Philosophy Compass 7/2 (2012): 127–138, 10.1111/j.1747-9991.2011.00467.x

Bayesian Models of Cognition: What's Built in After All?

Amy Perfors*
University of Adelaide

Abstract

This article explores some of the philosophical implications of the Bayesian modeling paradigm. In particular, it focuses on the ramifications of the fact that Bayesian models pre-specify an inbuilt hypothesis space. To what extent does this pre-specification correspond to simply "building the solution in"? I argue that *any* learner (whether computer or human) must have a built-in hypothesis space in precisely the same sense that Bayesian models have one. This has implications for the nature of learning, Fodor's puzzle of concept acquisition, and the role of modeling in cognitive science.

1. Introduction

In recent years, Bayesian modeling has become an increasingly popular tool in cognitive science. It has been used to explore questions in topics as wide-ranging as causal learning and reasoning (e.g., Pearl 2000; Steyvers *et al.* 2003; Griffiths and Tenenbaum 2009), decision making (e.g., Lee 2006), concept learning and representation (e.g., Anderson 1991; Griffiths *et al.* 2007; Kemp *et al.* 2007; Sanborn *et al.* 2010), reasoning about agents (e.g., Baker *et al.* 2007; Feldman and Tremoulet 2008) theory learning and representation (e.g., Goodman *et al.* 2007; Griffiths *et al.* 2010), and language (e.g., Griffiths and Kalish 2007; Xu and Tenenbaum 2007; Feldman and Griffiths 2009; Frank *et al.* 2009; Goldwater *et al.* 2009; Perfors *et al.* 2011b). The popularity of the Bayesian approach has led many questions about its utility and relevance.[1] What are its strengths and weaknesses as a framework for understanding the pressing questions in cognitive science? As with any modeling approach, it makes certain assumptions about – and has certain implications for – cognition and cognitive development. One central issue in cognitive science is the question of innateness: to what extent, and in what way, is human knowledge and behavior dependent on inborn or unlearned principles or constraints? The goal of this paper is to explore how Bayesian modeling touches on these fundamental issues.

Bayesian models are neither strongly nativist nor strongly empiricist, primarily because they are flexible enough to capture either type of theory. As a result, the relevance of the approach to questions of innateness does not derive from it taking a strong theoretical stance one way or another. Rather, because Bayesian models require the explicit specification of details that may be important but easy to overlook when a theory is verbally specified, they can clarify which factors are necessary to specify (in *any* model) as well as what the effects of such specification might be. Moreover, because the models can simulate behavior that is far more complex than we can predict through simple introspection, they are a useful tool for understanding what kind of learning is actually possible given different assumptions about the learner and the data available.

In this paper I focus on one primary area in which the Bayesian framework may especially illuminating: the implications of the fact that in Bayesian models, hypothesis spaces

must be explicitly specified. Because of this, Bayesian models might appear to be strongly nativist, but I argue instead that *all* models and learners must (implicitly or explicitly) build in hypothesis spaces in just this way. One implication is that any conception of learning which relies on not having a built-in hypothesis space is incoherent: the interesting question for cognitive scientists is the nature and extent of this built-in knowledge. I devote much of the paper to exploring this idea and its ramifications. First, though, I briefly review the basics of Bayesian modeling.[2]

2. The Bayesian Approach: A Brief Introduction

One of the primary aims of Bayesian modeling is to explore questions on Marr's computational level (1982). At this level the focus is on understanding how different characteristics of the learner, combined with differences in the type and nature of the data available, affect the nature of learning. What are the abstract goals of the cognitive system? What problem does it solve? How do the constraints under which it solves that problem affect what is learned? Bayesian modeling is also focused on the question of *why*: Why does the cognitive system have these goals? What would a good solution look like, and why would it be good?

The main idea of the Bayesian approach is to use the mathematics of probability theory to yield normative answers to these questions, and to use those answers as a standard against which human performance can be compared. Within probability theory, degrees of belief in some hypothesis or theory h are represented using real numbers ranging from 0 to 1, and Bayes' Rule (Equation 1 below) describes how to update beliefs in response to new data. According to Bayes' Rule, the posterior probability of a hypothesis h given some data d, denoted $P(h \mid d)$, is proportional to the prior probability of that hypothesis $P(h)$ and the likelihood of observing that data if that hypothesis were true $P(d \mid h)$. All Bayesian models thus capture a natural tradeoff between prior probability and likelihood, which often has an intuitive interpretation balancing between a sense of plausibility based on background knowledge on one hand and the data-driven sense of a "suspicious coincidence" on the other. If multiple hypotheses in some hypothesis space – some set of mutually exclusive hypotheses $\mathscr{H}$ – are being compared, then the posterior probability of each individual hypothesis h_i is calculated in the same way, and the hypothesis with the highest posterior probability is identified by dividing each by the total probability of all of the hypotheses in question:

$$P(h_i|d) = \frac{P(d|h_i)P(h_i)}{\sum_{h_j \in \mathscr{H}} P(d|h_j)P(h_j)} \tag{1}$$

