Opinion TRENDS in Cognitive Sciences Vol.xx No.xx Monthxxxx

Full text provided by www.sciencedirect.com

Special Issue: Probabilistic models of cognition

Theory-based Bayesian models of inductive learning and reasoning

Joshua B. Tenenbaum[1], Thomas L. Griffiths[2] and Charles Kemp[1]

[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI, USA

Inductive inference allows humans to make powerful generalizations from sparse data when learning about word meanings, unobserved properties, causal relationships, and many other aspects of the world. Traditional accounts of induction emphasize either the power of statistical learning, or the importance of strong constraints from structured domain knowledge, intuitive theories or schemas. We argue that both components are necessary to explain the nature, use and acquisition of human knowledge, and we introduce a theory-based Bayesian framework for modeling inductive learning and reasoning as statistical inferences over structured knowledge representations.

Introduction

Human cognition rests on a unique talent for extracting generalizable knowledge from a few specific examples. Consider how a child might first grasp the meaning of a common word, such as 'horse'. Given several examples of horses labeled prominently by her parents, she is likely to make an inductive leap that goes far beyond the data observed. She could now judge whether any new entity is a horse or not, and she would be mostly correct, except for the occasional donkey, deer or camel. The ability to generalize from sparse data is crucial not only in learning word meanings, but in learning about the properties of objects, cause–effect relations, social rules, and many other domains of knowledge.

This article describes recent research that seeks to understand human inductive learning and reasoning in computational terms (see also Conceptual Foundations Editorial by Chater, Tenenbaum and Yuille in this issue). The goal is to build broadly applicable, quantitatively predictive models that approximate optimal inference in natural environments, and thereby explain why human generalization works the way it does and how it can succeed given such sparse data [1,2]. Our focus is on computational-level theories [3], characterizing the functional capacities of human inference rather than specific psychological processes that implement those functions.

Most previous accounts of inductive generalization represent one of two approaches. The first focuses on relatively domain-general, knowledge-independent statistical mechanisms of inference, based on similarity, association, correlation or other statistical metrics [1,4–13]. This approach has led to successful mathematical models of human generalization in laboratory tasks, but fails to account for many important phenomena of learning and reasoning in complex, real-world domains, such as intuitive biology, intuitive physics or intuitive psychology. The second approach aims to capture more of the richness of human inference, by appealing to sophisticated domain-specific knowledge representations, or intuitive theories [14–20]. An intuitive theory may be thought of as a system of related concepts, together with a set of causal laws, structural constraints, or other explanatory principles, that guide inductive inference in a particular domain. However, theory-based approaches to induction have been notoriously difficult to formalize, particularly in terms that make quantitative predictions about behavior or can be understood in terms of rational statistical inference.

We will argue for an alternative approach, where structured knowledge and statistical inference cooperate rather than compete, allowing us to build on the insights of both traditions. We cast induction as a form of Bayesian statistical inference over structured probabilistic models of the world. These models can be seen as probabilistic versions of intuitive theories [14,18,20] or schemas [21,22], capturing the knowledge about a domain that enables inductive generalization from sparse data. This approach has only become possible in recent years, as advances in artificial intelligence [23] and statistics [24] have provided essential tools for formalizing intuitive theories and theory-based statistical inferences. The influence is bidirectional, as these Bayesian cognitive models have led to new machine-learning algorithms with more powerful and more human-like capacities [25,26].

Theory-based Bayesian models

Theory-based Bayesian models of induction focus on three important questions: what is the content of probabilistic theories, how are they used to support rapid learning, and how can they themselves be learned? The learner evaluates hypotheses h about some aspect of the world – the meaning of a word, the extension of a property or category, or the presence of a hidden cause – given observed data x and subject to the constraints of a

Corresponding author: Tenenbaum, J. (jbt@mit.edu).

www.sciencedirect.com 1364-6613/$ - see front matter © 2006 Elsevier Ltd. All rights reserved. doi:10.1016/j.tics.2006.05.009

background theory T. Hypotheses are scored by computing posterior probabilities via Bayes' rule:

$$P(h|x,T) = \frac{P(x|h,T)P(h|T)}{\sum_{h' \in H_T} P(x|h',T)P(h'|T)} \quad (1)$$

The likelihood $P(x|h,T)$ measures how well each hypothesis predicts the data, and the prior probability $P(h|T)$ expresses the plausibility of the hypothesis given the learner's background knowledge. Posterior probabilities $P(h|x,T)$ are proportional to the product of these two terms, representing the learner's degree of belief in each hypothesis given both the constraints of the background theory T and the observed data x (see the Technical Introduction to this special issue by Griffiths and Yuille for further background: Supplementary material online) Adopting this Bayesian framework is just the starting point for our cognitive models. The challenge comes in specifying hypothesis spaces and probability distributions that support Bayesian inference for a given task and domain. In theory-based Bayesian models, the domain theory plays this crucial role.

More formally, the domain theory T generates a space H_T of candidate hypotheses, such as all possible meanings for a word, along with the priors $P(h|T)$ and likelihoods $P(x|h,T)$. Prior probabilities and likelihoods are thus not simply statistical records of the learner's previous observations, as in some Bayesian analyses of perception and motor control [27,28], or previous Bayesian analyses of inductive reasoning [29]. Neither are they assumed to share a single universal structure across all domains, as in Shepard's pioneering Bayesian analysis of generalization [30]. Rather, they are products of abstract systems of knowledge that go substantially beyond the learner's direct experience of the world, and can take qualitatively different forms in different domains.

We will distinguish at least two different levels of knowledge in a theory (Figure 1). Although intuitive theories may well be much richer than this picture suggests, we focus on the minimal aspects of theories needed to support inductive generalization. The base level of a theory is a structured probabilistic model that defines a probability distribution over possible observables – entities, properties, variables, events. This model is typically built on some kind of graph structure capturing relations between observables, such as a taxonomic hierarchy or a causal network, together with a set of numerical parameters. The graph structure determines qualitative aspects of the probabilistic model; the numerical parameters determine more fine-grained quantitative details. At a higher level of knowledge are abstract principles that generate the class of structured models a learner may consider, such as the specification that a given domain is organized taxonomically or causally. Inference at all levels of this theory hierarchy (Figure 1) – using theories to infer unobserved aspects of the data, learning structured models given the abstract domain principles of a theory, and learning the abstract domain principles themselves – can be carried out in a unified and tractable way with hierarchical Bayesian models [24].

