A tutorial introduction to Bayesian models of cognitive development

Amy Perfors

School of Psychology, University of Adelaide

Joshua B. Tenenbaum

Brain and Cognitive Sciences, Massachusetts Institute of Technology

Thomas L. Griffiths

Fei Xu

Department of Psychology, University of California, Berkeley

Abstract

We present an introduction to Bayesian inference as it is used in probabilistic models of cognitive development. Our goal is to provide an intuitive and accessible guide to the *what* and the *why* of the Bayesian approach: what sorts of problems and data the framework is most relevant for, and how and why it may be useful for dvelopmentalists. We emphasize a qualitative understanding of Bayesian inference, but also include information about additional resources for those interested in the mathematical foundations, machine learning details, or cognitive science applications in more depth. We also discuss some important interpretation issues that often arise when evaluating Bayesian models in cognitive science.

Keywords: Bayesian models; cognitive development

1 Introduction

One of the most basic questions in cognitive development is how we learn so much from such apparently limited evidence. In learning about causal relations, object categories or their properties, in acquiring language or constructing intuitive theories, children routinely draw inferences that go beyond the data they observe. Probabilistic models provide a general-purpose computational framework for exploring how human learners might make these inductive leaps, explaining them as forms of Bayesian inference.

This paper presents a tutorial overview of the Bayesian framework for studying cognitive development. Our goal is to provide an intuitive and accessible guide to the *what* and the *why* of the Bayesian approach: what sorts of problems and data the framework is most relevant for, and how and why it may be useful for developmentalists. We consider three general inductive problems that learners face, each grounded in one or two specific developmental challenges:

1. Inductive generalization from examples, with a focus on learning the referents of words for object categories.

2. Acquiring inductive constraints, tuning and shaping priors from experience, with focus on learning to learn categories.

3. Learning inductive frameworks, constructing or selecting appropriate hypothesis spaces for inductive generalization, with applications to acquiring intuitive theories of mind and inferring hierarchical phrase structure in language.

We also discuss several general issues as they bear on the use of Bayesian models: assumptions about optimality, biological plausibility, and what idealized models can tell us about actual human minds.

2 Basic Bayes: Inductive generalization from examples

The most basic question the Bayesian framework addresses is how to update beliefs and make inferences in light of observed data. In the spirit of Marr's (1982) computational-level of analysis, it begins with understanding the logic of the inference made when generalizing from examples, rather than the algorithmic steps or specific cognitive processes involved. A central assumption is that degrees of belief can be represented as probabilities: that our conviction in some hypothesis h can be expressed as a real number ranging from 0 to 1, where 0 means something like "h is completely false" and 1 that "h is completely true." Such an assumption turns the mathematics of probability theory into an engine of inference, a means of weighing each of a set of mutually exclusive and exhaustive hypotheses $\mathcal{H}$ to determine which best explain the observed data. Bayes' Rule tells us to compute the posterior probability $P(h_i|D)$, or degree of belief in hypothesis h_i conditional on observing the data D, as follows:

$$P(h_i|D) = \frac{P(D|h_i)P(h_i)}{\sum_{h_j \in \mathcal{H}} P(D|h_j)P(h_j)}. \quad (1)$$

The score of each hypothesis h_i depends on the product of (1) its prior probability $P(h_i)$, or how much we believe in h_i independent of (or prior to) observing the data D; and (2) the likelihood $P(D|h_i)$, which captures the probability with which we would expect to observe D assuming h_i were true. As we will see below, the product of priors and likelihoods often has an intuitive interpretation in terms of balancing between a general sense of plausibility based on background knowledge and the data-driven sense of a "suspicious coincidence." In other words, it captures the tradeoff between the complexity of an explanation and how well it fits the observed data.

The score of each hypothesis is computed relative to the sum of scores for all hypotheses in the learner's hypothesis space $\mathcal{H}$, ensuring that posterior probabilities sum to 1 over all hypotheses. This represents what has been called 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, they must decrease beliefs in all other hypotheses. By contrast, if the data strongly disfavor all but one hypothesis, then (to paraphrase Sherlock Holmes) whichever remains, however implausible *a priori*, is very likely to be the truth.

The Bayesian framework is generative, meaning that observed data are assumed to be generated by some underlying process or mechanism explaining why the data occurs in the patterns it does. For instance, words in a language may be generated by a grammar of some sort, in combination with social and pragmatic factors. In a physical system, observed events may be generated by some underlying network of causal relations. The job of the learner is to evaluate different hypotheses about the underlying nature of the generative process, and to make predictions based on the most likely ones. A probabilistic model is simply a specification of the generative processes at work, identifying the steps (and associated probabilities) involved in generating data. Both priors and likelihoods are typically describable in generative terms.

We can illustrate this in the context of a schematic example, depicted graphically in Figure 1. The dots represent individual datapoints (e.g., words or events) generated independently from some unknown process (e.g., a language or a causal network) that we depict in terms of a region or subset of space: the process generates datapoints randomly within its region, never outside. Each hypothesis encodes a different idea about which subset of space the data are drawn from. According to the hypothesis shown in Figure 1a, for example, the data are generated by a relatively simple process, operating over a single rectangle in the space. If a learner believes this hypothesis provides a good account of the observed data, and thus places high posterior probability on it, then she should predict that a new datapoint is far more likely to be observed at position a than at position b, even if she had previously seen data at neither point. In this sense, inferring the hypotheses most likely to have generated the observed data

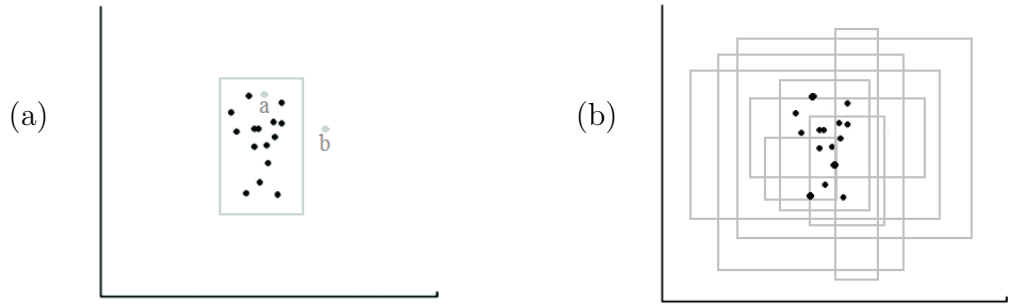

Figure 1: (a) Example data and hypothesis. Graphical representation of data and one possible hypothesis about the how that data was generated. This hypothesis suggests that there is one moderately-sized generative process, depicted as a single grey rectangle. Under this hypothesis, a new datapoint in position a is far more likely than one in b. (b) Hypothesis space for this example data. Hypotheses consist of rectangles; some are closely supported by the data and some are not.

guides the learner in generalizing beyond the data to new situations.

The hypothesis space $\mathcal{H}$ can be thought of as the set of all possible hypotheses, defined by the structure of the problem that the learner can entertain. Figure 1b shows a possible hypothesis space for our example, consisting of all possible rectangles in this space. Note that this hypothesis space is infinite in size, although just a few representative hypotheses are shown.

The hypothesis space here is assumed to be defined by the nature of the learning problem, and thus provided to the learner *a priori*. For instance, in our example, the hypothesis space would be constrained by the range of possible values for the lower corner (x and y), length (l), and width (w) of rectangular regions. Such constraints, even if built into the model, need not be very strong or very limiting: for instance, one might simply specify that the range of possible values for x, y, l, and w lies between 0 and some extremely large number like 10^9, or be drawn from a probability distribution with a very long tail. In this sense, the prior probability of a hypothesis $P(h_i)$ is also given by a probabilistic generative process – a process operating "one level up" from the process indexed by each hypothesis that generates the observed datapoints. We will see below how these hypothesis spaces and priors need not be built in, but can be constructed or modified from experience.

In our example the hypothesis space has a very simple structure, but because a Bayesian model can be defined for any well-specified generative framework, inference can operate over any representation that can be specified by a generative process.This includes, among other possibilities, probability distributions in a space (appropriate for phonemes as clusters in phonetic space); directed graphical models (appropriate for causal reasoning); abstract structures including taxonomies (appropriate for some aspects of conceptual structure); objects as sets of features (appropriate for categorization and object understanding); word frequency counts (convenient for some types of semantic representation); grammars (appropriate for syntax); argument structure frames (appropriate for verb knowledge); Markov models (appropriate for action planning or part-of-speech tagging); and even logical rules (appropriate for some aspects of conceptual knowledge).

This representational flexibility allows us to move beyond some of the traditional dichotomies that have shaped decades of research in cognitive development: structured knowledge vs. probabilistic learning (but not both), or innate structured knowledge vs. learned unstructured knowledge (but not the possibility of knowledge that is both learned and structured). As a result of this flexibility, traditional critiques of connectionism that focus on their inability to adequately capture compositionality and systematicity (e.g., Fodor & Pylyshyn, 1988) do not apply to Bayesian models. In fact, there are several recent examples of Bayesian models that embrace language-like or compositional representations in domains ranging from causal induction (Griffiths & Tenenbaum, 2009) to grammar learning (Perfors, Tenenbaum, & Regier, submitted) to theory acquisition (Kemp, Tenenbaum, Niyogi, & Griffiths, 2010).