The denominator in Equation 1 provides a normalizing term which is the sum of the probability of each of the possible hypotheses under consideration; this ensures that Bayes' Rule will reflect the proportion of all of the probability that is assigned to any single hypothesis h_i, and (relatedly) that the posterior probabilities of all hypotheses sum to one. This captures what we might call the "law of conservation of belief": a rational learner has a fixed "mass" of belief to allocate over different hypotheses, and the act of observing data just pushes this mass around to different regions of the hypothesis space. If the data lead us to strongly believe one hypothesis, we must decrease our degree of belief in all other hypotheses. By contrast, if the data strongly disfavor all but one hypothesis, then whichever remains, however, implausible *a priori*, is very likely to be the truth.

We can illustrate how this might work with a simple example. Suppose your friend Elmo has two bags: bag *A* contains four jellybeans (one black, one red, and two yellow) and bag *B* contains two (both yellow). Elmo then pulls a single yellow jellybean out of one of them. What is the probability that it is bag *B*? If h_A is the hypothesis that it is bag A and h_B is the hypothesis that it is bag B, then a prior capturing the fact that Elmo likes each bag equally is given by $P(h_A)=P(h_B)=0.5$. The likelihood of one yellow jellybean is 50% for bag *A* (two out of four jellybeans are yellow) and 100% for bag *B* (all jellybeans are yellow). Thus, the posterior probability of bag *B* given one yellow jellybean is 66.7%:

$$P(h_B|d) = \frac{(1.0)(0.5)}{(1.0)(0.5) + (0.5)(0.5)} = 0.667 \quad (2)$$

A bias for a certain kind of simplicity – a type of automatic Ockham's Razor – emerges naturally out of Bayesian modeling (MacKay 2003). As data accumulates, hypotheses that license more specific predictions will tend to be preferred over hypotheses that are consistent with many possibilities. Imagine that Elmo draws jellybeans two more times from the same bag, replacing the old draw and mixing thoroughly in between, and each time it is yellow. Intuitively it feels now feels even *more* likely to be bag B; this is because if it were bag A, it is an increasingly suspicious coincidence that all of the jellybeans are yellow. Bayes' Rule captures this: $P(h_B|d)$ given these three independent draws is now 88.9%.

Although this example is trivial, it captures an important characteristic of Bayesian modeling: the assumption that data are generated by some underlying process or mechanism. In this example, the data (jellybeans) are generated (drawn from) bags containing different jellybeans. In cognitive models, sentences may be generated by a grammar, observed events may be generated by some underlying network of causal relations, and words might be generated by a lexicon. A learner's goal is to evaluate different hypotheses about the underlying nature of the generative process, and to make predictions based on the most likely ones. A model is simply a specification of the generative process at work, identifying the steps (and associated probabilities) involved in generating data. Both priors and likelihoods are typically straightforward to define given a generative model; see Perfors *et al.* (2011a) for more details.

We can illustrate this with a slightly more realistic model. The left side of Fig. 1 shows three data points generated by a hidden process of some sort. Figure 1 also shows two possible hypotheses about that process, each corresponding to a Gaussian distribution (specified by its mean and variance) that contains the data points. Hypothesis h_A is consis-

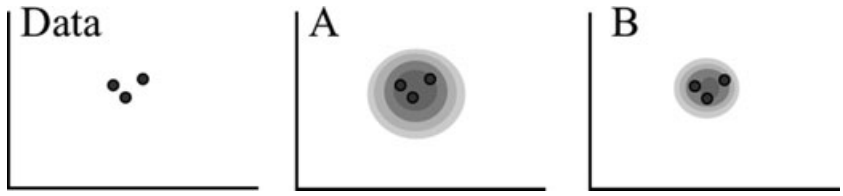

Fig.1. Example of generative model with two hypotheses. The left panel shows three data points generated by some sort of hidden process. *A* and *B* represent two hypotheses about the nature of the hypothesis. Both are Gaussian distributions consistent with the data, but hypothesis *A* predicts a wider variety of data points while *B* is more tightly clustered around the data. A Bayesian model would calculate the probability of each hypothesis based on its likelihood (degree of fit to the data) and prior probability.

tent with more distinct data points, while h_B is more tightly located around the observed data. The probability of each hypothesis can be calculated based on the likelihood of the data (in this case, the probability that each Gaussian would have generated each point) and the prior probability of each hypothesis (which would be set by the modeler, probably in the form of separate priors over expected means and variances). Although this example is still simplistic, it is slightly more cognitively interesting than the jellybean one: many simple categorization problems correspond to learning the nature of the distribution responsible for generating the exemplars within a category.

