of the hypothesis space in inductive reasoning, so that the same model can predict human responses in other inductive tasks? First of all, the assumption that all participants started to learn from the pre-specified set of candidate hypotheses may be questionable. Is it possible that people (and the model) learn to develop these categories through experience? Can people always distinguish the various stimuli along different dimensions in ways that are consistent with the categories constructed by the modeler? The assumption of a constant hypothesis space regardless of what stimuli were presented also seems inconsistent with the idea that the set of hypotheses may change during learning. Although it is possible to partially overcome this apparent difficulty by defining the hypothesis space as the set of all possible knowledge states (instead of the ones adopted by the learner at a particular time) and define higher levels of abstract knowledge to constrain the prior probabilities for the hypotheses (Kemp, Perfors, & Tenenbaum, 2004), at some level these "super priors" have to be grounded on empirically testable assumptions at the psychological or implementational level to avoid an infinite regress. Another problem with this argument is that it makes the theory difficult to falsify, as different priors and structures can be created to explain the posteriors. One obvious way to provide a stronger ground for the Bayesian approach is to put more effort into integrating multiple modeling approaches and show, for example, how Bayesian inference and the different assumed structures can be related to other process models and/or how they could be implemented in the brain.

2.2. The likelihood function

Given a set of priors, Bayesian models require a likelihood function that estimates the probability that an observation comes from one of the hypotheses in the hypothesis space. In Griffiths et al. (this issue), the likelihood function was constructed based on the assumption that the learner sampled an example from the set of possible members in each of the six types of categories. In Griffiths et al.'s model, the likelihood function was developed by analogizing the inference process to the idealized situation of a person drawing marbles (observation of an amoeba) from six different bags (the 6 types of categories) without replacement. As discussed earlier, one advantage of developing models at this high level of abstraction is that it allows the model to be applicable across a wide range of domain such as word-learning, inductive inference, or concept formation. However, an obvious disadvantage is that it is not clear that the model or the experiments are addressing issues and tasks relevant to inductive inferences per

se, including learning difficulties related to different stimuli and structures, possibly mediated by the inherent logical complexities (Feldman, 2000) or other cognitive demands such as memory limits. These assumptions at the process level not only provide useful information about human inductive inferences, but they are also empirically testable in ways that would directly advance our knowledge about the human inductive process. Besides, it is rarely the case that cognitive scientists will be just interested in *what* categories people can learn or not *a priori*. The more interesting questions seem to be how and when people will generalize from one situation to another based on the judgment that both situations are likely to belong to a category of situations having the same consequences, *why* and *how* people have trouble learning some categories, and whether we can overcome these difficulties. The answers to these questions demand our understanding of how the uncertainty about the distribution of consequential stimuli is represented and processed in the psychological space (Shepard, 1987). With a lack of directly testable assumptions at the level of psychological representation and process, what a computational theory *cannot* predict is why different learning difficulties arise for the various types of categories and what cognitive limits prevent people from attaining optimality, and what can be done to overcome these difficulties.

It seems that the likelihood function provides a perfect setting to combine the strength of a computational theory and the explanatory power of psychological representations and processes. For example, Shepard (1987) showed that the mapping of stimuli in different dimensions (such as brightness, loudness) into the psychological space may provide strong hints on people's judgments on whether an observation is consistent with a hypothesis, or how other cognitive limits may play a role as participants perform the mental operations required to generate a response. This mapping from external stimuli to psychological representations is precisely what the likelihood function is supposed to capture, and many behavioral differences could be explained by the different mappings induced by the interaction between external stimuli and internal representations. Indeed, different representations and processes could elicit very different responses, and may impose vastly different learning difficulties that cannot be easily captured by an abstract model, as evidenced by the different learning difficulties in isomorphic tasks (Kotovsky, Hayes, & Simon, 1985; Kotovsky & Simon, 1990).