2.1 A case study: learning names for object categories

To illustrate more concretely how this basic Bayesian analysis of inductive generalization applies in cognitive development, consider the task a child faces in learning names for object categories. As Quine (1960) pointed out, learning the names for object categories

is a classic problem of induction. Even an apparently simple word like "dog" can refer to a potentially infinite number of hypotheses, including all dogs, all Dalmatians, all mammals, all animals, all pets, all four-legged creatures, all dogs except Chihuahuas, all things with fur, all running things, etc. Despite the sheer number of possible extensions of the word, young children are surprisingly adept at acquiring the meanings of words – even when there are only a few examples, and even when there is no systematic negative evidence (Markman, 1989; Bloom, 2000).

How do children do this? One suggestion is that infants are born equipped with strong prior knowledge about what sort of word meanings are natural (Carey, 1978; Markman, 1989), which constrains the possible hypotheses considered. For instance, even if a child is able to rule out part-objects as possible extensions, she cannot know what level of a taxonomy the word applies: whether "dog" actually refers to dogs, mammals, Dalmatians, canines, or living beings. One solution would be to add another constraint – the presumption that count nouns map preferentially to the basic level in a taxonomy (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). This preference would allow children to learn names for basic-level categories, but would be counter-productive for every other kind of word.

Xu and Tenenbaum (2007) present a Bayesian model of word learning that offers a precise account of how learners could make meaningful generalizations from one or a few examples of a novel word. This problem can be schematically depicted as in Figure 2a[1] (and equivalently, as dots, as in Figure 2b): the child, shown one or many objects with a given label, must decide which hypothesis about possible extensions of the label is best. The hypotheses are overlapping, which is appropriate for the problem of inference in a hierarchical taxonomy as well as many other semantic domains. Intuitively, we would expect that when given one object, a reasonable learner should not strongly prefer any of the hypotheses that include it, though the more restricted ones might be slightly favored (as in Figure 2b). If the learner were shown *three* examples, as in Figure 2c, we would expect the most closely-fitting hypothesis to be much more strongly preferred.

[1]Picture reproduced from Xu and Tenenbaum (2007).

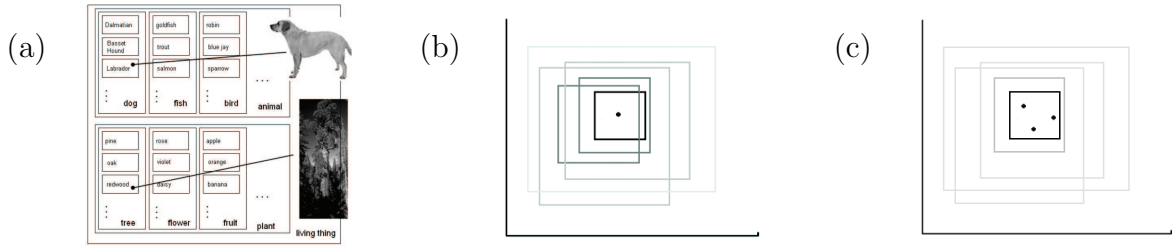

Figure 2: Learning object words. (a) Schematic view of hypotheses about possible extensions considered by the learner; because the taxonomy is hierarchical, the hypotheses are nested within each other. (b) Equivalent hypothesis space depicted as a two-dimensional dot diagram; hypotheses with higher probability are darker rectangles. With one data point, many hypotheses have some support. (c) With three examples, the most restrictive hypothesis is much more strongly favored.

For instance, given one Dalmatian as an example of a fep, it is unclear whether fep refers to Dalmatians, dogs, mammals, or animals. But if given three Dalmatians only, it would be odd indeed if fep meant something more general than Dalmatian.

A Bayesian analysis explains this pattern of inference via the likelihood $p(d|h)$, the probability of observing the data d assuming hypothesis h is true. This account is illustrated intuitively in Figure 2b and c. In general, more restrictive hypotheses, corresponding to smaller regions in the data space, receive more likelihood for a given piece of data. If a small hypothesis is the correct extension of a word, then it is not too surprising that the examples occur where they do; a larger hypotheses could be consistent with the same data points, but explains less well exactly why the data fall where they do. More formally, if we assume data sampled uniformly at random from all cases consistent with the concept, then the probability of any single data point d consistent with h is inversely proportional to the size of the region h picks out – call this the "size of h". Assuming further that multiple data points are generated independently from the concept, the likelihood of h with n consistent examples is inversely proportional to the size of h, raised to the nth power. Thus the preference for smaller consistent hypotheses over larger hypotheses increases exponentially with the number of examples. This assumption is often referred to as the size principle (Tenenbaum & Griffiths, 2001).

The size principle explains how it is possible to make strong inferences based on very

few examples. It also captures the notion of a suspicious coincidence: as the number of examples increases, hypotheses that make specific predictions – those with more explanatory power – tend to be favored over those that are more vague. This provides a natural solution to the "negative evidence" problem: deciding among hypotheses given positive-only examples. As the size of the data set approaches infinity, a Bayesian learner rejects larger or more overgeneral hypotheses in favor of more precise ones. With limited amounts of data, the Bayesian approach can make more subtle predictions, as the graded size-based likelihood trades off against the preference for simplicity in the prior. The likelihood in Bayesian learning can thus be seen as a principled quantitative measure of the weight of implicit negative evidence – one that explains both how and when overgeneralization should occur.

The results of the Xu and Tenenbaum (2007) experiment reflect this idea. Adults and 3- and 4-year-old children were presented with 45 objects distributed across three different superordinate categories (animals, vegetables, and vehicles), including many basic-level and subordinate-level categories within those. Subjects were then shown either one or three labelled examples of a novel word such as "fep", and were asked to pick out the other "feps" from the set of objects. Both children and adults responded differently depending on how many examples they were given. Just as in Figure 2, with one example, subjects and the model both showed graded generalization from subordinate to superordinate matches. By contrast, when given three examples, generalizations became much sharper and were usually limited to the most restrictive level.

So far we have illustrated how Bayesian inference can capture generalization from just a few examples, the simultaneous learning of overlapping extensions, and the use of implicit negative evidence. All of these are important, but it is also true that we have built in a great deal, including a restricted and well-specified hypothesis space. Very often, human learners must not make reasonable specific generalizations within a set hypothesis space, they also much be able to make generalizations *about* what sort of generalizations are reasonable. We see an example of this in the next section.

3 Acquiring inductive constraints

One of the conclusions of Quine's famous thought experiment was the need for generalizations about generalizations, or inductive constraints, of some sort. The core problem he raised is how induction is justified based on a finite sample of any kind of data, and the inevitable conclusion is that there must be some kind of constraint that enables learning to occur. Nearly every domain studied by cognitive science yields evidence that children rely on higher-level inductive constraints. Children learning words prefer to apply them to whole objects rather than parts (Markman, 1990). Babies believe that agents are distinct from objects in that they can move without contact (Spelke, Phillips, & Woodward, 1995) and act in certain ways in response to goals (Woodward, 1998; Gergely & Csibra, 2003). Confronted with evidence that children's behavior is restricted in predictable ways, the natural response is to hypothesize the existence of innate constraints, including the whole object constraint (Markman, 1990) core systems of object representation, psychology, physics, and biology (Carey & Spelke, 1996; Spelke & Kinzler, 2007; Carey, 2009), and so on. Given that they appear so early in development, it seems sensible to postulate that these constraints are innate rather than learned.

However, it may be possible for inductive constraints to be learned, at least in some cases. For instance, consider the problem of learning that some features "matter" for categorizing new objects while others should be ignored (e.g., Nosofsky, 1986). Acquiring higher-level abstract knowledge would enable one to make correct generalizations about an object from a completely novel category, even after seeing only one example. A wealth of research indicates that people are capable of acquiring this sort of knowledge, both rapidly in the lab (Perfors & Tenenbaum, 2009) and over the course of development (Landau, Smith, & Jones, 1988; L. Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). Children also acquire other sorts of inductive constraints over the course of development, including the realization that categories may be organized taxonomically (Rosch, 1978), that some verbs occur in alternating patterns and others

don't (e.g., Pinker, 1989) or that comparative orderings should be transitive (Shultz & Vogel, 2004).

How can an inductive constraint be learned, and how might a Bayesian inference explain this? Is it possible to acquire an inductive constraint faster than the specific hypotheses it is meant to constrain? If not, how can we explain people's learning in some situations? If so, what principles explain this acquisition?

A familiar example of the learning of inductive constraints was provided by N. Goodman (1955). Suppose we have many bags of colored marbles and discover that some bags have black marbles while others have white marbles. However, every bag is uniform in color; no bag contains both black and white marbles. Once we realize this, we have acquired knowledge on two levels: the item-based knowledge about the color of marbles in each particular bag, but also the higher-level knowledge (called, following Goodman, an *overhypothesis*) that bags tend to be uniform in color.

We can illustrate a similar idea in the rectangle world by imagining a learner who is shown the schematic data in Figure 3a. Having seen point a only, the learner has no way to decide whether b or c is more likely to be in the same category or region as a. However, if the learner has also seen the data in Figure 3b, it might infer both first-order and second-order knowledge about the data set. First-order learning refers to the realization that the rectangular regions constitute the best explanation for the datapoints seen so far; second-order (overhypothesis) learning would involve realizing that the regions tend to be long, thin, and oriented along the y-axis. Just as learning the how categories are organized helps children generalize from new items, this type of higher-order inference helps with the interpretation of novel data, leading to the realization that point b is probably in the same region as a but point c is not, even though b and c are equidistant from a.