All Bayesian models calculate the posterior probability of a set of hypotheses and make predictions on the basis of those probabilities. Different Bayesian models differ on almost every other particular, and the interesting cognitive questions often center on the questions or assumptions incorporated into those particulars. One of the main points of difference is simply what the hypothesis space is. In the examples so far, the hypothesis spaces each contained only two hypotheses. However, hypothesis spaces can be any size, and are generally defined by the structure of the problem and the nature of the representation used. For instance, in the jellybean case, it is assumed that learners can represent jellybeans, bags, colors, and distributions of jellybeans within bags. This results in a hypothesis space of possible bags whose size is affected by the number of different colors jellybeans can have: such a space could be finite (if the bags were constrained to be of finite size) or infinite (if they were not). In the Gaussian example, the hypothesis space is defined by the dimension(s) of the metric space in which the data points exist, as well as the boundaries, if any. In problems of more cognitive interest, hypotheses might consist of grammars, words and their mappings, networks of cause and effect, graph structures, predicate logics, probability distributions in a metric space, and more. As hypothesis spaces get increasingly large, it is rare for the posterior probability of each individual hypothesis to be calculated. Instead, the hypothesis space is searched using computational techniques that sample hypotheses from the space and are guaranteed to eventually converge on the true distribution (see, Gilks *et al.* 1996; Doucet *et al.* 2001; Gelman *et al.* 2004).

It is clear that an intrinsic aspect of Bayesian modeling involves the specification of a hypothesis space. But, for a learner, where does that space come from? If Bayesian modeling simply consists of evaluating the probability of hypotheses in a pre-specified space, in what sense does it capture *learning* at all? In the next sections I consider this and related questions, including the nature of prior probabilities, what it means to be ''learnable'', and the role of modeling in cognitive science in general.

3. Pre-specification of the Hypothesis Space: What Does it Mean to Learn?

In some sense, Bayesian models do not appear to be learning at all. The entire hypothesis space, as well as the evaluation mechanism for comparing hypotheses (including the prior), has been given by the modeler; all the model does is search among and chose from hypotheses that already exist. Isn't development and learning – particularly the sort of learning that children perform over the first years of their life – something more than this? Shouldn't it encompass the discovery of some sort of genuine novelty? Our intuitive notion of learning certainly does not appear at first glance to be captured by a model that simply does hypothesis testing within an already-specified hypothesis space.

The same intuition lies at the core of Fodor's famous puzzle of concept acquisition (Fodor 1975, 1981). His essential point is that one cannot acquire new concepts via hypothesis testing because one must possess them in order to test them in the first place.

Therefore, except for those concepts that can be created by composing them from others (which Fodor believes excludes most lexical concepts), all concepts (including CARBURETOR and GRANDMOTHER) must be innate.

However, this intuition is misleading. To understand why this is, it is helpful to make a distinction between two separate notions of what it means to build in a hypothesis space. A hypothesis space is ''built in'' in a trivial sense if the model is equipped with the representational capacity to represent any of the hypotheses in the space: given this capacity, even if the model is not currently evaluating or considering any given hypothesis, that hypothesis is in some sense latent in that space. Let us denote the space of all hypotheses that can be represented as the *latent hypothesis space*. The ability to represent possible hypotheses in a latent hypothesis space is necessary for learning of any sort, in any model or being. The latent hypothesis space can be contrasted with the hypotheses that are being explicitly considered or evaluated – the hypotheses that are being actively represented and manipulated by the conceptual system – which I refer to as the *explicit hypothesis space*.

To make this distinction clear, we can use the analogy of a standard English typewriter with an infinite amount of paper. It is capable of producing a certain space of documents, which includes things like *The Tempest* and does not include, say, a Vermeer painting or a poem written in Russian. This typewriter represents the generative model for the hypothesis space of a Bayesian learner: each possible document that can be typed is a hypothesis, the infinite set of documents producible by the typewriter is the latent hypothesis space,[3] and the documents that have actually been typed out so far make up the explicit hypothesis space. Is there a difference between documents that have been created by the typewriter and documents that exist only in the latent hypothesis space? Of course there is: documents that have been created can be manipulated in all sorts of ways (reading, burning, discussing, editing) that documents latent in the space cannot.

In the same way, there may be a profound difference between hypotheses that can be evaluated by the learner and hypotheses that are simply latent in the space: the former can be manipulated by the cognitive system – evaluated, used in inference, compared to other hypotheses – but the latter cannot. A concept, when used in the sense of something that can be manipulated by the cognitive system in precisely these ways, is thus part of a learner's explicit hypothesis space. In this conceptualization, hypothesis generation thus describes the process by which hypotheses move from the latent space to the explicit space – the process by which our typist decides what documents to produce – and hypothesis testing describes the process of deciding which of the documents produced should be preferred (by whatever standard). Learning, then, corresponds to the entire process of hypothesis generation and testing: it would never involve new hypotheses being added to the latent hypothesis space. This is the part that doesn't ''feel'' like learning, because all of the hypotheses are built into the latent space from the beginning.