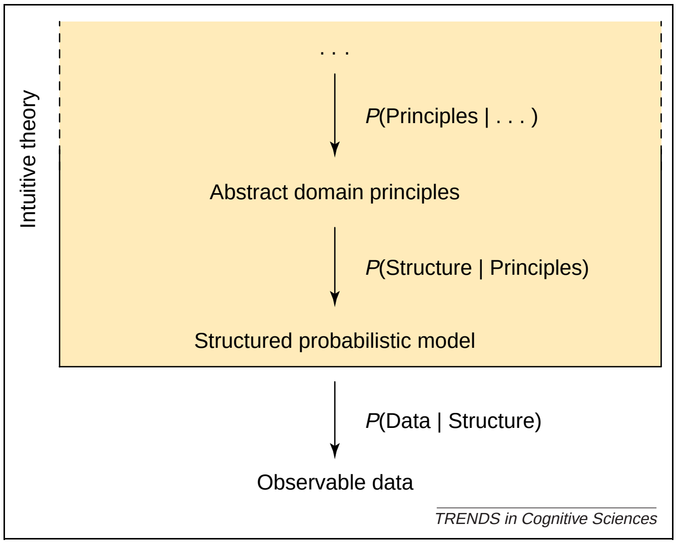
Intuitive theory
. . .
P(Principles | . . .)
Abstract domain principles
P(Structure | Principles)
Structured probabilistic model
P(Data | Structure)
Observable data
TRENDS in Cognitive Sciences

Figure 1. A hierarchical Bayesian framework for theory-based induction. The learner observes data about the world (e.g. examples of objects that a word refers to) and must predict other unobserved data (e.g. which other objects the word can refer to). The learner's intuitive theory generates hypotheses that can explain the observed data and that support the desired predictions. The theory represents knowledge on at least two levels of abstraction: a structured probabilistic model generates expectations about the probability of possible data sets, while more abstract domain principles generate the space of possible structures that the learner may consider. Each level generates the hypotheses and probability distributions that support learning at the level below. Priors for abstract domain principles can come from multiple sources, including higher-level domain knowledge or domain-general conceptual resources.

The following sections describe theory-based Bayesian models for several important inductive tasks, contrasting them with alternative approaches emphasizing either statistical learning or structured knowledge alone. We begin with the task of learning words or category labels, and focus on the lowest level of inference: theory-based generalization. Then we illustrate the full hierarchical approach in two other domains, property induction and causal inference.

Learning names for things

Behavioral studies of human inductive generalization arguably began with the study of category learning [31]. The basic experimental task presents learners with a set of objects or visual stimuli, and a verbal label (e.g. 'blicket') that applies to a subset of the objects. Learners observe several examples of blickets, and perhaps negative examples (non-blickets), and must then infer which other objects the label applies to.

These artificial category-learning tasks abstract the essence of the problem children face in learning words for kinds of things, and formal models of category-learning and word-learning have developed in parallel. They typically rely on bottom-up general-purpose statistical mechanisms, either explicitly probabilistic [1,32] or framed in terms of similarity or association [12,13,8]. These models assume relatively simple notions of categories and how labels relate to categories: for instance [32], each object belongs to a single category, and each label picks out a unique category, so each object receives exactly one label. However, people's representations of categories and word meanings are considerably more structured, reflecting their intuitive domain theories. The need for a more theory-based approach has often been pointed out [18,14,33,16,20], but rarely pursued by formal modelers.

Insights from both these traditions come together in a Bayesian framework [34–36]. In terms of Equation 1, hypotheses about the meaning of a novel label refer to subsets of objects – candidate extensions for a word's meaning or a category to be labeled. Abstract knowledge about category structure, word usage, and word-category mappings generates the priors and likelihoods for these hypotheses. Tenenbaum and Xu [35,36] focus on learning names for object-kind concepts, which are typically organized into a tree-structured taxonomy with labels at various levels [37,16]. Accordingly, the hypothesis space of candidate word meanings consists of all subtrees in a tree-structured taxonomy of objects (in Figure 2a, subtrees correspond to *basset hounds*, *dogs*, *animals*, etc.). Other logically possible subsets of objects not corresponding to subtrees are effectively assigned zero (or very low) prior probability. The prior can be further restricted to favor mappings of words onto basic-level categories [37], or to disfavor mapping two words onto exactly the same concept [38]. The likelihood embodies a pragmatic assumption that words will be used by a competent and cooperative speaker [17], and that the objects labeled are a fair random sample from the set of objects that the word applies to.

These priors and likelihoods combine to explain how children generalize object labels from one or several examples. Figure 2b shows that generalization follows a gradient according to taxonomic distance, which sharpens up given multiple examples to focus on the most specific consistent taxonomic category: e.g. *basset hounds* if the examples are all basset hounds, or *dogs* if the examples are all dogs but different kinds of dogs. Generalization along taxonomic contours derives from the tree-structured prior. In principle, a prior could be defined over other representational structures, such as a Euclidean space recovered from similarity judgments, as is common in similarity-based models [12]. But to date, tree-structured priors have been the basis for the most accurate Bayesian models of word-learning, consistent with a proposed taxonomic bias in children's word learning [16]. The sharpening of generalization with more examples derives from the likelihood: a single example is not highly diagnostic about the scope of the word's extension,