A certain kind of Bayesian model, known as a hierarchical Bayesian model (HBM), can learn overhypotheses by not only choosing among specific hypotheses, but by also making higher-order generalizations about those hypotheses. As we've already seen, in

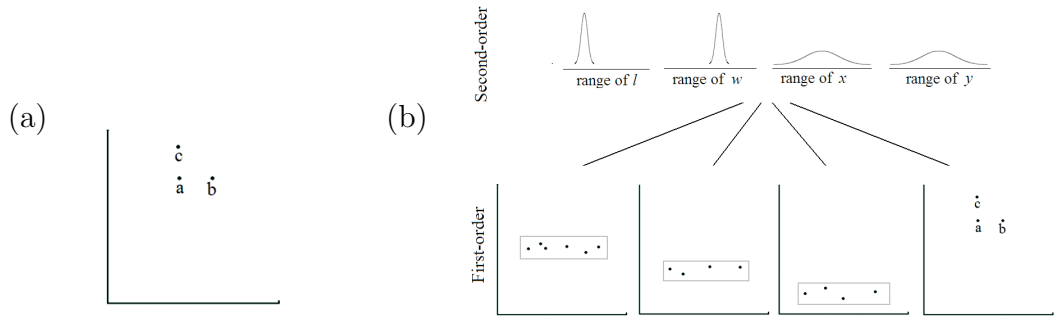

Figure 3: Learning higher-order information. (a) Given point a, one cannot identify whether b or c is more likely. (b) Given additional data, a model that could learn higher-order information about hypotheses might realize that regions tend to be long, thin rectangles oriented along the y axis: that the length l tends to be short, the width w tends to be long, and the location (x and y coordinates) can be anywhere. If this is the case, points a and b are probably within the same region, but a and c are not.

a non-hierarchical model, the modeler sets the range of the parameters that define the hypotheses. In a hierarchical model, the modeler instead specifies *hyperparameters* – parameters defining the parameters – and the model learns the range of the parameters themselves. So rather than being given that the range of each of the x, y, l, and w values lies between 0 and 10^9, a hierarchical model might instead learn the typical range of each (e.g., that l tends to be short while w tends to be long, as depicted in Figure 3b) while the modeler specifies the range of the ranges. In principle, the level at which learning occurs could be extended upward even more, until the knowledge at the highest level is weak or general enough that it can be plausibly assumed to be innate.

We use a hierarchical Bayesian model (HBM), which supports the simultaneous acquisition of multiple levels of knowledge: both concrete and item-specific, as well as abstract and general. Several Bayesian models describing how categorization biases might emerge exist (Navarro, 2006; Kemp, Perfors, & Tenenbaum, 2007; Griffiths, Sanborn, Canini, & Navarro, 2008; Heller, Sanborn, & Chater, 2009). What all of these models have in common is that they can perform inference on multiple levels of abstraction. When presented with data consisting of specific objects and features, these models are capable of making generalizations about the specific objects as well as the appropriate generalizations about categorization in general. For instance, children in

an experiment by L. Smith et al. (2002) were presented with four novel concepts and labels and acquired a bias to assume not only that individual categories like chairs tend to be organized by shape, but also that solid categories *in general* are as well. A model can make the same generalization on the basis of the same data (Kemp et al., 2007).

A surprising effect of learning in hierarchical models is that, quite often, the higher-order abstractions are acquired before all of the specific lower-level details: just as children acquire some categorization biases even before they have learned all categories, the model might realize that l tends to be short and w tends to be long before the size and location of each rectangular region is learned with precision. This effect, which we might call the "blessing of abstraction", is somewhat counterintuitive. Why are higher-order generalizations like this sometimes easier for a Bayesian learner to acquire?

One reason is that the higher-level hypothesis space is often smaller than the lower-level one. As a result, the model has to choose between fewer options at the higher level, which may require less evidence. In our rectangle example, the higher-level knowledge may consist of only three options: l and w are approximately equal, l is smaller than w, or w is smaller than l. Even if a learner doesn't know whether l is 10 units or 11 units long and w is 20 or 22, it might be fairly obvious that l is smaller than w.

More generally, the higher-level inference concerns the lower-level hypothesis space (and is therefore based on the data set as a whole), whereas the lower-level inference is only relevant for specific datapoints. A single datapoint is informative only about the precise size and location of a single region. However, it – and every other single datapoint – is informative about all of the higher-level hypotheses. There is, in effect, more evidence available to the higher levels than the lower ones, and they are therefore learned more quickly.

This has interesting implications for the study of learnability and questions of innateness. The basic motivation for positing innate constraints on cognitive development is that without these constraints, children would be unable to infer the specific knowledge that they seem to acquire from the limited data available to them. What is critical

to the argument is that some constraints are present prior to learning some of the specific data, not that those constraints must be innate. Approaches to cognitive development that emphasize learning from data typically view the course of development as a progressive layering of increasingly abstract knowledge on top of more concrete representations; under such a view, learned abstract knowledge would tend to come in after more specific concrete knowledge is learned, so the former could not usefully constrain the latter. This view is sensible in the absence of learning mechanisms that can explain how abstract constraints could be learned together with (or before) the more specific knowledge they are needed to constrain. However, the hierarchical Bayesian framework provides such a learning mechanism. If an abstract generalization can be acquired very early and can function as a constraint on later acquisition of specific data, it may function effectively as if it were an innate domain-specific constraint, even if it is in fact not innate and instead is acquired by domain-general induction from data.

In sum, then, hierarchical Bayesian models offer a valuable tool for exploring questions of innateness due to the ability to limit built-in knowledge to increasingly abstract levels and thereby learn inductive constraints at other levels. As we will see in the next section, the Bayesian framework is also a useful way of approaching these questions for another reason – their ability to evaluate the rational tradeoff between the simplicity of a hypothesis and its goodness-of-fit to the evidence in the world. Because of this, Bayesian learners can make inferences that otherwise appear to go beyond the amount of evidence available.

4 Developing inductive frameworks

The hierarchical Bayesian models described above explain the origins of inductive biases and constraints by tuning priors in response to data observed from multiple settings or contexts. But the acquisition of abstract knowledge often appears more discrete or qualitative – more like constructing an appropriate hypothesis space, or selecting an appropriate hypothesis space from a higher level "hypothesis space of hypothesis spaces".

Consider the "theory theory" view of cognitive development. Children's knowledge about the world is organized into intuitive theories with a structure and function analogous to scientific theories (Carey, 1985; Gopnik & Meltzoff, 1997; Karmiloff-Smith, 1988; Keil, 1989). The theory serves as an abstract framework that guides inductive generalization at more concrete levels of knowledge, by generating a space of hypotheses. Intuitive theories have been posited to underlie real-world categorization (Murphy & Medin, 1985), causal induction (Waldmann, 1996; Griffiths & Tenenbaum, 2009), biological reasoning (Atran, 1995; Inagaki & Hatano, 2002; Medin & Atran, 1999), physical reasoning (McCloskey, 1983) and social interaction (Nichols & Stich, 2003; Wellman, 1990). For instance, an intuitive theory of mind generates hypotheses about how a specific agent's behavior might be explained in particular situations – candidate explanations framed in terms of mental states such as goals, beliefs, or preferences. Under this view, cognitive development requires recognizing that a current theory of a domain is inadequate, and revising it in favor of a new theory that makes qualitative conceptual distinctions not made in the earlier theory (Carey, 1985; Gopnik, 1996). Probabilistic models provide a way to understand how such a process of theory change might take place, and in particular how a learner might weigh the explanatory power of alternative theories against each other.

4.1 Trading off parsimony and goodness-of-fit

One of the most basic challenges in choosing between theories (or grammars, or other kinds of inductive frameworks) is trading off the parsimony, or simplicity, of a theory with how well it fits the observed data. To take a developmental example inspired by one of the papers that appears in this special issue (Lucas et al., submitted), we can imagine a child choosing between two theories of human choice behavior. Under one theory, everybody shares essentially the same assumptions about what kinds of things are desirable, such as having the same preferences for different kinds of food (and hence has the same preferences as the child). Under the other theory, different people can

possess different preferences. The developmental data suggest that a transition between these two theories occurs when children are between 14 and 18 months of age (Repacholi & Gopnik, 1997). However, the second theory is significantly more complex than the first, with the information required to specify the preferences of everybody the child knows increasing with the number of people. This extra complexity makes the theory more flexible, and thus better able to explain the pattern of choices a group of people might make. Even if it were the case that everybody shared the same preferences, any random variation in people's choices could be explained by the more complex theory in terms of different people having different preferences. So, how can the child know when a particular pattern of choices should lead to the adoption of this more complex theory?

Developing intuitive theories requires trading off parsimony with goodness-of-fit. A more complex theory will always fit the observed data better, and thus needs to be penalized for its additional flexibility. While our example focuses on the development of theories of preference, the same problem arises whenever theories, grammars or other inductive frameworks that differ in complexity need to be compared. Just as a higher-order polynomial is more complicated but can fit a data set more precisely, so too can a highly expressive theory or grammar, with more internal degrees of freedom, fit a body of data more exactly. How does a scientist or a child recognize when to stop positing ever more complex epicycles, and instead adopt a qualitatively different theoretical framework? Bayesian inference provides a general-purpose way to formalize a rational tradeoff between parsimony and fit.

As we saw earlier, goodness-of-fit for a hypothesis h is captured by the likelihood term in Bayes' Rule, or $p(d|h)$, while the prior $p(h)$ reflects other sources of a learner's beliefs. Priors can take various forms, but in general, a preference for simpler or more parsimonious hypotheses will emerge naturally without having to be engineered deliberately. This preference derives from the generative assumptions underlying the Bayesian framework, in which hypotheses are themselves generated by a stochastic process that

produces a space of candidate hypotheses and $p(h)$ reflects the probability of generating h under that process.