However, this intuitive feeling is misleading. If we take ''learning'' to mean ''discovering something that was not built into the latent hypothesis space'', then there are only two conclusions possible. Either the hypotheses appear in the latent hypothesis space completely arbitrarily, or *nothing* can ever be learned.

How is this so? Imagine that we could explain how a new hypothesis could be added to a latent hypothesis space; such an explanation would have to make reference to some rules or some kind of process for adding things, since that is what an explanation is. That process and those rules, however, would implicitly define a ''meta'' latent space of their own. And because this meta-space is pre-specified (implicitly, by that process or set of rules) in the exact same way the original hypothesis space was pre-specified (implicitly, by the original generative process), the hypotheses within it are ''innate'' in precisely the

same way that the hypotheses in the original latent space were. In general, the only way for something to be learned but not built into the latent hypothesis space is for it to be able to spring into a hypothesis space in such a way that is essentially random (i.e., unexplainable via some process or rule). If this is truly what learning is, it seems to preclude the possibility of studying it scientifically; but luckily, this is not what most of us generally mean by learning.

One implication of this is that Fodor's point is true but trivial. If one understands ''concept'' to mean ''something that can be represented by the brain'', then all of our concepts *are* innate – they exist in the latent hypothesis space of possible things the brain can represent (a space implicitly defined by the structure of the brain). In the more interesting sense, where ''having concept'' means that the concept is available at the cognitive level – it is capable of being manipulated by the cognitive system – then it need not be innate (i.e., having always been available at that level).

But isn't the process of hypothesis generation – the movement of a hypothesis from the latent space to the explicit space – itself a type of manipulation? It is indeed, but there is an important distinction between the two ways that hypotheses can be manipulated. Hypotheses in the explicit hypothesis space can be manipulated in all of the ways we intuitively think a cognitive system manipulates ideas: testing them, comparing them, etc. In terms of Bayesian modeling, these are the hypotheses that are evaluated using Bayes' Rule and used to generate predictions or inferences. However, a second type of manipulation involves searching the latent hypothesis space and identifying the hypotheses that will become explicit hypotheses. The reason this is a different kind of manipulation is that what is being manipulated is not the hypotheses themselves but the primitives that define those hypotheses in the first place.

To make sense of this, note that the discussion of latent spaces and meta-latent spaces actually implies a *hierarchy* of hypothesis spaces. On any given level of abstraction, as we have seen, the latent space is defined by some set of primitives or rules. In the typewriter example, the primitives might be the paper, the different keys on the typewriter, and the physical constraints on the movement of the typewriter (tapping of keys); this permits documents with rows of letters and spaces, but not documents that aren't made of paper, or letters floating in the air, or oil paintings. In the Gaussian example, the primitives would be the mean and variance (along with factors like the dimensions of the space, real numbers, the equation of a Gaussian, and certain inherent notions of magnitude that are necessary for such an equation to make sense).

These primitives correspond to the *hypotheses* of a higher-level space, as Fig. 2 illustrates. For instance, a typewriter is just one possible way that metal, plastic, and ink can be arranged and used to yield documents; different arrangements – like a Cyrillic keyboard – would define a different space of possible documents. Thus, an English-language typewriter and a Cyrillic typewriter correspond to two of many possible hypotheses in the space of document-producing machines. Similarly, Gaussians are but one option out of a space of ''possible equations''; a different choice of equations would define a different set of solutions. In principle, this could extend even higher, since ''the space of all equations'' itself is only defined based on the primitives of our mathematics: different elementary operations would define an entirely different set of equations to choose from. These examples are purely illustrative, and it is possible to invent spaces whose details might vary considerably;[4] the point here is simply to demonstrate how hypothesis spaces, by their nature, are located in a hierarchy of different levels of abstraction.

The notion of a hierarchy of hypothesis makes clear how hypothesis generation might work in principle. During hypothesis testing, hypotheses at one level are *compared as*