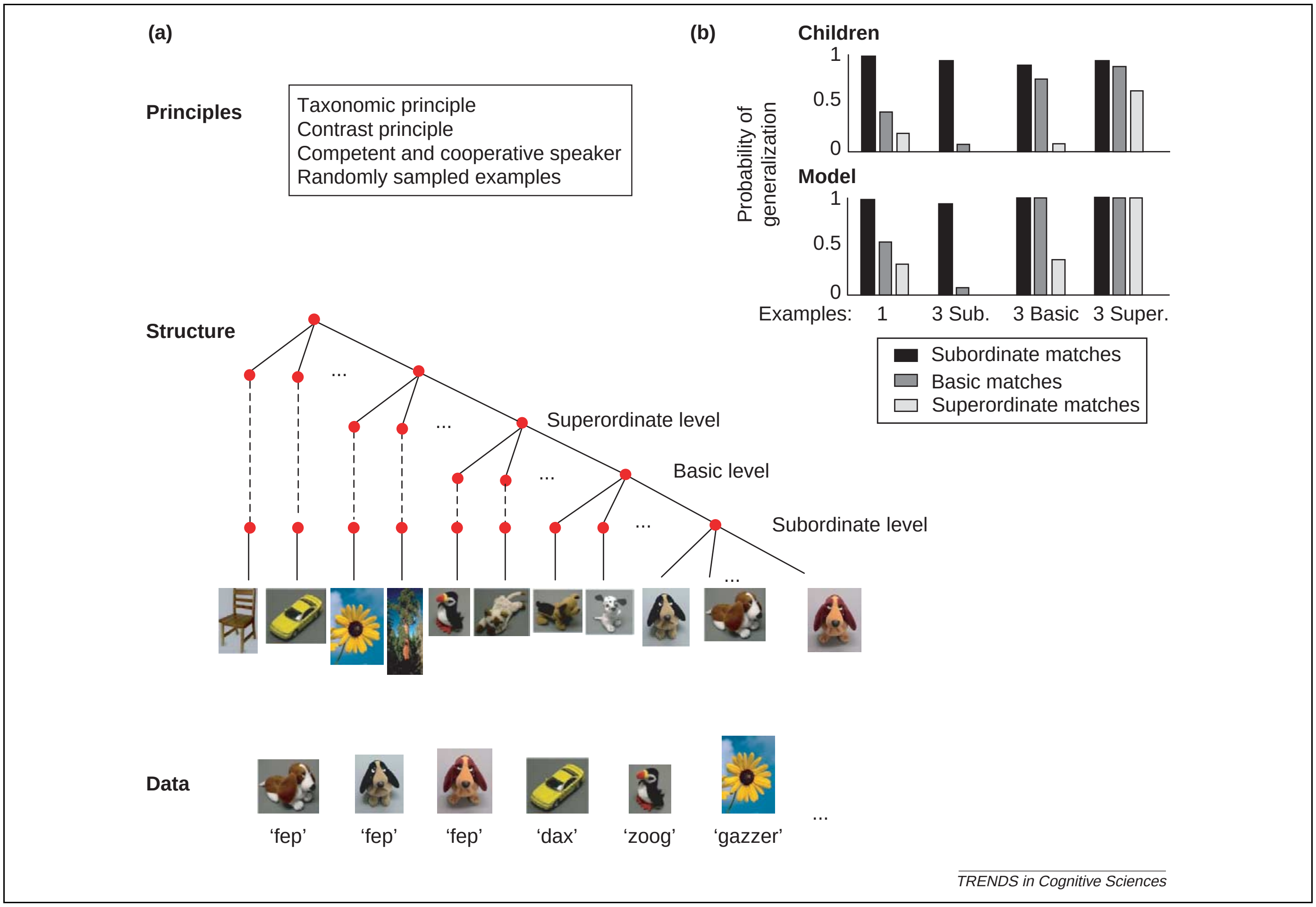

Figure 2. Theory-based Bayesian word learning. (a) Learning words for object categories, from examples of object–label pairs. Objects are given unfamiliar labels to illustrate the problem faced by a child learner. A tree-structured taxonomy of categories comprises the hypothesis space of word meanings: each node in the tree (red dot) is a possible extension of a word. Abstract principles constrain the structure of this hypothesis space, and generate the priors and likelihoods necessary to evaluate these hypotheses given data [35]. **(b)** Comparison of the model's predictions with 4-year-old children's patterns of generalization (preliminary findings from [36]). For both children and the model, the probability of generalization decreases with taxonomic distance to the examples, and this gradient becomes sharper as more examples are observed. See text for more discussion.

but observing several examples drawn at random, it would be a highly suspicious coincidence for all examples to fall within a given taxonomic category (e.g. *basset hounds*) if the word in fact had a much broader extension (e.g. *dogs*), so the most specific consistent hypothesis is strongly preferred.

The tendency for smaller, more specific hypotheses to be increasingly preferred over larger, more general hypotheses as more examples are observed is a general principle of Bayesian learning when randomly sampled examples are assumed. Tenenbaum and Griffiths [34] referred to this as the 'size principle' and showed how it could potentially explain a wide range of phenomena in category learning, generalization, and similarity judgment, which were not previously unified under a single rational-inference account. The random-sampling assumption is not always valid, of course, and the size principle may be accordingly defeasible. Xu and Tenenbaum (unpublished data) have found its effects are reduced or eliminated when word learners (children or adults) are given examples that are clearly not drawn as independent random samples. This Bayesian framework has been extended to learning other aspects of linguistic meaning, using differently structured hypothesis spaces appropriate for learning verb frames [39], adjectives [40], or anaphora resolution [41]. There are also clear connections to Bayesian syntactic acquisition (see Chater and Manning, this issue [42]).

These Bayesian analyses have focused on learning at the bottom level of Figure 1 – learning about which words can refer to which entities, situations or properties. Future work should explore learning the higher-level knowledge that supports these inferences – for instance, how people learn the principles that structure word-category mappings, or the relevant taxonomic tree of categories [43,16,15]. The following section describes a closely related learning task where Bayesian inferences at higher levels of Figure 1 have been analyzed more systematically.

Reasoning about hidden properties

Many kinds of predicates may be true of a given entity. Some of these predicates correspond to category labels (*is a horse*, *is a fish*) but many correspond to properties, such as *is brown*, *has a spleen*, or *can fly*. Property induction has been the subject of numerous behavioral experiments and formal models. In a typical task, learners find out that one or more categories have a novel property, and must decide how to extend the property to other categories in the domain. For instance, subjects might be told that gorillas and lions carry a certain gene, and asked to judge how likely it is that monkeys also carry this gene [4,5].

Theory-based property induction

The most systematic studies of property induction have used biological species and blank properties: properties like *has the T4 gene* that are unfamiliar but recognizably biological. A tree-based Bayesian model [44] similar to the Tenenbaum and Xu [35] word-learning model accounts well for judgments about blank biological properties. The model assumes that species are organized into a tree-structured taxonomy [15], and that properties are generated by a mutation process over this tree (Figure 3a, left). The mutation process generates a prior over candidate extensions of predicates that is more flexible than traditional symbolic semantic hierarchies (or the word-learning prior in [35]): properties that pick out a single subtree of the taxonomy are favored, but polyphyletic properties (those that arise independently in two or more subtrees) are also allowed (Figure 3b). This model approximates optimal inference for biological species and their properties, which are in fact generated by a stochastic branching process – the process of evolution. Both the tree structure and some mutation-like process for generating property distributions seem to be important; Bayesian models using a range of other priors have consistently correlated less well with people's judgments [45,44,46].

Previous computational models for blank-property induction have used more generic knowledge representations, such as pairwise similarities [5] or collections of features [6]. In comparison with these models, theory-based approaches have clear advantages in explaining inferences about other kinds of predicates, where more specialized prior knowledge is involved. Qualitatively different patterns of generalization have been found for anatomical properties, behavioral properties and disease properties [29,47,48]. To cite one classic example, given that Poodles can bite through wire, it seems likely that German Shepherds can bite through wire, but knowing that Dobermans can bite through wire provides less support for the same conclusion about German Shepherds [48]. These inferences cannot rely on similarity, because German Shepherds are more similar to Dobermans than to Poodles. Other inferences appear to rely on asymmetric causal relations: for example, a disease carried by salmon is more likely to be found in grizzly bears than vice versa [47].

Bayesian models can account for these different patterns of generalization by using different priors [29]. In terms of Figure 1, inferences about different kinds of observable predicates are based on different kinds of structured probabilistic models, which are in turn governed by different abstract domain principles. Several examples are shown in Figure 3a. In each case, the taxonomic tree structure and the mutation process of the default model are replaced by a differently structured graph and a different kind of stochastic process over that structure. For inferences about disease predicates (e.g. *carries Leptospirosis*), the prior is generated by a noisy transmission process over a directed food-web network (Figure 3a, middle). This prior captures the asymmetry of generalizations in this domain – that a prey species is more likely to share a disease with its predators than vice versa – and accurately predicts people's inductive judgments about disease predicates (Figure 3c). A predicate like *can bite through wire* or *weighs more than an anvil* corresponds to an unknown threshold along some known dimension (e.g. strength or size). A linearly ordered graph can represent the relevant dimension, and judgments about threshold predicates [48,7] can be modeled