To illustrate, consider the three hypotheses shown in Figure 4. We expand on our previous example by now stipulating that individual hypotheses may include more than one rectangular subregion. As a result, hypotheses are generated by first choosing a number of rectangular subregions and then choosing l, w, x, and y for each subregion. The first choice of how many subregions could be biased towards smaller numbers, but it need not be. Simpler hypotheses, corresponding to those with fewer subregions, would still receive higher prior probability because they require fewer choice points in total to generate. The simplest hypothesis A, with one subregion, can be fully specified by making only four choices: l, w, x, and y. Hypothesis C, at the other extreme, contains twenty-one distinct rectangular subregions, and therefore requires 84 separate choices to specify, four for each subregion. Intuitively, the more complicated a pattern is, the more "accidental" it is likely to appear; the more choices a hypothesis requires, the more likely it is that those choices could have been made in a different way, resulting in an entirely different hypothesis. More formally, because the prior probability of a hypothesis is the product of the probabilities for all choices needed to generate it, and the probability of making any of these choices in a particular way must be less than one, a hypothesis specified by strictly more choices will in general receive strictly lower prior probability.

There are other ways of generating the hypotheses shown in Figure 4 – for instance, we could choose the upper-right and lower-left corners of each rectangular subregion, rather than choosing one corner, a height and a width. These might generate quantitatively different prior probabilities but would still give a qualitatively similar tradeoff between complexity and fit. The "Bayesian Ockham's razor" (MacKay, 2003) thus removes much of the subjectivity inherent in assessing simplicity of an explanation.[2]

[2]That said, it is always possible to imagine bizarre theories, generating hypotheses from very different primitives than we typically consider, in which hypotheses that are intuitively more complex receive higher (not lower) prior probabilities. For instance, suppose that the hypotheses shown in Figure 4 were generated not by choosing the dimensions of one or more rectangles from some generic

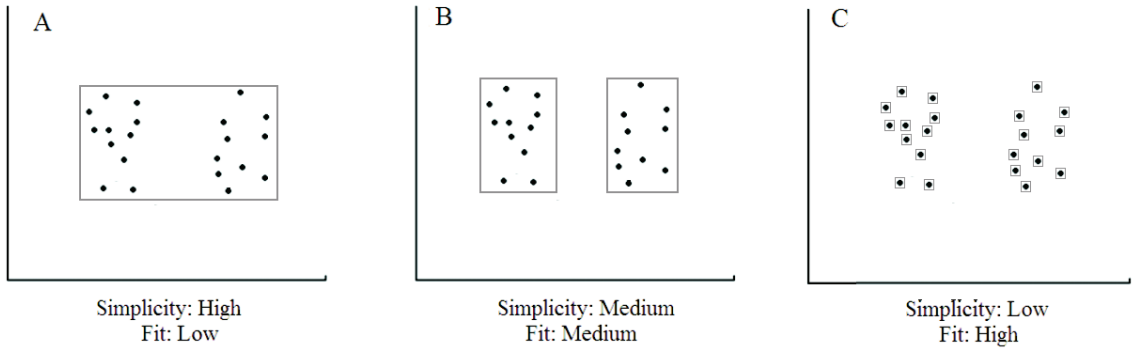

Figure 4: Hypothesis A is too simple, fitting the observed data poorly; C fits closely but is too complex; while B is "just right." A Bayesian analysis naturally ensures that the best explanation of the data optimizes a tradeoff between complexity and fit, as in B.

Note that in any of these generative accounts where hypotheses are generated by a sequence of choices, earlier or higher-up choices tend to play more important roles because they can affect the number and the nature of choices made later on or lower down. The initial choice of how many rectangular subregions to generate determines how many choices about positions and side lengths are made later on. Perhaps we could have also chosen initially to generate circular subregions instead of rectangles; then each subregion would involve only three choices rather than four. These higher-level "choices that control choices" characterize the learner's "hypothesis space of hypothesis spaces"; they embody a more discrete, qualitative version of the hierarchical Bayesian ideas introduced in the previous section. They capture the role that intuitive theories or grammars play in providing frameworks for inductive inference in cognition, or the analogous role that higher-level frameworks or paradigms play in scientific theory building (Henderson, Goodman, Tenenbaum, & Woodward, 2010).

The logic outlined in the preceding paragraphs has been used to analyze develop-

distribution, but by starting with just the twenty-one small rectangles in Figure 4C, and then making choices about whether to add or remove rectangles to or from this set. In that case, hypothesis C would almost certainly have higher prior probability than A or B. Because the simplicity of a hypothesis is only meaningful relative to the primitives out of which hypotheses are generated, the decision of which primitives to include in a probabilistic model of cognition is a crucial choice, which we consider in more depth later. For now, we simply note that this is a key concern for any cognitive modeler, Bayesian or otherwise inclined. It can be seen as a virtue of the Bayesian framework that it forces us to make these choices and their consequences explicit, and that it provides a tool to evaluate the primitives we choose.

mental theory transitions in several settings. Elsewhere in this issue, Lucas et al. (submitted) show that the change from believing everybody shares the same preferences (analogous to hypothesis A in Figure 4) to believing everybody has different preferences (analogous to hypothesis C in Figure 4) can be produced simply by providing more data, a mechanism that we discuss in more detail in the next section. Goodman et al. (2006) show that the same approach can be used to explain the development of understanding of false beliefs, with a theory in which the beliefs that people maintain are influenced by their access to information being more complex but providing a better fit to the data than a theory without this principle. Schmidt, Kemp, and Tenenbaum (2006) demonstrated that a high-level theory about the properties of semantic predicates known as the M-constraint (essentially the constraint that predicates respect the structure of an ontological hierarchy Sommers, 1971; Keil, 1979) can be induced from linguistic data consistent with that theory, providing an alternative to the idea that this constraint is innate. Work by Perfors, Tenenbaum, and Regier (2006); Perfors et al. (submitted) reanalyzes one version of a famous "poverty of stimulus" argument, and demonstrates that highly abstract and universal features of language – in particular, the principle that grammars incorporate hierarchical phrase structure – need not be built in as a language-specific bias but instead can be inferred on the basis of only a few hours of child-directed speech. This is because hierarchical grammars offer a more parsimonious explanation of the observed sentences: the grammars are shorter, with fewer non-terminals and fewer rules – that is, fewer choice points.

4.2 Adapting Ockham's Razor to the data

A key advantage of Bayesian approaches over earlier approaches to selecting grammars or theories based on data can be seen in how they adapt the preference for simpler hypotheses as the amount of observed data increases. In language acquisition, a traditional solution to the problem of constraining generalizing in the absence of negative evidence is the Subset Principle (Wexler & Culicover, 1980; Berwick, 1986): learners

should choose the most specific grammar consistent with the observed data. In scientific theorizing, the classical form of Ockham's Razor speaks similarly: entities should not be multiplied beyond necessity. The difficulty with these approaches is that because their inferential power is too weak, they require additional constraints in order to work – and those constraints often apply only in a way we can recognize *post hoc*. In Figure 4, for instance, the preference for hypothesis B over A can be explained by the Subset Principle, but to explain why B is better than C (a subset of B), we must posit that C is ruled out a priori by some innate constraint; it is just not a natural hypothesis and should never be learnable, regardless of the data observed.

A Bayesian version of Ockham's Razor, in contrast, will naturally modulate the tradeoff between simplicity and goodness-of-fit based on the available weight of data, even if the data are always generated by the same underlying process. This adaptiveness is intuitively sensible and critical for human learning. Consider Figure 5, which shows three data sets generated from the same underlying process but varying in the amount of data observed. The best hypothesis fits the five datapoints in data set 1 quite loosely, but because there are so few points this does not impose a substantial penalty relative to the high prior probability of the hypothesis. Analogously, early on in development children's categories, generalizations, and intuitive theories are likely to be more coarse than those of adults, blurring distinctions that adults consider highly relevant and therefore being more likely to over-generalize.[3] As data accumulate, the relative penalty

[3]Adopting a sequence of ever more complex theories as the relevant data come to light seems like a plausible account of cognitive development, but it appears to be at odds with the familiar phenomenon of U-shaped learning curves (e.g., Marcus et al. (1992); see also Siegler (2004) for an overview). A U-shaped learning pattern occurs when a learner initially appears to have correctly acquired some piece of knowledge, producing it without error, but then follows this by an interval of incorrect performance marked by overgeneralization before eventually self-correcting. It may be possible to understand U-shaped acquisition patterns by considering a learner who can simply memorize individual datapoints in addition to choosing among hypotheses about them. In our example, memorizing a datapoint would require two choices to specify – its x and y coordinates – but even the simplest hypothesis would require at least four (x, y, l, and w). Moreover, a single datapoint also has the highest possible likelihood, since it predicts the data (itself) exactly. A data set with only one or a few datapoints, therefore, would be preferred in both the prior *and* the likelihood. Only as the number of datapoints increases would the penalty in the prior become high enough to preclude simply memorizing each datapoint individually: this is when overgeneral, highly simple hypotheses begin to be preferred. Thus, whether a U-shaped pattern occurs depends on the tradeoff in complexity that it takes to represent individual datapoints as opposed to entire hypotheses: if it is cheaper to memorize a few datapoints, then that would have