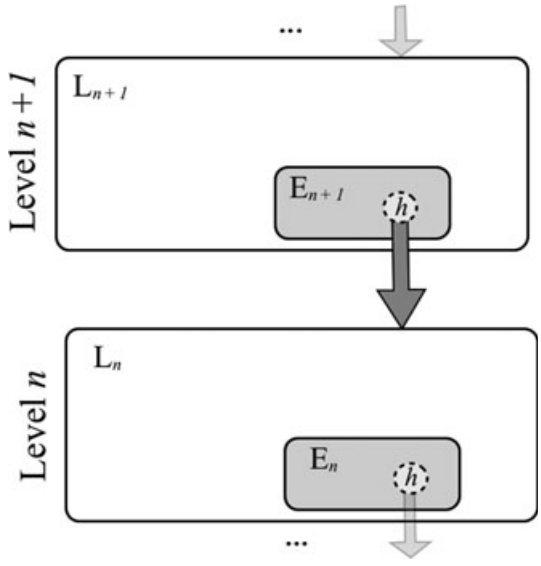

Fig. 2. Schematic illustration of a hierarchy of hypothesis spaces. The explicit hypothesis space at level n, E_n, is a subset of the latent hypothesis space at that level (L_n). That latent hypothesis is implicitly defined based on primitives which correspond to a single hypothesis h at some higher level L_{n+1}. In principle, the number of levels could extend further in either direction.

hypotheses with hypotheses at the same level, but during hypothesis generation a hypothesis is *constructed* by manipulating the hypotheses at a higher level (i.e., the primitives that define the latent space). Every Bayesian model searches the latent hypothesis space by doing some sort of manipulation of the primitives that help to define that space. For instance, in the Gaussian example each hypothesis is a Gaussian with some mean and variance, and individual hypotheses about Gaussians are compared based on their probability as calculated by Bayes' Rule. However, new hypotheses are *found* by manipulating the values of the primitives: for instance, by adding or subtracting from the means or variances of existing hypotheses. Within the typewriter analogy, new documents would be generated by pressing different keys in different orders.

What does this all mean in terms of Fodor's argument? Fodor argues that hypothesis testing requires that the concept (i.e., the hypothesis) be represented explicitly – that is, that it be in the explicit hypothesis space. His conclusion from this is that one must therefore already explicitly possess concepts. I am arguing that the conclusion doesn't follow from the premise: hypothesis testing does indeed require that the hypothesis be in the explicit hypothesis space, but hypotheses don't *enter* the explicit space (i.e., come to be possessed) through hypothesis testing – or at least not hypothesis testing at *that* level. They enter the explicit hypothesis space through testing at a level whose primitives do not correspond to verbalizable concepts.

There is thus no reason that hypotheses at one level should decompose or be expressible in terms of hypotheses at the same level. That is, we shouldn't expect concepts to be decomposable *into other concepts* (i.e., other hypotheses that can be manipulated by the cognitive system). They might be (and probably are) decomposable into primitives of some sort, but because those are on a different level, they shouldn't be available to the cognitive system for verbalization: verbalization is a particular kind of manipulation by the cognitive system that applies only to hypotheses at that level. This reasoning reveals

another facet of the same problem with Fodor's argument that we have seen already: our inability to decompose concepts into other concepts does not imply that they must all be innate, except in the most trivial and uninteresting sense of the word ''innate'' (i.e., the system has the representational capacity to capture them).

So far I have suggested that Bayesian modeling makes clear how it may be possible to have true learning even when a hypothesis space is fully pre-specified – indeed, I argue that a pre-specification of a latent hypothesis space is *necessary* for learning. The next section briefly discusses some of the implications of this reasoning.

4. Additional Implications

If the notion of ''learning'' as ''discovering something that was not built into the latent hypothesis space'' is incoherent, then *every* learning system – including the brain – must come equipped with a latent hypothesis space consisting of everything that it can possibly represent. In some respects this is not a controversial point, but it is an easy point to forget when evaluating Bayesian models: the fact that hypothesis spaces are clearly defined within the Bayesian framework makes them appear more ''innate'' than if they were simply implicit in the model. But even neural networks – which are often believed to presume very little in the way of innate knowledge – implicitly define hypotheses and hypothesis spaces via their architecture, functional form, learning rules, etc. In fact, neural networks can be viewed as implementations of Bayesian inference (e.g., Funahashi 1998; McClelland 1998), corresponding to a computational-level model whose hypothesis space is a set of continuous functions (e.g., Funahashi 1989; Stinchcombe and White 1989) This is a large space, but Bayesian inference can entertain hypothesis spaces that are equivalently large.

If any model, including the brain, is equipped with a latent hypothesis space defined by its primitives and rules, doesn't this imply that the best way to study the brain is by seeking to understand those primitives and rules? In the brain, investigating the primitives would correspond to investigating the neuronal level. However, the primitives used in most Bayesian models (so far) do not consist of anything that looks like neurons; they are instead usually chosen for their mathematical elegance and ability to produce solutions that seem well-matched to the learning problem. Does this not suggest that Bayesian modeling is fundamentally approaching the issues from the wrong direction?