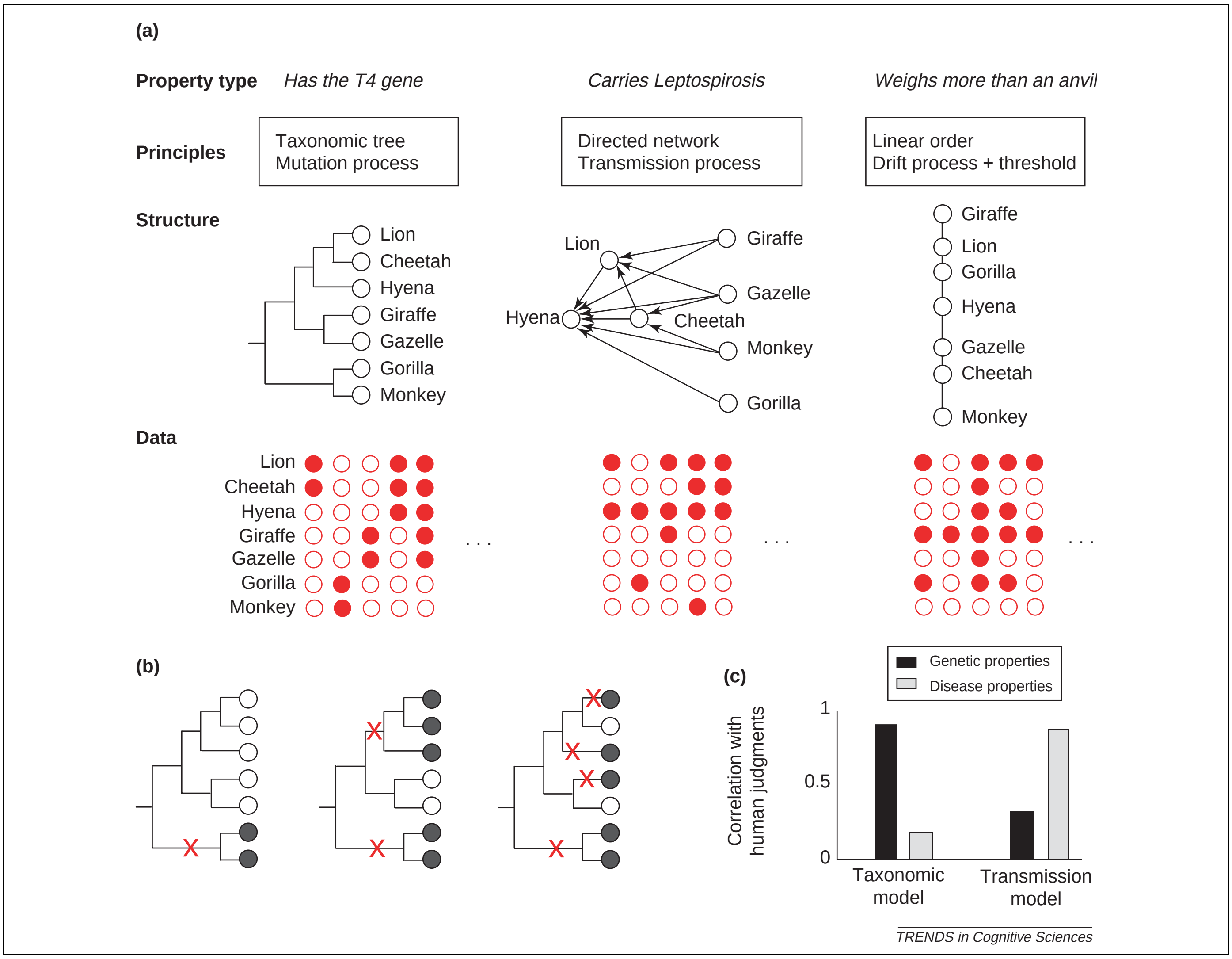

Figure 3. Theory-based Bayesian property induction. (a) Three models for property induction: a taxonomic model (left), a food-web model (centre) and a dimensional-threshold model (right). Each model assumes that properties are generated by a different probabilistic process over a different kind of graph structure, and each model is appropriate for a different kind of property. In the taxonomic model, properties are generated by a mutation process: there is a small probability of a property appearing at any point along any branch of the tree. In the food-web model, properties are generated by a causal transmission process: there is a small probability of a property arising spontaneously in any species, and a high probability of transmitting that property from the species to each of its predators. In the dimensional-threshold model, properties are generated by a random-drift process biased such that species towards one end of the dimension are increasingly likely to have the novel property. The 'Data' level of the figure shows properties with high prior probability under each of these models: e.g. the dimensional-threshold model favors hypotheses that include all species beyond some point in the linear order. **(b)** Three possible outcomes if a property is generated by the taxonomic model, shown in order of decreasing prior probability: properties are most likely if they can be explained by a small number of mutations, and if those mutations occur on long branches. **(c)** The importance of domain theories in Bayesian models of property induction is illustrated by a double dissociation in model predictions for two different kinds of properties (preliminary findings from [70]). Participants learned both a taxonomy and a food web over a set of species and were asked to make inductive judgments about either genetic or disease properties. The Bayesian taxonomic model correlated strongly with judgments for genetic properties but not disease properties, and vice versa for the Bayesian food-web model.

assuming a prior generated by a random drift process over that graph (Figure 3a, right).

Learning theories to support property induction

If differently structured theories are necessary to account for inferences about different kinds of predicates, it becomes even more pressing to explain how these theories could be acquired. Bayesian approaches can address this question at all levels of the theory hierarchy in Figure 1, and we illustrate by showing how the taxonomic theory might be acquired (Figure 4). First, consider the problem of learning the tree structure given raw observable data, in the form of a large collection of species-property pairs (e.g. *lions have sharp teeth*, *chimps have hair*, etc.). There are many different ways to organize species into a tree (Figure 4a), but we can search for the tree that maximizes the likelihood $P(\text{Data}|\text{Structure})$ for the dataset of observed properties [49,46]. Intuitively, the best choice allows features to vary smoothly over the tree: for example, because gorillas and monkeys share many properties, these species should be located nearby in the tree.

This approach to learning a structured probabilistic model relies crucially on abstract knowledge at the highest level in Figure 1: a 'taxonomic principle' specifying that living kinds should be represented by a tree structure. Could such an abstract domain principle itself by acquired? Abstract knowledge of this sort is often thought to be innate [15,50], perhaps because it seems so remote from the data of experience. Given an appropriate hypothesis space,