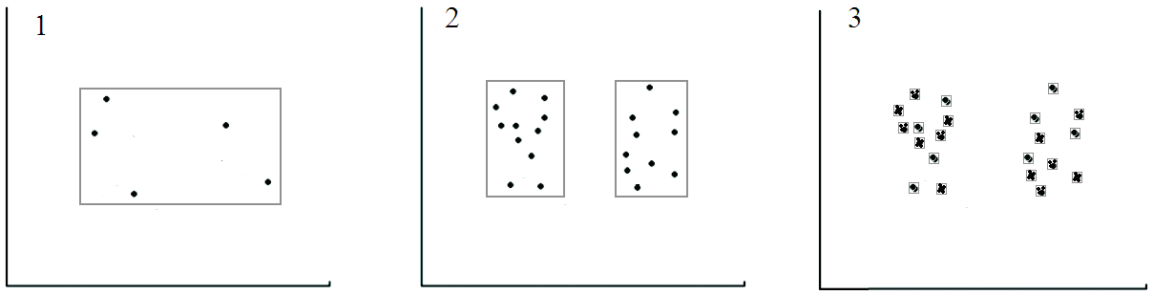

Figure 5: Role of data set size. Three datasets with increasing numbers of datapoints and their corresponding best hypotheses. For data set 1, there are so few datapoints that the simplicity of the hypothesis is the primary consideration; by dat aset 3, the preferred hypothesis is one that fits the clustered datapoints quite tightly.

imposed for poor fit is greater, since it applies to each datapoint that is not predicted accurately by the hypothesis. More complex hypotheses become more plausible, and even hypothesis C that looked absurd on the data in Figure 4 could become compelling given a large enough data set, containing many datapoints all clustered into the twenty-one tiny regions exactly as the theory predicts. The Subset Principle is not flexible in the same way. Being able to explain the process of development, with different theories being adopted by the child at different stages, requires being able to adapt to the data. This property makes it possible for the gradual accumulation of data to be the driving force in theory change, as in the examples discussed above.

Looking at Figure 5, one might guess that as the data increase, the most complex hypotheses will eventually always be preferred. This is not true in general, although as data accumulate a Bayesian learner will tend to consider more complex hypotheses. Yet the preferred hypothesis will be that which best compresses the data, and ultimately, this will be the hypothesis that is closest to the true generative process (MacKay, 2003).[4] In other words, if the data are truly generated by a process corresponding to

both a higher prior and likelihood than would an extremely vague, overly general hypothesis.

[4]Technically, this result has been proven for information-theoretic models in which probabilities of data or hypotheses are replaced by the lengths (in bits) of messages that communicate them to a receiver. The result is known as the "MDL Principle" (Rissanen, 1978), and is related to Kolmogorov complexity (Solomonoff, 1964; Kolmogorov, 1965). The Bayesian version applies given certain assumptions about the randomness of the data relative to the hypotheses and the hypotheses relative to the prior (Vitànyi & Li, 2000). Both versions apply only to the hypotheses in the hypothesis space: if no hypothesis corresponding to the true data generating process exists in the space, then it will never be considered, much less ultimately preferred. Thus, the hypothesis that is preferred by the model in the limit of infinite data is the "best" hypothesis only in the sense that it is closest to the true data

twenty-one different rectangular regions, then the points will increasingly clump into clusters in those regions, and hypothesis C will eventually be preferred. But if the data are truly generated by a process inhabiting two larger regions, then hypothesis B would still have a higher posterior probability as more data accumulate.

5 Some general issues

Several issues are typically raised when evaluating Bayesian modelling as a serious computational tool for cognitive science. Bayesian reasoning characterizes "optimal" inference: what does this mean? How biologically plausible are these models, and how much does this matter? And finally, where does it all come from – the hypothesis space, the parameters, the representations? The answers to each of these questions affect what conclusions about actual human cognition we can draw on the basis of Bayesian models; we therefore consider each in turn.

5.1 Optimality: What does it mean?

Bayesian probability theory[5] is not simply a set of *ad hoc* rules useful for manipulating and evaluating statistical information: it is also the set of unique, consistent rules for conducting plausible inference (Jaynes, 2003). In essence, it is a extension of deductive logic to the case where propositions have degrees of truth or falsity – that is, it is identical to deductive logic if we know all the propositions with 100% certainty. Just as formal logic describes a deductively correct way of thinking, Bayesian probability theory describes an inductively correct way of thinking. As Laplace (1816) said, "probability theory is nothing but common sense reduced to calculation."

generating process out of all of the hypotheses considered.

[5]Bayesian methods are often contrasted to so-called "frequentist" approaches, which are the basis for many of the standard statistical tests used in the social sciences, such as t-tests. Although frequentist methods often correspond to special cases of Bayesian probability theory, Bayesian methods have historically been relatively neglected, and often attacked, in part because they are viewed as unnecessarily subjective. This perception is untrue – Bayesian methods are simply more explicit about the prior information they take into account. Regardless, the issue of subjectivity seems particularly irrelevant for those interested in modelling human cognition, where accurately capturing "subjective belief" is part of the point.

What does this mean? If we were to try to come up with a set of desiderata that a system of "proper reasoning" should meet, they might include things like consistency and qualitative correspondence with common sense – if you see some data supporting a new proposition A, you should conclude that A is more plausible rather than less; the more you think A is true, the less you should think it is false; if a conclusion can be reasoned multiple ways, its probability should be the same regardless of how you got there; etc. The basic axioms and theorems of probability theory, including Bayes' Rule, emerge when these desiderata are formalized mathematically (Cox, 1946, 1961), and correspond to common-sense reasoning and the scientific method. Put another way, Bayesian probability theory is "optimal inference" 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, 1937).

Even if the Bayesian framework captures optimal inductive inference, does that mean it is an appropriate tool for modelling human cognition? After all, people's everyday reasoning can be said to be many things, but few would aver that it is always optimal, subject as it is to emotions, heuristics, and biases of many different sorts (e.g., Tversky & Kahneman, 1974). However, even if humans are non-optimal thinkers in many ways – and there is no reason to think they are in *every* way – it is impossible to know this without being able to precisely specify what optimal thinking would amount to. Understanding how humans *do* think is often made easier if one can identify the ways in which people depart from the ideal: this is approximately the methodology by which Kahneman and Tversky derived many of their famous heuristics and biases, and the flexibility of the Bayesian approach makes it relatively easy to incorporate constraints based on memory, attention, or perception directly into one's model.

Bayesian modelling, in fact, is an implementation of scientific inquiry that operates on Marr's third (computational) level, which seeks to understand cognition based on what its goal is, why that goal would be appropriate, and the constraints on achieving that goal, rather than precisely how it is implemented algorithmically (Marr, 1982).

Understanding at this level is important because the nature of the reasoning may often depend more on the learner's goals and constraints than it does on the particular implementation. It can also enhance understanding at the other levels: for instance, analyzing connectionist networks as an implementation of a computational-level theory can elucidate what sort of computations they perform, and often explain why they produce the results they do (Hertz, Krogh, & Palmer, 1991; MacKay, 2003).

Being able to precisely specify and understand optimal reasoning is also useful for performing ideal learnability analysis, which especially important in the area of cognitive development. What must be "built into" the newborn mind in order to explain how infants eventually grow to be adult reasoners, with adult knowledge? One way to address this question is to establish the bounds of the possible: if some knowledge couldn't possibly be learned by an optimal learner presented with the type of data children receive, it is probably safe to conclude either that actual children couldn't learn it, either, or that some of the assumptions underlying the model are inaccurate. The tools of Bayesian inference are well-matched to this sort of problem, both because they force modelers to make all of these assumptions explicit, and also because of their representational flexibility and ability to calculate optimal inference.

5.2 Biological plausibility

Because cognitive scientists are ultimately interested in understanding human cognition, and human cognition is ultimately implemented in the brain, it is important that our computational-level explanations be realizable on the neurological level, at least potentially. This is one reason for the popularity of the Parallel Distributed Processing, or connectionist, approach, which was developed as a neurally inspired model of the cognitive process (Rumelhart & McClelland, 1986). Connectionist networks, like the brain, contain many highly interconnected, active processing units (like neurons) that communicate with each other by sending activation or inhibition through their connections. As in the brain, learning appears to involve modifying connections, and

knowledge is represented in a distributed fashion over the connections. As a result, representations degrade gracefully with neural damage, and reasoning is probabilistic and "fuzzy" rather than all-or-none.

In contrast, Bayesian models may appear implausible from the neurological perspective. One of the major virtues of Bayesian inference – the transparency of its computations and the explicitness of its representation – is, in this light, potentially a major flaw: the brain is many wonderful things, but it is neither transparent nor explicit. How could representations like grammars or logics be instantiated in our neural hardware? How could our cortex encode hypotheses and compare them based on a tradeoff between their simplicity and goodness-of-fit? Perhaps most problematically, how could the brain – with a processing speed that is orders of magnitude slower than that of modern computers – actually implement optimal inference, which requires a search of such enormous hypothesis spaces that even the fastest computers can take days or weeks, if they succeed at all?

These are good questions, but there is growing evidence for the relevance of Bayesian approaches on the neural level (e.g., Doya, Ishii, Pouget, & Rao, 2007). Probability distributions can in fact be represented by neurons, and they can be combined according to a close approximation of Bayes' Rule; posterior probability distributions may be encoded in populations of neurons in such a way that Bayesian inference is achieved simply by summing up firing rates (Pouget, Dayan, & Zemel, 2003; Ma, Beck, Latham, & Pouget, 2006). Spiking neurons can be modelled as Bayesian integrators accumulating evidence over time (Deneve, 2004; Zemel, Huys, Natarajan, & Dayan, 2005). Recurrent neural circuits are capable of performing both hierarchical and sequential Bayesian inference (Deneve, 2004; Rao, 2004, 2007). Even specific brain areas have been studied: for instance, there is evidence that the recurrent loops in the visual cortex integrate top-down priors and bottom-up data in such a way as to implement hierarchical Bayesian inference (T. Lee & Mumford, 2003).