This question is a topic of much debate in cognitive science (Griffiths *et al.* 2010; McClelland *et al.* 2010). My view, and that of many computational modelers, is that both directions are necessary. That said, there are several reasons why a computational-level approach may be useful. First, we currently have very little idea about how the actual primitives of the brain combine to form anything approaching conceptual knowledge.[5] Models that build in primitives and explore what emerges are useful for showing *that* something can be learned by neurons with those characteristics – but it can be difficult to interpret what precisely was learned or why. Furthermore, it is not even clear that we could learn how the primitives underlie conceptual knowledge without having some sense of what that conceptual knowledge *is*. Bayesian modeling focuses on that critical question, seeking to characterize the abstract nature of the problems facing the learner, what representational abilities the learner must have in order to deal with those problems, and what a good solution would look like. It is true that Bayesian models assume that the way in which the brain *realizes* those representations is at least somewhat irrelevant to answering questions on the computational level, but if this assumption is incorrect, the best way to realize that is to develop the models and see where they fail and why.

On a more general level, because all models implicitly define a hypothesis space, it does not make sense to compare models according to whether they build hypothesis spaces in. More interesting questions are: What is the size of the latent hypothesis space defined by the model? How strong or inflexible is the prior? (All models and learners that generalize at all, not just Bayesian ones, define a prior: even the assumption that all hypotheses are equally likely is a prior of its own.) A model that compares two hypotheses is far more restrictive (i.e., builds far more in) than a model that compares all possible Gaussians; a model that places high prior probability on one specific hypothesis builds more in than one that ''spreads'' probability mass evenly over many. Although it is straightforward to compare Bayesian models according to these metrics, it is more difficult to compare models in which the priors and hypotheses are implicit in the architecture or learning rule. Nevertheless, these are precisely the questions that should be asked.

The notion of a hierarchy of hypothesis spaces illuminates another way in which a model might build very little in: by simultaneously searching over hypotheses on multiple levels. All things being equal, searching in higher-level spaces opens up many additional hypotheses on lower levels and can result in hypotheses that look extremely novel on that level (just as adding a Cyrillic keyboard would make the resulting documents appear very novel relative to any English ones). Bayesian models that search over multiple levels of hypotheses are called hierarchical Bayesian models, and are increasingly common (e.g., Lee 2006; Kemp *et al.* 2007; Heller *et al.* 2009; Kemp *et al.* 2010; Perfors *et al.* 2010; Perfors *et al.* 2011b). For instance, a hierarchical version of the model in the Gaussian example might compare different higher-level hypotheses about the *shape* of the distribution (Gaussian, multinomial, etc). The modeler would specify how the likelihoods and priors for each distribution type would be calculated and the model would evaluate which specific distributions of each type best captured the data, as well as which distribution type *itself* was best. Knowledge is still built in, but the hypothesis space is much larger and the constraints built into the model are weaker. One interesting and unexpected outcome of these models is that successful learning in such spaces does not necessarily require more data than learning in smaller spaces (Kemp *et al.* 2010; Perfors *et al.* 2011b).

There is one other important respect in which Bayesian models have implications for innateness. In many Bayesian models, adequately sampling from a hypothesis space is a difficult technical problem, but is not the cognitive problem of interest; the goal is to identify the hypotheses with the highest probability, derive predictions on that basis, and ascertain how well that matches up with human behavior. This ''ideal learning'' approach is helpful for identifying what optimal[6] behavior would look like, given the data and the assumptions made explicit in the model. This is useful in two ways: first, by examining models that incorporate different assumptions, or given different data, we can rigorously investigate the effect of those assumptions or that data. Second, by comparing model performance to human behavior, we can investigate to what extent to what extent human learning can be understood as optimal performance – and, if it falls short, in what ways it does so.

However, there is another sense in which we might want to investigate the learnability of some knowledge: can a learner efficiently search the hypothesis space and find it? If this is the question, then suddenly many details matter. Where in the hypothesis space do searches begin? How long does it take to search efficiently? How is the search done? Although still in its infancy, some research within the Bayesian framework is beginning to address these kinds of questions. Rational process models attempt to capture different sorts of capacity constraints or resource limitations by limiting the breadth or extent of the search of the hypothesis space (e.g., Levy *et al.* 2008; Vul *et al.* 2009; Sanborn *et al.* 2010). Some models also investigate the effects of different types of memory constraints

on what inferences become favored (Perfors 2011) or what sort of representations are sensible (Navarro and Perfors 2011). It remains to be seen to what extent the possibilities that are easy to capture within the Bayesian framework correspond to the limitations people actually have, but this is a promising area for further research.

Short Biography

Amy Perfors is a lecturer in the School of Psychology at the University of Adelaide in Australia (equivalent to an assistant professor in the United States). Her research interests include language acquisition and evolution, conceptual representation and learning, and hypothesis generation and testing. She employs a combination of methodologies, including experiments on adults (and sometimes children) as well as computational modeling, generally within the Bayesian framework. She did her PhD research at MIT with Joshua Tenenbaum, graduating in 2008 with a thesis entitled *Learnability, representation, and language: A Bayesian approach*. In her spare time Amy enjoys violent sports (playing, not watching), learning to spell the Australian way with all of those extra u's, and playing scrabble.