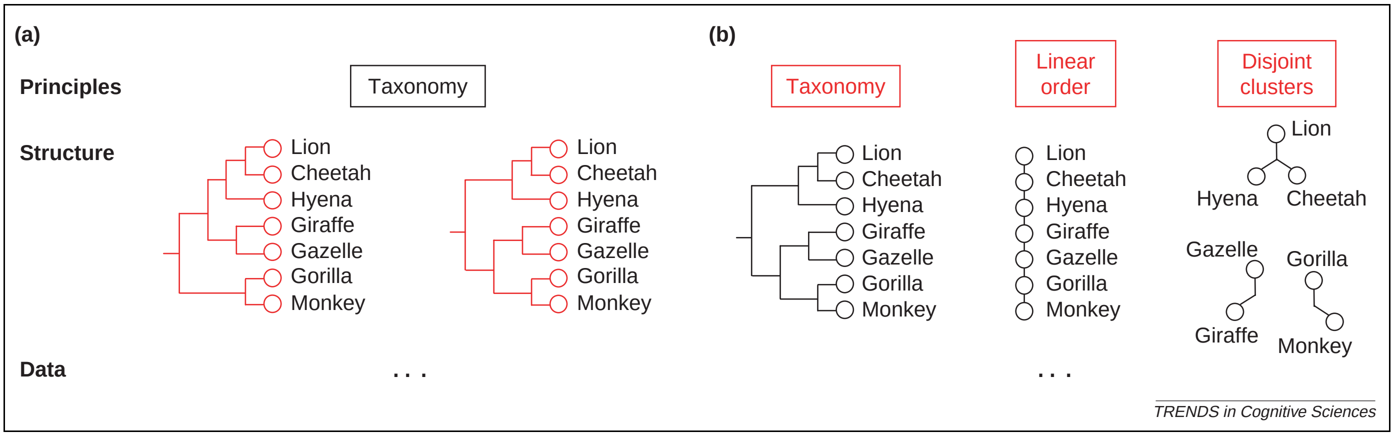

Figure 4. Learning a theory for how biological properties are distributed over species. (a) Given abstract domain knowledge that species should be organized in a taxonomic tree, with properties varying smoothly over that tree, a Bayesian learner can infer the tree structure that best explains a set of observed properties. Two ways to organize animal species into a taxonomy are shown. The preferred structure will be the tree over which observed properties vary most smoothly. **(b)** Animal species may be organized according to various different structural principles, such as the three shown here. Bayesian inference in the hierarchical framework of Figure 1 can select the organizing principles best supported by a set of observed properties.

however, Bayesian inference can account for knowledge acquisition at any level of abstraction. Suppose that the learner has access to a repertoire of different structural forms that includes taxonomic trees as well as other basic alternatives, such as one-dimensional orders and disjoint clusters (Figure 4b). Choosing the best form involves a trade-off between complexity and fit to the data, which can be formalized in terms of the hierarchical Bayesian framework of Figure 1. Kemp *et al.* [46] showed that under this trade-off, the judged properties of biological species are better accounted for by trees than either linear orders or clusters.

In summary, the hierarchical Bayesian framework of Figure 1 supports a unified learning model that takes as input data a collection of species-property pairs, and subsequently discovers the taxonomic principle, discovers the best tree structure for the species in the dataset, and makes predictions about how to generalize new, sparsely observed properties. It thus explains how the theory in Figure 3a (left) might be discovered from raw data, and likewise could explain the origins of hypothesis spaces and inductive biases for word learning in young children discussed in the previous section. Theories can also be acquired through other processes, such as explicit instruction; for example, food-web relations are typically learned that way. Unlike similarity-based accounts of property induction, a hierarchical Bayesian approach naturally accommodates explicit instructions at any level of abstraction, as children typically receive from parents or in school: for example, *dolphins breath air* (at the level of observable data), *dolphins are mammals* (at the level of structure), and *living things can be organized into a tree* (at the level of abstract principles). This approach could explain how hearing a single statement about domain structure (e.g. *dolphins are mammals*) might lead to dramatic changes in inferences about unobserved properties.

Causal learning and reasoning

The role of intuitive theories in learning and reasoning has been most prominently studied in the context of causal cognition [33,18,51,19]. For many authors, causality is central to the notion of a theory. Carey, for instance, suggests that a theory comprises ‘a set of phenomena that are in its domain, the causal laws and other explanatory mechanisms in terms of which the phenomena are accounted for, and the concepts in terms of which the phenomena and explanatory apparatus are expressed’ ([33], p. 394). The hierarchical Bayesian framework of Figure 1 gives a unified account of inference at all these levels, suggesting how causal models can be used to predict and explain observable events, how domain-specific principles guide construction of these models, and how that abstract domain knowledge could itself be learned.

At the lowest level of the hierarchy are inferences about variables characterizing observable events. For instance, a doctor might observe certain aspects of a patient's state, such as symptoms and risk factors, and want to predict others, such as diseases or future symptoms. These tasks have often been viewed as bottom-up statistical inferences [52,1], but there is evidence to suggest that these predictions are often driven by causal knowledge [53–55]. Recent work has tried to explain the relevant knowledge and inferences in rational terms using the formalism of causal graphical models [56,57]. These models constitute a particular kind of structured probabilistic model at the middle level of Figure 1. For example, Figure 5a shows a causal model that could be used to make inferences about the diseases of patients based on their symptoms and risk factors.

Theory-based induction of causal structure

Many recent studies have examined how people learn these causal models. Again there have been both bottom-up and top-down proposals. Bottom-up approaches detect statistical cues to causal structure, such as contingency between variables [9], normalized probabilistic contrast [10], or partial correlations [11]. These cues, when they can be detected reliably, allow causal structure to be learned for any kinds of variables, without substantive prior knowledge. However, both adults and young children