This work, though still in its infancy, suggests that concerns about biological plau-

sibility may not, in the end, prove to be particularly problematic. It may seem to us, used to working with serial computers, that searching these enormous hypothesis spaces quickly enough to perform anything approximating Bayesian inference is impossible; but the brain is a parallel computing machine made up of billions of highly interconnected neurons. The sorts of calculations that take a long time on a serial computer, like a sequential search of a hypothesis space, might be very easily performed in parallel. They also might not; but whatever the future holds, the indications so far serve as a reminder of the danger of advancing from the "argument from incredulity" to any conclusions about biological plausibility.

It is also important to note that, for all of their apparent biological plausibility, neural networks are unrealistic in important ways, as many modelers acknowledge. Units in neural networks are assumed to have both excitatory and inhibitory connections, which is not neurally plausible. This is a problem because the primary learning mechanism, backpropagation, relies on the existence of such connections (Rumelhart & McClelland, 1986; Hertz et al., 1991). There is also no analogue of neurotransmitters and other chemical transmission, which play an important role in brain processes (Gazzaniga, Ivry, & Mangun, 2002). These issues are being overcome as the state of the art advances (see Rao, Olshausen, and Lewicki (2002) for some examples), but for the models most commonly used in cognitive science – perceptrons, multilayered recurrent networks, and Boltzmann machines – they remain a relevant concern.

Different techniques are therefore biologically plausible in some ways and perhaps less so in others. Knowing so little about the neurological mechanisms within the brain, it is difficult to characterize how plausible either approach is or how much the ways they fall short impact their utility. In addition, biological plausibility is somewhat irrelevant on the computational level of analysis. Even if it turned out that there was no possible way the brain could implement anything even heuristically approximating Bayesian inference – which seems unlikely in light of current research – these models would still be useful for comprehending the goals and constraints faced by the cognitive

system and comparing actual human performance to optimal reasoning. To the extent that neural networks are relevant to the computational level, the same is true for them.

5.3 Where does it all come from?

For many, a more important critique is that, 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, has been given by the modeler; all the model does is choose among hypotheses that already exist. Isn't learning, particularly the sort of learning that children perform over the first years of their life, something more than this? Our intuitive notion of learning certainly encompasses a spirit of discovery that does not appear at first glance to be captured by a model that simply does hypothesis testing within an already fully-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 learn anything via hypothesis testing because one must possess it in order to test it in the first place. Therefore, except for those concepts that can be created by composing them from others, all concepts (including CARBURETOR and GRANDMOTHER) must be innate.

To understand how this intuition can be misleading, it is helpful to make a distinction between two separate notions of what it means to build in a hypothesis space. A trivial sense is to equip the model with the representational capacity to represent any of the hypotheses in the space: if a model has this capacity, even if it is not currently evaluating or considering any given hypothesis, that hypothesis is in some sense latent in that space. Thus, if people have the capacity to represent some given hypothesis, we say it can be found in their *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. We can contrast this with hypotheses that may be considered or evaluated: the hypotheses that can be manipulated by the conceptual system, which we refer to as the *explicit hypothesis space*.

As an analogy, consider a standard English typewriter with an infinite amount of paper. There is a space of documents that it is capable of producing, which includes things like *The Tempest* and does not include, say, a Vermeer painting or a poem written in Russian. This typewriter can easily be formalized in Bayesian terms where each possible document is a hypothesis, the infinite set of documents producible by the typewriter is its latent hypothesis space[6], and the documents actually produced 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 have been considered 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. Hypothesis generation would describe 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. Hypothesis testing would describe the process of deciding which of the documents produced should be preferred (by whatever standard). Learning, then, would correspond to the entire process of hypothesis generation and testing – and hence would never involve new hypotheses being a! dded to the latent hypothesis space. This is what some object to: it doesn't "feel" like learning, since in some sense everything is already "built in."

However, this intuitive feeling is misleading. If we take "learning" to mean "learning in the Fodorian sense" or, equivalently, "not built into the latent hypothesis space", then there are only two conclusions possible. Either the hypotheses appear in the latent hypothesis space completely randomly, or *nothing* can ever be learned. In other words, there is no interpretation of "learning in the Fodorian sense" that allows for an

[6]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. Indeed, 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.

interesting model or theory of learning to emerge.

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. 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 original hypotheses were. In general, the only way for something to be learned in the Fodorian sense – the sense that underlies this critique – is for them 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 is that *every* model, even the brain, must come equipped with a latent hypothesis space that consists of everything that it can possibly represent and compute; all learning must happen within this space. This is not a novel or controversial point – all cognitive scientists accept that *something* must be built in – but it is often forgotten; 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 rule, etc. In fact, neural networks can be viewed as implementations of Bayesian inference (e.g., Funahashi, 1998; McClelland, 1998; MacKay, 2003), corresponding to a computational-level model whose hypothesis space is a set of continuous functions (e.g., Funahashi, 1989; Stinchcombe & White, 1989). This is a large space, but Bayesian inference can entertain hypothesis spaces that are equivalently large.

Does this mean that there is no difference between Bayesian models and neural networks? In one way, the answer is yes: because neural networks are universal approximators, it is always possible to construct one that is an implementation of a Bayesian model. In practice, however, the answer is usually no: the two methods have very different strengths and weaknesses, and therefore their value as modelling tools varies depending on the questions being asked.[7] Bayesian models optimally trade off between simplicity and goodness-of-fit; neural network models perform a similar tradeoff, but generally non-optimally and in a more *ad hoc* manner, avoiding overfitting by limiting the length of training and choosing appropriate weights, learning rules, and network architecture.[8] In the Bayesian framework, what is built in is the generative process, which implicitly defines the assignment of prior probabilities, the representation, and the size of the hypothesis space; in the PDP framework, these things are built in through choices about the architecture, weights, learning rule, training procedure, etc.

It is therefore incorrect to say one framework assumes more innate knowledge than another: *specific models* within each may assume more or less, but it can be quite difficult to compare them precisely, in part because neural networks incorporate it implicitly. Which model assumes more innate knowledge is often not even the interesting question. A more appropriate one might be: *what* innate knowledge does it assume? Instead of asking whether one representation is a stronger assumption than another, it is often more productive to ask which predicts human behavior better. The answer will probably depend on the problem and the domain, but the great advantage of computational modelling is that it allows us to explore this dependence precisely.

[7] Many PDP modelers see the two approaches as complementary (e.g., Rogers & McClelland, 2004), and we agree with this view.

[8] There is an interesting subfield called Bayesian neural networks studying how to construct models that make these choices for themselves, pruning connections in a Bayes-optimal way (e.g., MacKay, 1995; Neal, 1994, 1996).

6 Conclusion

Bayesian inference is a useful approach in cognitive science, especially when used in conjunction with other types of computational modelling and experimental work. Its representational flexibility makes the Bayesian approach applicable to a wide variety of learning problems, and its transparency makes it easy to be clear about what assumptions are being made and what is being learned. The framework is valuable for defining an optimal standard as well as for exploring and illustrating the tradeoff between simplicity and goodness-of-fit. As a result, it has the potential to explain many aspects of human cognition and development.

References

Alishahi, A., & Stevenson, S. (2008). A computational model for early argument structure acquisition. *Cognitive Science*, *32*.

Andrews, M., & Vigliocco, G. (2009). Learning semantic representations with hidden markov topics models. In *31st Annual Conference of the Cognitive Science Society.*

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*(3), 463-498.

Atran, S. (1995). Classifying nature across cultures. In E. E. Smith & D. N. Osherson (Eds.), *Thinking: An invitation to cognitive science* (Vol. 3, p. 131-174). Cambridge, MA: MIT Press.

Baker, C., Tenenbaum, J. B., & Saxe, R. (2007). Goal inference as inverse planning. *Proceedings of the 29th Annual Conference of the Cognitive Science Society.*

Berger, J. (1993). *Statistical decision theory and Bayesian analysis.* New York: Springer.

Berwick, R. (1986). Learning from positive-only examples: The subset principle and three case studies. *Machine Learning*, *2*, 625–645.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Box, G., & Tiao, G. (1992). *Bayesian inference in statistical analysis*. New York: Wiley.

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 264–293). Cambridge, MA: MIT Press.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Carey, S. (2009). *Origin of concepts*. Oxford University Press.

Carey, S., & Spelke, E. (1996). Science and core knowledge. *Philosophy of Science*, *63*, 515–533.

Cox, R. (1946). Probability, frequency, and reasonable expectation. *American Journal of Physics*, *14*, 1–13.

Cox, R. (1961). *The algebra of productive inference*. Baltimore, MD: Johns Hopkins University Press.

de Finetti, B. (1937). Prevision, its logical laws, its subjective sources. In H. Kyburg & H. Smokler (Eds.), *In studies in subjective probability* (2nd ed.). New York: J. Wiley and Sons.

de Finetti, B. (1974). *Theory of probability* (2nd ed.). New York: J. Wiley and Sons.

Deneve, S. (2004). Bayesian inference in spiking neurons. *Advances in Neural Information Processing Systems*, *17*.

Doucet, A., Freitas, N. d., & Gordon, N. (2001). *Sequential monte carlo in practice*. Springer-Verlag.