Notes

⋆ Correspondence: Level 4, Hughes Building, University of Adelaide, Adelaide, SA 5005, Australia. Email: amy.perfors@adelaide.edu.au.

[1] Special issues of two journals have been devoted to Bayesian modeling (*Cognition* 120, 2011; *Trends in Cognitive Science* 10.7, 2006). See also Griffiths *et al.* (2010); McClelland *et al.* (2010), and Jones and Love (2011) and replies for additional papers exploring some of the fundamental issues involved in interpreting and applying these models.

[2] Perfors *et al.* (2011a) contains a more thorough overview.

[3] Note that the latent hypothesis space does not need to be completely enumerated in order to exist; it must simply be defined by some sort of process or procedure. In the case of Bayesian models, that process is the generative model for the data, from which the priors and likelihoods are defined. As noted earlier in practice, exhaustive hypothesis enumeration is intractable for all but the simplest models; most perform inference via guided search, and only a subset of the hypotheses within the space are actually evaluated.

[4] It is of course possible to define hypotheses that do not serve as primitives in any lower-level spaces. However, the reverse is not true: the primitives of any space can be viewed as one of many hypotheses in a higher-level space.

[5] Connectionist models only loosely approximate the behavior of neurons in the brain.

[6] Bayesian reasoning is ''statistically optimal'' in the sense that a non-Bayesian reasoner attempting to predict the future will always be out-predicted by a Bayesian reasoner in the long run (de Finetti 1980). As a first approximation, just as formal logic describes a deductively correct way of thinking, Bayesian probability theory describes an inductively correct way of thinking. For further discussion of these issues please see Jeffreys (1931, 1939); Cox (1946, 1961); Jaynes (2003).

Works Cited

Anderson, J. 'The Adaptive Nature of Human Categorization.' *Psychology Review* 98.3 (1991): 409–29.

Baker, C., J. B. Tenenbaug and R. Saxe. 'Goal Inference as Inverse Planning.' *Proceedings of the 29th Annual Conference of the Cognitive Science Society*. Eds. D. McNamara and J. Trafton. Austin, TX: Cognitive Science Society, 2007.

Cox, R. *The Algebra of Productive Inference*. Baltimore, MD: Johns Hopkins University Press, 1961.

——. 'Probability, Frequency, and Reasonable Expectation.' *American Journal of Physics* 14 (1946): 1–13.

Doucet, A., N. D. Freitas and N. Gordon. *Sequential Monte Carlo in Practice*. New York: Springer-Verlag, 2001.

Feldman, J. and Tremoulet, P. *The Attribution of Mental Architecture from Motion: Towards a Computational Theory* (Tech. Rep. No. RuCCS TR-87). Rutgers University, 2008.

Feldman, N. and T. L. Griffiths. 'Learning Phonetic Categories by Learning a Lexicon.' *31st Annual Conference of the Cognitive Science Society*. Eds. N. Taatgen, H. van Riji, L. Schomaker and J. Nerbonne. Austin, TX: Cognitive Science Society, 2009.