3. What can we learn from a computational theory?

A common reservation to the Bayesian approach is the assumption that people are rational agents trying to optimize their performance. This assumption is perhaps reasonable in many cases, but in other cases it is clear that behavior does sometimes deviate from the Bayesian norms (e.g., Gigerenzer & Todd, 1999). A common response to this reservation is that Bayesian models aim at prescribing the computational properties of cognition, but not describing the mechanisms. In other words, Bayesian approaches never assume that people actually carry out those complex Bayesian computations in their head, but that the priors and likelihoods indicate the *nature* of the computations that can be implemented in various process models. Although there is little doubt that Bayesian approaches shed light on the computational properties of many aspects of cognition, it is sometimes difficult to know exactly what the model reveals.

I focus on two potential problems. One is that the computational theory relies much on the assumption of Bayesian optimization and the careful construction of the priors and likelihoods by the skilled modeler. The interdependency between the optimization assumption and the specific properties of the chosen priors and likelihoods make it hard to tease them apart and thus cannot be easily tested empirically. In fact, the conclusions that Bayesian modelers draw from their optimization models seldom depend solely on the optimizing assumptions, but they do depend critically on the assumptions in the priors and likelihoods. In many cases, changes in these assumptions of priors and likelihoods (while retaining the Bayesian optimization assumption) may lead to different conclusions.

One possible solution to this first problem is to construct the model based on a consistent set of general assumptions. In other words, instead of constructing priors and likelihoods specific for a given task, the same set of assumptions can be applied *across* different tasks and domains, so that the nature of the computation can be directly explained with respect to the set of general assumptions, and are therefore directly testable by comparing results across tasks and domains. One example is the rational analysis by Anderson (1990), who developed a number of Bayesian models to explain a range of general phenomena of cognition based on the core assumption that the priors and likelihoods conform to the statistical structures of the task environment. The nature of the computations prescribed by the Bayesian models can therefore be explained by the adaptive nature of behavior to the environment. In situations when the data deviated from the predictions, the deviations can be traced back to the assumptions of the environment to help determine their cause.

The second problem is that psychological representations and processes may directly change the computational problem faced by the learner. This notion can be traced back to Simon's (1956) notion of bounded rationality where he argued that the "bounds" of cognition need to be included as parts of the information-processing problem imposed by a task environment to a person. So while the conventional wisdom in cognitive science is that people are not rational, when these bounds are taken into consideration, behavior can still be shown to be *satisfactorily* rational. The point is that we need to make additional assumptions about these bounds in order to define the computational problem, and these bounds often depend on the specific psychological representations and processes used by the person. This does not imply that the pure computational aspect of cognition is not important, but that only in the simplest cases will the system be predictable by a pure computational theory. In most other cases, effects of specific representations and processes will "show through" in the data, as in the higher than expected Type VI response proportions in Griffiths et al.'s (this issue) study (See also Fu & Gray, 2006; Fu & Pirolli, 2007). Indeed, Griffiths et al. explained the discrepancies between their data and previous findings by speculating that different memory demands may influence how easily the task can be learned. If a researcher uses their Bayesian model to analyze an unfamiliar task as suggested by Griffiths et al. and found that people have difficulties learning a particular category structure, can the researcher easily tease apart whether the difficulty can be attributed to different priors (inductive biases), violation of the Bayesian assumption, or psychological limits and constraints? It seems that more assumptions at the representational or process levels would be needed to generalize their findings to other tasks, so as to make their model useful as an exploratory tool to explain human inductive reasoning.

4. Validating a computational theory

Similar to most modeling approaches, Griffiths et al. (this issue) validate the predictions of their model by showing their fit to human data. The validation of the model by Griffiths et al. illustrates two important issues related to testing models developed at the computational level. The first issue concerns the kind of mutual constraints provided by the model and the data that are needed to validate a computational theory. The second issue concerns the importance of identifying where and when the model does not fit human data.

Because models developed at the computational level do not make any specific claims about the representations and processes that generate the responses, modelers are in general satisfied with the observation that the model predictions match the general trends as shown in the human data. For example, Griffiths et al. (this issue) conclude that their model matches the human data well based on the observation that the model predicts most of the up and down trends across trials as their respondents learned to do the task, as confirmed by the high linear correlation between model predictions and human data. Because it is obvious that some trends are easier to generate than others, an important consideration is the a priori likelihood that one is able to predict those general trends. The moral is that a model is usually more convincing and useful if it can predict trends that have low a priori likelihood to appear in other data sets, and if the trends cannot be easily accounted for by other computational theories. I will discuss how the data from Griffiths et al.'s studies score on this measure.