Dowman, M. (2000). Addressing the learnability of verb subcategorizations with Bayesian inference. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*.

Doya, K., Ishii, S., Pouget, A., & Rao, R. (2007). *Bayesian brain: Probabilistic approaches to neural coding*. Cambridge, MA: MIT Press.

Feldman, N., & Griffiths, T. L. (2009). Learning phonetic categories by learning a lexicon. In *31st Annual Conference of the Cognitive Science Society.*

Feldman, N., Morgan, J., & Griffiths, T. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752–782.

Fisher, R. (1933). Probability, likelihood, and quantity of information in the logic of uncertain inference. *Proceedings of the Royal Society*, *146*, 1–8.

Fodor, J. (1975). *The language of thought*. New York, NY: Thomas Y. Crowell Company.

Fodor, J. (1981). *Representations: philosophical essays on the foundations of cognitive science*. Cambridge, MA: MIT Press.

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

Frank, M., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2007). Modeling human performance in statistical word segmentation. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society.*

Frank, M., Goodman, N., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578-585.

Frank, M., Ichinco, D., & Tenenbaum, J. B. (2008). Principles of generalization for learning sequential structure in language. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society.*

Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, *2*, 183–192.

Funahashi, K. (1998). Multilayer neural networks and bayes decision theory. *Neural Networks*, *11*(2), 209–213.

Gazzaniga, M., Ivry, R., & Mangun, G. (2002). *Cognitive neuroscience: The biology of the mind* (2nd ed.). New York, NY: W.W. Norton & Company.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis* (2nd ed.). Chapman & Hall.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the naive theory of rational action. *Trends in Cognitive Science*, *7*(7), 287–292.

Gilks, W., Richardson, S., & Spiegelhalter, D. (1996). *Markov chain monte carlo in practice*. Chapman & Hall.

Goldwater, S., Griffiths, T., & Johnson, M. (2006). Interpolating between types and tokens by estimating power law generators. *Advances in Neural Information Processing Systems*, *18*.

Goldwater, S., & Griffiths, T. L. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. *Proceedings of the 45th Annual Conference of the Association for Computational Linguistics*.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2007). Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21-54.

Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

Goodman, N., Griffiths, T. L., Feldman, J., & Tenenbaum, J. B. (2007). A rational analysis of rule-based concept learning. *Proceedings of the 29th Annual Conference of the Cognitive Science Society*.

Goodman, N. D., Baker, C. L., Bonawitz, E. B., Manisinghka, V. K., Gopnik, A., Wellman, H., et al. (2006). Intuitive theories of mind: A rational approach to false belief. In *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Gopnik, A. (1996). The scientist as child. *Philosophy of Science*, *63*, 485-514.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological*

Review, *111*(1), 1–30.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

Griffiths, T. L., & Kalish, M. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*, 441-480.

Griffiths, T. L., Sanborn, A., Canini, K., & Navarro, D. (2008). Categorization as nonparameteric bayesian density estimation. In M. Oaksford & N. Chater (Eds.), *The probabilistic midn: Prospects for bayesian cognitive science*. Oxford: Oxford University Press.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*, 661-716.

Grünwald, P., Myung, J., & Pitt, M. (2005). *Advances in minimum description length: Theory and applications*. Cambridge, MA: MIT Press.

Heller, K., Sanborn, A., & Chater, N. (2009). Hierarchical learning of dimensional biases in human categorization. In *Advances in neural information processing systems* (Vol. 22). Cambridge, MA: MIT Press.

Henderson, L., Goodman, N. D., Tenenbaum, J. B., & Woodward, J. F. (2010). The structure and dynamics of scientific theories: A hierarchical Bayesian perspective. *Philosophy of Science*, *77*(2).

Hertz, J., Krogh, A., & Palmer, R. (1991). *Introduction to the theory of neural computation: Santa Fe institute studies in the science of complexity* (Vol. 1). Reading, MA: Perseus Books.

Hsu, A. (2009). Differential use of implicit negative evidence in generative and discriminative language learning. *Advances in Neural Information Processing Systems*, *22*.

Inagaki, K., & Hatano, G. (2002). *Young children's thinking about biological world*.

New York: Psychology press.

Jaynes, E. (2003). *Probability theory: The logic of science*. Cambridge: Cambridge University Press.

Jeffreys, H. (1939). *Theory of probability*. Oxford: Clarendon Press.

Karmiloff-Smith, A. (1988). The child is a theoretician, not an inductivist. *Mind and Language*, *3*, 183-195.

Keil, F. C. (1979). *Semantic and conceptual development: An ontological perspective*. Cambridge, MA: Harvard University Press.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2004). Learning domain structures. *Proceedings of the 26th Annual Conference of the Cognitive Science Society*.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Kemp, C., Tenenbaum, J. B., Niyogi, S., & Griffiths, T. L. (2010). A probabilistic model of theory formation. *Cognition*, *114*(2), 165-196.

Kolmogorov, A. (1965). Three approaches to the quantitative definition of information. *Problems of Information and Transmission*, *1*(1), 1–7.

Landau, B., Smith, L., & Jones, S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*, 299–321.

Laplace, P. S. (1816). *A philosophical essay on probabilities*. Dover Publications.

Lee, M. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science*, *30*(3), 555-580.

Lee, M., Fuss, I., & Navarro, D. (2006). A Bayesian approach to diffusion models of decision-making and response time. In *Advances in neural information processing systems.*

Lee, P. (1997). *Bayesian statistics*. New York: Wiley.

Lee, T., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex.

Journal of the Optical Society of America A, *20*, 1434–1448.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., et al. (submitted). The child as econometrician: A rational model of preference understanding in children. *Cognition*.

Ma, W., Beck, J., Latham, P., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438.

MacKay, D. (1995). Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, *6*, 469–505.

MacKay, D. (1998). Introduction to Monte Carlo methods. In M. Jordan (Ed.), *Learning in graphical models* (pp. 175–204). Cambridge, MA: MIT Press.

MacKay, D. (2003). *Information theory, inference, and learning algorithms*. Cambridge, UK: Cambridge University Press.

Marcus, G., Pinker, S., Ullman, M., Hollander, M., Rosen, T., Xu, F., et al. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, *57*, 1–178.

Markman, E. (1989). *Categorization and naming in children*. MIT Press.

Markman, E. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*, 57–77.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt & Company.

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

McCloskey, M. (1983). Intuitive physics. *Scientific American*, *284*, 114-123.

Medin, D. L., & Atran, S. (Eds.). (1999). *Folkbiology*. Cambridge, MA: Bradford Books.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence.

Psychological Review, *92*, 289-316.

Navarro, D. (2006). From natural kinds to complex categories. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the cognitive science society* (p. 621-626).

Navarro, D., & Griffiths, T. L. (2007). A nonparametric Bayesian method for inferring features from similarity judgments. *Advances in Neural Information Processing Systems*, *19*.

Navarro, D., Griffiths, T. L., Steyvers, M., & Lee, M. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*, 101-122.

Neal, R. (1993). *Probabilistic inference using Markov chain Monte Carlo methods* (Tech. Rep. No. CRG-TR-93-1). University of Toronto.

Neal, R. (1994). *Priors for infinite networks* (Tech. Rep. No. CRG-TR-94-1). University of Toronto.

Neal, R. (1996). *Bayesian learning for neural networks*. Springer.

Nichols, S., & Stich, S. (2003). *Mindreading: An integrated account of pretense, self-awareness and understanding other minds*. Oxford: Oxford University Press.

Nosofsky, J. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39-57.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Perfors, A., & Tenenbaum, J. B. (2009). Learning to learn categories. In N. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), *Proceedings of the 31st annual conference of the cognitive science society* (p. 136-141).

Perfors, A., Tenenbaum, J. B., & Regier, T. (2006). Poverty of the stimulus? A rational approach. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.

Perfors, A., Tenenbaum, J. B., & Regier, T. (submitted). The learnability of abstract

syntactic principles. *Cognition*.

Perfors, A., Tenenbaum, J. B., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Pouget, A., Dayan, P., & Zemel, R. (2003). Inference and computation with population codes. *Annual Reviews in Neuroscience*, *26*, 381–410.

Quine, W. (1960). *Word and object*. Cambridge, MA: MIT Press.

Rao, R. (2004). Bayesian computation in recurrent neural circuits. *Neural Computation*, *16*, 1–38.

Rao, R. (2007). Neural models of Bayesian belief propagation. In K. Doya, S. Ishii, A. Pouget, & R. Rao (Eds.), *Bayesian brain: Probabilistic approaches to neural coding* (pp. 239–267). Cambridge, MA: MIT Press.

Rao, R., Olshausen, B., & Lewicki, M. (2002). Probabilistic models of the brain: Perception and neural function.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology*, *33*(1), 12-21.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471.

Rogers, T., & McClelland, J. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Rosch, E. (1978). Principles of categorization. In E. Rosch & L. B. (Eds.), *Cognition and categorization* (p. 27-48). Lawrence Erlbaum.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Roy, D., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2006). Learning annotated hierarchies from relational data. *Advances in Neural Information Processing Systems*, *19*.

Rumelhart, D., & McClelland, J. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1). Cambridge, MA: MIT Press.

Savage, L. (1954). *Foundations of statistics*. New York, NY: J. Wiley & Sons.

Schmidt, L., Kemp, C., & Tenenbaum, J. B. (2006). Nonsense and sensibility: Inferring unseen possibilities. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*.