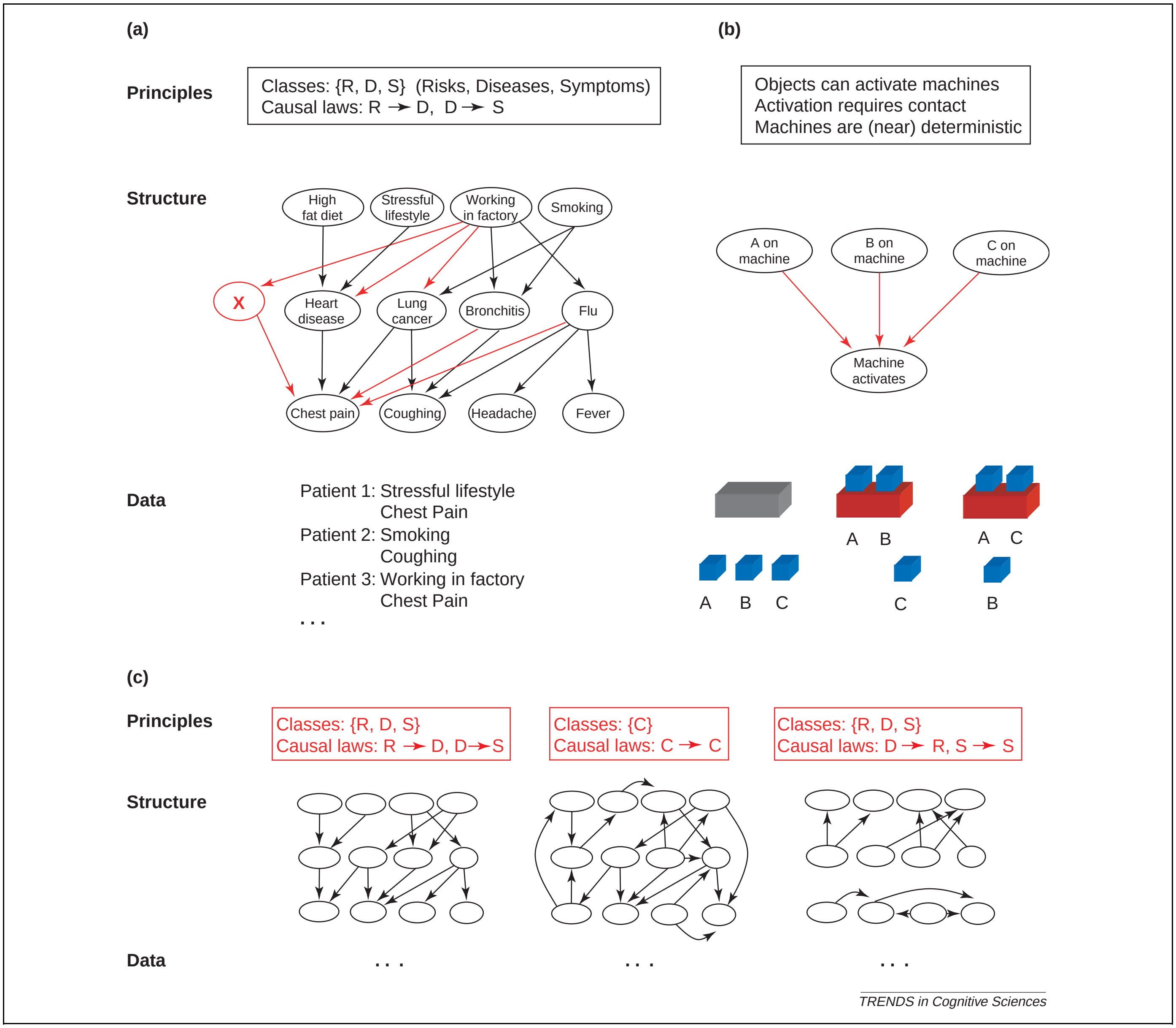

Figure 5. Theory-based Bayesian causal induction. Abstract causal principles constrain the causal structures that may be learned to capture dependencies among observable variables. **(a)** Abstract knowledge in a simple medical domain can be represented using a 'graph schema', a probabilistic generative grammar for graphs. Variables fall into three classes – risk factors, diseases, and symptoms – with causal influences only from risks to diseases and diseases to symptoms. Given a newly observed correlation (e.g. between working in a factory and chronic chest pain), the graph schema generates a constrained set of hypotheses for how that data might be explained (shown in red). In the simplest hypotheses, a disease known to be caused by working in a factory might cause chest pain, or a disease known to cause chest pain might actually be produced by working in a factory. Failing these possibilities, the learner could posit a new disease X, which has chest pain as a symptom and is caused by working in a factory. Other hypotheses that may be simpler a priori but which violate the theory would never be considered, such as a direct causal link from working in a factory to chest pain, or from chest pain to working in a factory. **(b)** The abstract knowledge that supports causal learning in a simple physical system, the 'blicket detector' [11,63], can be formalized using probabilistic predicate logic [64,62]. The theory includes several principles: there is some probability ρ that any object has the power to activate the machine; an efficacious object will activate the machine upon contact with probability near 1; activation has no other causes. Possible causal relations, given a sequence of interactions between blocks and the machine, are shown in red. In a context in which most objects have failed to activate the machine (ρ is small), both people and the theory-based Bayesian model infer that object A probably activates the machine, whereas B and C probably do not [63,64]. **(c)** The infinite relational model [26] supports a hierarchical Bayesian approach to learning simple forms of abstract causal theories. Graph schemas with different numbers of classes and appropriately defined causal laws can be inferred to explain different kinds of causal network structures that a learner might encounter.

frequently make correct inferences about causal structure from just a handful of observations, far too few to compute contingency or correlation reliably. These inferences must be supported by more abstract prior knowledge, such as knowledge about the kinds of causal mechanisms or structures in a domain, which has been the focus of most top-down approaches to causal learning [58–60]. The theory-based Bayesian framework once again provides the means to integrate bottom-up and top-down influences. In terms of Figure 1, P(Data|Structure) measures the likelihood for a particular causal model – how well that causal structure explains the observed pattern of events. P(Structure|Principles) measures the causal model's prior probability – how well it fits with the learner's abstract domain knowledge. The best causal model maximizes the product of these two probabilities.

If a causal model represents the most basic kind of causal theory [51,54], the abstract domain knowledge that allows these models to be learned can be thought of as a higher-level theory – a 'theory for theories'. Cognitive developmentalists [19] have often emphasized the importance of larger-scale 'framework' theories, which constrain the specific theories a learner considers, but they have not been treated computationally. The concept of probabilistic models for structured representations is well-developed in computational linguistics [42], where abstract syntactic principles may be formalized in terms of probabilistic grammars that generate admissible syntactic structures in a language with varying probabilities. Several proposals for formalizing abstract causal theories have been inspired by this linguistic analogy [61]. These 'causal grammars' [62] share the idea of generating causal graphical models based on an ontology, which identifies the types of entities in a domain and the predicates that can apply to them, and a set of causal laws, which specify the form of allowed causal relationships between predicates. Loosely speaking, the ontology generates the variables that appear in causal graphical models, and the causal laws generate the edges and associated conditional probabilities.

Figure 5 sketches two examples of this approach to theory-based causal induction [61,62]. First, consider learning a causal network relating risk factors, diseases, and symptoms, given data on patients' states. The task is greatly simplified by an ontology that divides the variables into three classes – diseases, symptoms, and risk factors – and causal laws defined over that ontology, specifying that direct influences only exist between risk factors and diseases, and between diseases and symptoms. These principles can be formalized using a kind of probabilistic graph grammar that [62] call a 'graph schema'. The grammar places strong constraints on the causal structures a learner must consider, and thus allows strong inferences about causal structure to be drawn from sparse data (Figure 5a).

Second, consider recent studies exploring theory-based causal learning using simple physical systems such as the 'blicket detector' [11,63,64]. Learners are shown a number of objects, along with a machine that 'activates' (lights up and makes noise) whenever certain blocks are placed on it. After observing several trials in which various combinations of objects are placed on the machine, participants are asked which objects have the hidden causal power to activate the machine. Figure 5b shows how these judgments can be modeled as theory-based Bayesian inferences guided by several domain principles, such as the 'activation law': the machine activates if and only if it is in contact with an object that has the hidden causal power. This theory can be cast more formally in probabilistic predicate logic [62]. Unlike bottom-up approaches to causal learning, this account naturally explains many findings where adults and young children make correct inferences from just a few trials, even for objects that have never appeared on the machine alone [63,64]. This general framework for theory-based causal induction has been used to model how people learn more complex physical systems with hidden variables and dynamic causes [65], and how people choose optimal experiments to perform in causal learning [66].

Learning abstract causal theories

The question of how abstract causal principles, or 'framework theories', might themselves be learned is a major open question in both artificial intelligence and cognitive development. For some simple kinds of framework knowledge, such as the probabilistic graph grammars discussed above, it is possible to formulate the learning problem as a Bayesian inference that can be approximated with tractable search algorithms. The infinite relational model (IRM) [26] assumes that variables come in one or more classes, with relations between them depending on these classes. The model can be used to infer the number of classes, which variables are in which classes, and what kinds of relationships hold between classes, directly from data (Figure 5c). This approach is capable of learning the abstract principles of the disease theory shown in Figure 5a, but not the richer theories based on probabilistic predicate logic needed to explain inferences in some other systems such as the blicket detector (Figure 5b). Several methods for learning in probabilistic logical systems have recently been introduced in artificial intelligence [67,68], however, and these methods could provide the basis for more powerful models of human theory acquisition.