In the studies by Griffiths et al. (this issue), almost one half of the responses in all category types were close to zero (see Figs. 5, 6, 7, & 8 in that article). Even for those that were non-zero, learning seems to follow simple up and down trends (and because they are choice proportions they are inherently tied to each other as one goes up the other goes down). The simplicity in the patterns of data provides few constraints on possible theories for these learning curves or, in other words, it is not clear what theories these data rule out. In fact, one may easily show that a set of simple linear equations can predict the same trends. For the same reason, the model does not provide enough constraints on the possible outcomes, as most of the predictions by Griffiths et al.'s model were simple monotonic functions that stay flat after the first few blocks (e.g., the model predictions shown in their Fig. 8), it is therefore hard to know what outcomes their model truly predicts and what outcomes it can rule out.

The second issue concerns the situation when model predictions deviate from human data. For example, in all three experiments in Griffiths et al. (this issue), there was an uptrend in the responses initiated by the Type IV chain, but the model consistently predicted a downtrend across trials. Was the discrepancy due to the wrong likelihood function or that learning did not conform to the Bayesian assumption? What can we learn from the discrepancy? Another example was the finding that the Type VI responses were found to be higher than found in previous studies. Indeed, in many situations one can learn more when a model deviates from human behavior than when it fits the data. Drawing an example from economics, the assumption that laborers are all rational agents will predict that there should be no unemployment. The apparent discrepancy with real-world observations led to the Keynesian explanation that labor fails to distinguish between increases in real wages and money wages. This failure is a violation of the assumption that people are rational agents, but it is precisely this irrationality that explains the real-world phenomenon: unemployment (Simon, 1996). Similarly, Griffiths

et al. speculate that the higher proportion of the Type VI responses found in their studies compared to previous studies was due to differences in memory demands. However, without further testing of their models, it is hard to know whether it was truly due to cognitive limits, violation of the Bayesian assumption, or differences in priors or likelihoods. The example does seem to suggest that the integration of Bayesian approaches with representational and process accounts of cognition will be an important contribution.

5. Conclusion

Bayesian modeling has a long history in cognitive science and has led to significant insights into various aspects of cognition. However, although powerful and important, the Bayes's rule can only take us so far. To add on to its already impressive contributions, integration with other approaches at the levels of representation, process, and implementation will be an important next step. To temper its sometimes expansive claims, it is also time to show what it *cannot* explain. Insight as to how human reality may differ from the Bayesian predictions may shed more light on human cognition than the simpler focus on what the Bayesian approach *can* explain.

Notes

1. In Simon's (1956) original analogy, the two blades refer to the structures of the task environment and the psychological bounds of the person. Although the Bayesian computational theory is not necessarily based on structures of the task environment, the implicit assumption about the underlying principle of rationality appears to be consistent.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature, 407,* 630–633.

Fu, W.-T., & Gray, W. D. (2006). Suboptimal tradeoffs in information-seeking. *Cognitive Psychology, 52,* 195–242.

Fu, W.-T., & Pirolli, P. (2007). SNIF-ACT: A cognitive model of user navigation on the World Wide Web. *Human–Computer Interaction, 22,* 355–412.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Kemp, C., Perfors, A., & Tenenbaum, J.B. (2004). Learning domain structures. In Forbus, Kenneth and Gentner, Dedre (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 672–678). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Kotovsky, K., Hayes, J. R., & Simon, H. A. (1985). Why are some problems hard? *Cognitive Psychology, 17,* 248–292.

Kotovsky, & Simon, H. A. (1990). What makes some problems really hard: Explorations in the problem space of difficulty. *Cognitive Psychology, 22*, 143–183.

Marr, D. (1982). *Vision*. San Fransciso: Freeman.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science, 237*, 1317–1323.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*, 129–138.

Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). Cambridge, MA: MIT Press.