Shafto, P., Kemp, C., Mansinghka, V., Gordon, M., & Tenenbaum, J. B. (2006). Learning cross-cutting systems of categories. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.

Shultz, T., & Vogel, A. (2004). A connectionist model of the development of transitivity. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th annual conference of the cognitive science society* (p. 1243-1248).

Siegler, R. (2004). U-shaped interest in U-shaped development – and what it means. *Journal of Cognition and Development*, *5*(1), 1–10.

Sivia, D. (1996). *Data analysis: A Bayesian tutorial*. Oxford: Oxford University Press.

Smith, K. (2009). Iterated learning in populations of bayesian agents. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society.*

Smith, L., Jones, S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.

Solomonoff, R. (1964). A formal theory of inductive inference, parts 1 and 2. *Information and Control*, *7*(1–22), 224-254.

Sommers, F. (1971). Structural ontology. *Philosophia*, *1*, 21-42.

Spelke, E., & Kinzler, K. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96.

Spelke, E., Phillips, A., & Woodward, A. (1995). Infants' knowledge of object motion and human action. In *Causal cognition: A multidisciplinary debate* (pp. 44–78). Oxford: Oxford University Press.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*, 453–489.

Stinchcombe, M., & White, H. (1989). Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions. *Proceedings of the International Joint Conference on Neural Networks*, 607–611.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–641.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *135*, 1124–1131.

van Dantzig, D. (1957). Statistical priesthood (Savage on personal probabilities). *Statistica Neerlandica*, *2*, 1–16.

Verma, D., & Rao, R. (2006). Goal-based imitation as probabilistic inference over graphical models. *Advances in Neural Information Processing Systems*, *18*.

Vitànyi, P., & Li, M. (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory*, *46*(2), 446–464.

Waldmann, M. R. (1996). Knowledge-based causal induction. In *The psychology of learning and motivation* (Vol. 34, p. 47-88). San Diego: Academic Press.

Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

Wexler, K., & Culicover, P. (1980). *Formal principles of language acquisition*. Cambridge, MA: MIT Press.

Woodward, A. (1998). Infants' ability to distinguish between purposeful and non-purposeful behaviors. *Infant Behavior and Development*, *22*(2), 145–160.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*. (in press)

Zemel, R., Huys, Q., Natarajan, R., & Dayan, P. (2005). Probabilistic computation in spiking populations. *Advances in Neural Information Processing Systems*, *17*.

A Appendix

This appendix contains additional references that may be useful to those interested in learning more about different aspects of Bayesian learning.

A.1 Glossary

This is a brief glossary of some of the terms that may be encountered when learning about Bayesian models.

Bayesian Ockham's Razor : Describes how a preference for "simpler" models emerges in a Bayesian framework.

Blessing of abstraction : The phenomenon whereby higher-level, more abstract knowledge, may be easier or faster to acquire than specific, lower-level knowledge. This term was coined by Noah Goodman.

Conditional distribution : The probability of one variable (e.g., a) given another (e.g., b), denoted $p(a|b)$.

Graphical model : A probabilistic model for which a graph denotes the conditional independence structure between random variables. A directed graphical model identifies which of the nodes are the parents, and thus enables the joint distribution to be factored into conditional distributions. A directed graphical model is also known as a Bayesian network.

Hierarchical Bayesian model (HBM) : A type of Bayesian model capable of learning at multiple levels of abstraction.

Hyperparameters : The higher-level parameters learned in a hierarchical Bayesian model. These parameters capture the *overhypothesis* knowledge and govern the choice of lower-level parameters.

Hypothesis space : The set of all hypotheses a learner could entertain. This is divided into the *latent hypothesis space*, which consists of all logically possible hypothesis spaces and is defined by the structure of the learning problem, and the *explicit hypothesis space*, which contains the hypotheses a learner has explicitly considered or enumerated.

Joint distribution : The probability of multiple variables (e.g., a and b) occurring jointly, denoted $p(a, b)$.

Likelihood : The probability of having observed some data d if some hypothesis h is correct, denoted $p(d|h)$.

Markov chain Monte Carlo (MCMC) : A class of algorithms for sampling probability distributions. It is generally used when the probability distributions are too complex to be calculated analytically, and involves a series of sampling steps. Metropolis-Hastings and Gibbs sampling are two common types of MCMC methods.

Markov model : A model which captures a discrete random process in which the current state of the system depends only on the previous state of the system, rather than on states before that.

Overhypothesis : A higher-level inductive constraint that guides second-order generalization (or above). The term originates from N. Goodman (1955).

Posterior probability : The degree of belief assigned to some hypothesis h after having seen some data d (combines the *likelihood* and the *prior*, denoted $p(h|d)$.

Prior probability : The degree of belief assigned to some hypothesis h before having seen any data, denoted $p(h)$.

Probability distribution : Defines either the probability of a random variable (if the variable is discrete) or the probability of the value of the variable falling in a particular interval (when the variable is continuous).

Size principle : The preference for smaller hypotheses over larger ones, all else being equal, naturally instantiated by the likelihood term.

Stochastic : Random.

A.2 Applications

Recent years have seen a surge of interest in applying Bayesian techniques to many different problems in cognitive science. Although an exhaustive overview of this research is beyond the scope of this paper, we list here some example references, loosely organized by topic, intended to give the interested reader a place to begin, and also to illustrate the flexibility and scope of this framework. In addition, *Trends in Cognitive Sciences* (2007) published a special issue (Volume 10, Issue 7) focused on probabilistic models in cognition.

1. Learning phonetic categories: Feldman, Morgan, and Griffiths (2009); Feldman and Griffiths (2009)

2. Acquisition and nature of causal reasoning: Pearl (2000); Steyvers, Tenenbaum, Wagenmakers, and Blum (2003); Gopnik et al. (2004); Griffiths and Tenenbaum (2009)

3. Abstract reasoning and representation based on graphical structures, including taxonomies: Kemp, Perfors, and Tenenbaum (2004); Roy, Kemp, Mansinghka, and Tenenbaum (2006); Schmidt et al. (2006); Xu and Tenenbaum (2007)

4. Abstract semantic representations: Navarro and Griffiths (2007); Griffiths, Steyvers, and Tenenbaum (2007); Andrews and Vigliocco (2009)

5. Category learning and categorization: Navarro (2006); Kemp et al. (2007); Shafto, Kemp, Mansinghka, Gordon, and Tenenbaum (2006); Griffiths et al. (2008); Perfors and Tenenbaum (2009); Heller et al. (2009)

6. Decision making: M. Lee (2006); M. Lee, Fuss, and Navarro (2006)

7. Grammar learning and representation: Dowman (2000); Perfors et al. (2006, submitted)

8. Individual differences: Navarro, Griffiths, Steyvers, and Lee (2006)

9. Language evolution: Griffiths and Kalish (2007); K. Smith (2009)

10. Morphological acquisition: Goldwater, Griffiths, and Johnson (2006); Frank, Ichinco, and Tenenbaum (2008)

11. Part-of-speech tagging: Goldwater and Griffiths (2007)

12. Planning and inferences about agents: Verma and Rao (2006); Baker, Tenenbaum, and Saxe (2007); Lucas et al. (submitted)

13. Reasoning using logical rules: N. Goodman, Griffiths, Feldman, and Tenenbaum (2007)

14. Theory learning: Kemp et al. (2010)

15. Verb learning: Alishahi and Stevenson (2008); Hsu (2009); Perfors, Tenenbaum, and Wonnacott (2010)

16. Word learning: Xu and Tenenbaum (2007); Andrews, Vigliocco, and Vinson (2009); Frank, Goodman, and Tenenbaum (2009)

17. Word segmentation: Goldwater, Griffiths, and Johnson (2007); Frank, Goldwater, Griffiths, and Tenenbaum (2007)

A.3 Mathematical foundations

The mathematical foundations of Bayesian inference extend back decades if not centuries. Sivia (1996) and P. Lee (1997) are good introductory textbooks; more advanced texts include Berger (1993) and Jaynes (2003).

As discussed briefly within the paper, Bayesian probability theory brings up several issues related to the subjectivity of the prior probability, relation to frequentist statistical approaches, and the interpretation and nature of probability in the first place. Classic work from a frequentist perspective includes Fisher (1933) and van Dantzig (1957), and from a Bayesian perspective Jeffreys (1939), Cox (1946), Savage (1954), and de Finetti (1974). Box and Tiao (1992) explores how the frequentist approach may be interpreted from a Bayesian perspective, and Jaynes (2003) provides a nice overview, bringing the threads of many of these arguments together.

There is a great deal of work exploring the relationship between Bayesian learning and information-theoretic or minimum description length (MDL) approaches. Vitànyi and Li (2000), Jaynes (2003), MacKay (2003) and Grünwald, Myung, and Pitt (2005) provide excellent discussions and overview of some of the issues that arise. More classic texts include Rissanen (1978), Solomonoff (1964), and Kolmogorov (1965).

A.4 Technical details

One of the largest areas of research in machine learning is focused on developing more effective techniques for searching the (sometimes quite large) hypothesis spaces defined by Bayesian models. One of the standard approaches includes Markov chain Monte Carlo (MCMC), which is introduced and explained in Neal (1993); MacKay (1998); Gilks, Richardson, and Spiegelhalter (1996) and Gelman, Carlin, Stern, and Rubin (2004) provide examples of how to incorporate these methods into Bayesian models. In addition, sequential Monte Carlo methods (e.g., Doucet, Freitas, & Gordon, 2001) provide a means to explore capacity limitations and a more "on-line" processing approach.