Conclusion

The theory-based Bayesian framework provides a formal means to address several fundamental questions about human cognition. What is the content and form of human knowledge, at multiple levels of abstraction? How can abstract domain knowledge guide learning of new concepts? How can abstract domain knowledge be learned? What conceptual resources must be built in innately? How do mechanisms of statistical learning and

Box 1. Questions for future research

- How might theory-based Bayesian models apply to other aspects of cognition where structured knowledge and inductive inference appear to interact, such as intuitive physics, intuitive psychology, or moral judgment?
- Can Bayesian models of concept learning and word learning be integrated with Bayesian models of syntactic acquisition [42] to give a unified approach to language development?
- How do Bayesian models of inductive inference relate to probabilistic models of (apparently) deductive thinking, such as hypothesis testing or syllogistic reasoning [2]?
- Could a hierarchical Bayesian approach provide insight into other functions of intuitive theories besides induction, such as analogy or explanation?
- Like other theoretical paradigms, the Bayesian approach is not meant to be falsifiable in the same sense that a specific computational model should be; it should be judged in terms of whether it leads to specific models with explanatory value across a range of different data sets. What aspects of intuitive theories and theory-based inference will prove difficult to explain from a Bayesian view?
- How could Bayesian approaches to induction – considered here strictly at the level of computational theory [3] – be implemented with tractable algorithms? How can they can be reconciled with psychological processes that have sometimes looked dramatically inconsistent with Bayesian principles [69]?
- Philosophers of science and developmental psychologists have often exchanged ideas about the structure, function, and origins of theories. Can we build hierarchical Bayesian models of scientific theories analogous to our models of intuitive theories?

inference interact with – and operate over – structured symbolic knowledge? Traditionally, computational accounts that aim to explain a broad spectrum of human cognition have focused exclusively on either sophisticated inference processes or sophisticated knowledge representations. Our view embraces both, and highlights their interactions.

It is far too soon to say what a mature computational theory of inductive learning and reasoning will look like. The real-world problems that children and adults face are still well beyond the scope of our models, and issues of algorithmic and psychological plausibility will have to be addressed (see Box 1, and Editorial 'Where next?' in this issue). Yet as future work on induction unfolds, one idea should play a crucial role in any explanatory account: probabilistic inference over hierarchies of increasingly abstract, flexibly structured representations of the world.

Acknowledgements

We thank the current and past members of the Computational Cognitive Science group at MIT for innumerable discussions of the work described here. We owe particular debts to Fei Xu, Patrick Shafto, and Sourabh Niyogi, who have collaborated closely on the projects described here. We acknowledge support from NTT Communication Sciences Laboratories, Mitsubishi Electric Research Labs, the National Science Foundation, the James S. McDonnell Foundation Causal Learning Collaborative, the Paul E. Newton Career Development Chair (J.B.T.), the Stanford Graduate Fellowships (T.L.G.), and the William Asbjornsen Albert Memorial Graduate Fellowship (C.K.).

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tics.2006.05.009

References

1 Anderson, J.R. (1990) *The Adaptive Character of Thought*, Erlbaum
2 Oaksford, M. and Chater, N. (1999) Ten years of the rational analysis of cognition. *Trends Cogn. Sci.* 3, 57–65
3 Marr, D. (1982) *Vision*, W.H. Freeman
4 Rips, L.J. (1975) Inductive judgments about natural categories. *J. Verbal Learn. Verbal Behav.* 14, 665–681
5 Osherson, D.N. *et al*. (1990) Category-based induction. *Psychol. Rev.* 97, 185–200
6 Sloman, S.A. (1993) Feature-based induction. *Cogn. Psychol.* 25, 231–280
7 Blok, S. *et al.* (2003) Probability from similarity. In *Working Papers of the 2003 AAAI Spring Symposium on Logical Formalization of Commonsense Reasoning*, pp. 43–50, AAAI Press
8 Regier, T. (2005) The emergence of words: attentional learning in form and meaning. *Cogn. Sci.* 29, 819–865
9 Shanks, D.R. (1995) *The Psychology of Associative Learning*, Cambridge University Press
10 Cheng, P. (1997) From covariation to causation: A causal power theory. *Psychol. Rev.* 104, 367–405
11 Gopnik, A. *et al*. (2004) A theory of causal learning in children: Causal maps and Bayes nets. *Psychol. Rev.* 111, 1–31
12 Nosofsky, R.M. (1986) Attention, similarity, and the identification-categorization relationship. *J. Exp. Psychol. Gen.* 115, 39–57
13 Kruschke, J.K. (1992) ALCOVE: An exemplar-based connectionist model of category learning. *Psychol. Rev.* 99, 22–44
14 Carey, S. (1985) *Conceptual Change in Childhood*, MIT Press
15 Atran, S. (1998) Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behav. Brain Sci.* 21, 547–609
16 Markman, E. (1989) *Naming and Categorization in Children*, MIT Press
17 Bloom, P. (2000) *How Children Learn the Meanings of Words*, MIT Press
18 Murphy, G.L. and Medin, D.L. (1985) The role of theories in conceptual coherence. *Psychol. Rev.* 92, 289–316
19 Wellman, H.M. and Gelman, S.A. (1992) Cognitive development: Foundational theories of core domains. *Annu. Rev. Psychol.* 43, 337–375
20 Gopnik, A. and Meltzoff, A.N. (1997) *Words, Thoughts, and Theories*, MIT Press
21 Rumelhart, D.E. (1980) Schemata: the building blocks of cognition. In *Theoretical Issues in Reading Comprehension* (Spiro, R.J. *et al*., eds), pp. 33–58, Erlbaum
22 Minsky, M. (1975) A framework for representing knowledge. In *The Psychology of Computer Vision* (Winston, P., ed.), pp. 211–277, McGraw-Hill
23 Russell, S.J. and Norvig, P. (2002) *Artificial Intelligence: A Modern Approach*, 2nd edn, Prentice-Hall
24 Gelman, A. *et al*. (2003) *Bayesian Data Analysis*, 2nd edn, Chapman & Hall
25 Kemp, C. *et al.* (2004) Semi-supervised learning with trees. In *Advances in Neural Information Processing Systems 16*, pp. 257–264, MIT Press
26 Kemp, C. *et al.* Learning systems of concepts with an infinite relational model. In *Proc. 21st Natl Conf. Artif. Intell.* (Gil, Y. and Mooney, R.J., eds), AAAI Press (in press)
27 Purves, D. *et al*. (2002) Why we see what we do. *Am. Sci.* 90, 236–243
28 Kording, K.P. and Wolpert, D. (2004) Bayesian integration in sensorimotor learning. *Nature* 427, 244–247
29 Heit, E. (2000) Properties of inductive reasoning. *Psychon. Bull. Rev.* 7, 569–592
30 Shepard, R.N. (1987) Towards a universal law of generalization for psychological science. *Science* 237, 1317–1323
31 Bruner, J.A. *et al*. (1956) *A Study of Thinking*, Wiley
32 Fried, L.S. and Holyoak, K.J. (1984) Induction of category distributions: A framework for classification learning. *J. Exp. Psychol. Learn. Mem. Cogn.* 10, 234–257
33 Carey, S. (1985) Constraints on semantic development. In *Neonate Cognition* (Mehler, J., ed.), pp. 381–398, Erlbaum
34 Tenenbaum, J.B. and Griffiths, T.L. (2001) Generalization, similarity, and Bayesian inference. *Behav. Brain Sci.* 24, 629–641
35 Tenenbaum, J.B. and Xu, F. (2000) Word learning as Bayesian inference. In *Proc. 22nd Annu. Conf. Cogn. Sci. Soc.* (Gleitman, L.R. and Joshi, A.K., eds), pp. 517–522, Erlbaum
36 Xu, F. and Tenenbaum, J.B. (2005) Word learning as Bayesian inference: evidence from preschoolers. In *Proc. 27th Annu. Conf. Cogn. Sci. Soc.* (Bara, B.G. *et al*., eds), pp. 2381–2386, Erlbaum
37 Rosch, E. (1978) Principles of categorization. In *Cognition and Categorization* (Rosch, E. and Lloyd, B.B., eds), pp. 27–48, Erlbaum
38 Clark, E.V. (1987) The principle of contrast: A constraint on language acquisition. In *Mechanisms of Language Acquisition* (MacWhinney, B., ed.), pp. 1–33, Erlbaum
39 Niyogi, S. (2002) Bayesian learning at the syntax-semantics interface. In *Proc. 24th Annu. Conf. Cogn. Sci. Soc.* (Gray, W. and Schunn, C., eds), pp. 697–702, Erlbaum
40 Dowman, M. (2002) Modelling the acquisition of colour words. In *Proc. 15th Austr. Joint Conf. Artif. Intell.: Adv. Artif. Intell.*, pp. 259–271, Springer-Verlag
41 Regier, T. and Gahl, S. (2004) Learning the unlearnable: the role of missing evidence. *Cognition* 93, 147–155
42 Chater, N. and Manning, C.D. (2006) Probabilistic models of language processing and acquisition. *Trends Cogn. Sci.* DOI:10.1016/j.tics.2006.05.006
43 Smith, L.B. *et al*. (2002) Object name learning provides on-the-job training for attention. *Psychol. Sci.* 13, 13–19
44 Kemp, C. and Tenenbaum, J.B. (2003) Theory-based induction. In *Proc. 25th Annu. Conf. Cogn. Sci. Soc.* (Alterman, R. and Kirsh, D., eds), pp. 658–663, Erlbaum
45 Tenenbaum, J.B. and Griffiths, T.L. (2001) The rational basis of representativeness. In *Proc. 23rd Annu. Conf. Cogn. Sci. Soc.* (Moore, J.D. and Stenning, K., eds), pp. 1036–1041, Erlbaum
46 Kemp, C. *et al*. (2004) Learning domain structures. In *Proc. 26th Annu. Conf. Cogn. Sci. Soc.* (Forbus, K. *et al*., eds), pp. 672–678, Erlbaum