de Finetti, B. 'Foresight Its Logical laws its Subjective Sources.' *Studies in Subjective Probability* (2nd ed.). Eds. H. Kyburg and H. Smokler. New York: J. Wiley and Sons, 1980. 53–118.
Fodor, J. *The Language of Thought*. New York: Thomas Y. Crowell Company, 1975.
——. *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, MA: MIT Press, 1981.
Frank, M., N. Goodman and J. B. Tenenbaum. 'Using Speakers' Referential Intentions to Model Early Cross-Situational Word Learning.' *Psychological Science* 20.5 (2009): 578–85.
Funahashi, K. 'Multilayer Neural Networks and Bayes Decision Theory.' *Neural Networks* 2 (1998): 209–13.
——. 'On the Approximate Realization of Continuous Mappings by Neural Networks.' *Neural Networks* 2 (1989): 183–92.
Gelman, A. et al. *Bayesian Data Analysis* (2nd ed.). New York: Chapman & Hall, 2004.
Gilks, W., S. Richardson and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. New York: Chapman & Hall, 1996.
Goldwater, S., T. L. Griffiths and M. Johnson. 'Bayesian Framework for Word Segmentation: Exploring the Effects of Context.' *Cognition* 112 (2009): 21–54.
Goodman, N. et al. 'A Rational Analysis of Rule-Based Concept Learning.' *Proceedings of the 29th Annual Conference of the Cognitive Science Society*. Eds. D. McNamara and J. Trafton. Austin, TX: Cognitive Science Society, 2007.
Griffiths, T. L. et al. 'Probabilistic Models of Cognition: Exploring Representations and Inductive Biases.' *Trends in Cognitive Sciences* 14.8 (2010): 357–64.
—— and M. KalishLanguage Evolution by Iterated Learning with Bayesian Agents.' *Cognitive Science* 31 (2007): 441–80.
——, M. Steyvers and J. B. Tenenbaum. 'Topics in Semantic Representation.' *Psychological Review* 114.2 (2007): 211–44.
—— and J. B. Tenenbaum. 'Theory-Based Causal Induction.' *Psychological Review* 116 (2009): 661–716.
Heller, K., A. Sanborn and N. Chater. 'Hierarchical Learning of Dimentional Biases Human Categorization.' *Advances in Neutral Information Processing Systems* (Vol. 22). Eds. Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams and A. Culotta. Cambridge, MA: MIT Press, 2009. 727–35.
Jaynes, E. *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press, 2003.
Jeffreys, H. *Scientific Inference*. Cambridge: Cambridge University Press, 1931.
——. *Theory of Probability*. Oxford: Clarendon Press, 1939.
Jones, M. and B. Love. 'Bayesian Fundamentalism or Enlightenment? On the Explanatory Status and Theoretical Contributions of Bayesian Models of Cognition.' *Behavioral and Brain Science* 34.4 (2011): 169–321.
Kemp, C. et al. 'A Probabilistic Model of Theory Formation.' *Cognition* 114.2 (2010): 165–96.
——, A. Perfors and J. B. Tenenbaum. 'Learning Overhypotheses with Hierarchical Bayesian Models.' *Developmental Science* 10.3 (2007): 307–21.
Lee, M. 'A Hierarchical Bayesian Model of Human Decision-Making on an Optimal Stopping Problem.' *Cognitive Science* 30.3 (2006): 555–80.
Levy, R., F. Reali and T. L. Griffiths. 'Modeling the Effects of Memory on Human Online Sentence Processing with Particle Filters.' *Advances in Neutral Information Processing Systems* (Vol. 21). Eds. D. Koller, D. Schuurmans, Y. Bengio, L. Bottou. Cambridge, MA: MIT Press, 2008. 937–44.
MacKay, D. *Information Theory Inference and Learning Algorithms*. Cambridge UK: Cambridge University Press, 2003.
Marr, D. Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information. San Francisco: WH Freeman & Company, 1982.
McClelland, J. 'Connectionist Models and Bayesian Inference.' *Rational Models of Cognition*. Eds. M. Oaksford and N. Chater. Oxford: Oxford University Press, 1998. 21–53.
—— et al. Letting Structure Emerge: Connectionist and Dynamic Systems Approaches to Cognition.' *Trends in Cognitive Sciences* 14.8 (2010): 348–56.
Navarro, D. and A. Perfors. 'Hypothesis Generation the Positive Test Strategy and Sparse Categories.' *Psychological Review* 118 (2011): 120–34.
Pearl, J. *Causality: Models Reasoning and Inference*. Cambridge: Cambridge University Press, 2000.
Perfors, A. 'Memory Limitations Alone Do Not Lead to Over-Regularization: An Experimental and Computational Investigation.' *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, 2011.
—— et al. 'A Tutorial Introduction to Bayesian Models of Cognitive Development.' *Cognition* 120 (2011a): 302–21.
——, J. B. Tenenbaum and T. Regier. 'The Learnability of Abstract Syntactic Principles.' *Cognition* 118.3 (2011b): 306–38.
——,—— and——. 'Variability Negative Evidence and the Acquisition of Verb Argument Constructions.' *Journal of Child Language* 37 (2010): 607–42.
Sanborn, A., T. L. Griffiths and D. Navarro. 'Rational Approximations to Rational Models: Alternative Algorithms for Category Learning.' *Psychological Review* 117 (2010): 1144–67.

Steyvers, M. et al. 'Inferring Casual Networks from Observations and Interventions.' *Cognitive Science* 27 (2003): 453–89.

Stinchcombe, M. and H. White. 'Universal Approximation using Feedforward Networks with Non-sigmoid Hidden Layer Activation Functions.' *Proceedings of the International Joint Conference on Neural Networks*. Washington D.C., June 18–22, (1989). 607–11.

Vul, E. et al. 'Explaining Human Multiple Object Tracking as a Resource-Constrained Approximate Inference in a Dynamic Probabilistic Model.' *Advances in Neural Information Processing Systems* (Vol. 22). Eds. Y. Bengio and D. Schuurmans, J. Lafferty, C. K. I. Williams and A. Culotta. Cambridge, MA: MIT Press, 2009. 1955–63.

Xu, F. and J. B. Tenenbaum. 'Word Learning as Bayesian Inference.' *Psychological Review* 114.2 (2007): 245–72.