47 Shafto, P. and Coley, J.D. (2003) Development of categorization and reasoning in the natural world: novices to experts, naive similarity to ecological knowledge. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 641–649

48 Smith, E.E. *et al*. (1993) Similarity, plausibility, and judgments of probability. *Cognition* 49, 67–96

49 Huelsenbeck, J.P. and Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755

50 Chomsky, N. (1980) *Rules and Representations*, Blackwell

51 Gopnik, A. and Glymour, C. (2002) Causal maps and Bayes nets: A cognitive and computational account of theory-formation. In *The Cognitive Basis of Science*, pp. 117–132, Cambridge University Press

52 Gigerenzer, G. and Todd, P. (1999) *Simple Heuristics that Make us Smart*, Oxford University Press

53 Krynski, T.R. and Tenenbaum, J.B. (2003) The role of causal models in statistical reasoning. In *Proc. 25th Annu. Conf. Cogn. Sci. Soc.* (Alterman, R. and Kirsh, D., eds), pp. 693–698, Erlbaum

54 Rehder, B. (2003) A causal-model theory of conceptual representation and categorization. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 1141–1159

55 Lagnado, D. and Sloman, S.A. (2004) Inside and outside probability judgment. In *Blackwell Handbook of Judgment and Decision Making* (Koehler, D.J. and Harvey, N., eds), pp. 157–176, Blackwell

56 Pearl, J. (2000) *Causality: Models, Reasoning and Inference*, Cambridge University Press

57 Glymour, C. (2001) *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*, MIT Press

58 Ahn, W. and Kalish, C.W. (2000) The role of mechanism beliefs causal reasoning. In *Explanation and Cognition* (Wilson, R. and Keil, F., eds), pp. 199–226, MIT Press

59 Shultz, T.R. (1982) Rules of causal attribution. *Monogr. Soc. Res. Child Dev.* 47 (Serial no. 194)

60 Waldmann, M.R. (1996) Knowledge-based causal induction. In *The Psychology of Learning and Motivation* (Vol. 34), pp. 47–88, Academic Press

61 Tenenbaum, J.B. *et al.* Intuitive theories as grammars for causal inference. In *Causal Learning: Psychology, Philosophy, and Computation* (Gopnik, A. and Schulz, L., eds), Oxford University Press (in press)

62 Griffiths, T.L. and Tenenbaum, J.B. Two proposals for causal grammars. In *Causal Learning: Psychology, Philosophy, and Computation* (Gopnik, A. and Schulz, L., eds), Oxford University Press (in press)

63 Sobel, D.M. *et al*. (2004) Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cogn. Sci.* 28, 303–333

64 Tenenbaum, J.B. and Griffiths, T.L. (2003) Theory-based causal induction. In *Advances in Neural Information Processing Systems 15*, pp. 35–42. MIT Press

65 Griffiths, T.L. *et al*. (2004) Using physical theories to infer hidden causal structure. In *Proc. 26th Annu. Conf. Cogn. Sci. Soc.* (Forbus, K. *et al*., eds), pp. 500–505, Erlbaum

66 Steyvers, M. *et al*. (2003) Inferring causal networks from observations and interventions. *Cogn. Sci.* 27, 453–489

67 De Raedt, L. and Dehaspe, L. (1997) Clausal discovery. *Mach. Learn.* 26, 99–146

68 Friedman, N. *et al*. (1999) Learning probabilistic relational models. In *Proc. 16th Int. Joint Conf. Artif. Intell. (IJCAI)* (Dean, T., ed.), pp. 1300–1309, Morgan Kaufmann

69 Tversky, A. and Kahneman, D. (1974) Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131

70 Shafto, P. *et al*. (2005) Context-sensitive induction. In *Proc. 27th Annu. Conf. Cogn. Sci. Soc.* (Bara, B.G. *et al*., eds), pp. 2003–2008, Erlbaum