Cognitive Development 27 (2012) 455–480

Contents lists available at SciVerse ScienceDirect

Cognitive Development

Theory learning as stochastic search in the language of thought

Tomer D. Ullman[a,*], Noah D. Goodman[b], Joshua B. Tenenbaum[a]

[a] *Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, USA*
[b] *Department of Psychology, Stanford University, Stanford, USA*

ARTICLE INFO

Keywords:
Conceptual change
Bayesian models
Algorithms
Language of thought
Intuitive theories

ABSTRACT

We present an algorithmic model for the development of children's intuitive theories within a hierarchical Bayesian framework, where theories are described as sets of logical laws generated by a probabilistic context-free grammar. We contrast our approach with connectionist and other emergentist approaches to modeling cognitive development. While their subsymbolic representations provide a smooth error surface that supports efficient gradient-based learning, our symbolic representations are better suited to capturing children's intuitive theories but give rise to a harder learning problem, which can only be solved by exploratory search. Our algorithm attempts to discover the theory that best explains a set of observed data by performing stochastic search at two levels of abstraction: an outer loop in the space of theories and an inner loop in the space of explanations or models generated by each theory given a particular dataset. We show that this stochastic search is capable of learning appropriate theories in several everyday domains and discuss its dynamics in the context of empirical studies of children's learning.

> *If a person should say to you "I have toiled and not found", don't believe. If they say "I have not toiled but found", don't believe. If they say "I have toiled and found", believe. - Rabbi Itz'hak, Talmud*

For the Rabbis of old, learning was toil, exhausting work – a lesson many scientists appreciate. Over recent decades, scientists have toiled hard trying to understand learning itself: what children know

[*] Corresponding author. Tel.: +1 617 452 3894.
E-mail address: tomeru@mit.edu (T.D. Ullman).

0885-2014/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cogdev.2012.07.005

when, and how they come to know it. How do children go from sparse fragments of observed data to rich knowledge of the world? From one instance of a rabbit to all rabbits, from occasional stories and explanations about a few animals to an understanding of basic biology, from shiny objects that stick together to a grasp of magnetism – children seem to go far beyond the specific facts of experience to structured interpretations of the world.

What some scientists found in their toil is themselves. It has been argued that children's learning is much like a kind of science, both in terms of the knowledge children create, its form, content, and function, and the means by which they create it. Children organize their knowledge into intuitive theories – abstract coherent frameworks that guide inference and learning within particular domains (Carey, 1985, 2009; Gopnik & Meltzoff, 1997; Murphy & Medin, 1985; Wellman & Gelman, 1992). Such theories allow children to generalize from given evidence to new examples, make predictions and plan effective interventions on the world. Children even construct and revise these intuitive theories using many of the same practices that scientists do (Schulz, 2012b): searching for theories that best explain the data observed, trying to make sense of anomalies, exploring further and even designing new experiments that could produce informative data to resolve theoretical uncertainty, and then revising their hypotheses in light of the new data.

Consider the following concrete example of theory acquisition, which we return to frequently. A child is given a bag of shiny, elongated, hard objects to play with and finds that some pairs seem to exert mysterious forces on each other, pulling or pushing apart when they are brought near enough. These are magnets, but she doesn't know what that would mean. This is her first exposure to the domain. To make matters more interesting, and more like the situation of early scientists exploring the phenomena of magnetism in nature, suppose that all of the objects have an identical metallic appearance, but only some of them are magnetic, and only a subset of those are actually magnets (permanently magnetized). She may initially be confused trying to figure out what interacts with what, but like a scientist developing a first theory, after enough exploration and experimentation, she might start to sort the objects into groups based on similar behaviors or similar functional properties. She might initially distinguish two groups, the magnetic objects (which can interact with each other) and the nonmagnetic ones (which do not interact). Perhaps then she will move on to subtler distinctions, noticing that this very simple theory doesn't predict everything she observes. She could distinguish three groups, separating the permanent magnets from the rest of the magnetic objects as well as from the nonmagnetic objects and recognizing that there will only be an interaction if at least one of the two magnetic objects brought together is a permanent magnet. With more time to think and more careful observation, she might even come to discover the existence of magnetic poles and the laws by which they attract or repel when two magnets are brought into contact. These are but three of a large number of potential theories, varying in complexity and power, that a child could entertain to explain her observations and make predictions about unseen interactions in this domain.

Our goal here is to explore computational models of how children might acquire and revise an intuitive theory such as this on the basis of domain experience. Any model of learning must address two kinds of questions: what and how? Which representations can capture the form and content of what the learner comes to know, and which principles or mechanisms can explain how the learner comes to know it, moving from one state of knowledge to another in response to observed data? Here we address the 'how' question. We build on much recent work addressing the 'what' question, which proposes to represent the content of children's intuitive theories as probabilistic generative models defined over hierarchies of structured symbolic representations (Kemp, Goodman, & Tenenbaum, 2008b; Tenenbaum, Griffiths, & Kemp, 2006; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). Previously the 'how' question has been addressed only at a very high level of abstraction, if at all: The principles of Bayesian inference explain how an ideal learner can successfully identify an appropriate theory, based on maximizing the posterior probability of a theory given data (as given by Bayes' rule). But Bayes' rule says nothing about the processes by which a learner could construct such a theory or revise it in light of evidence. Here our goal is to address the 'how' of theory construction and revision at a more mechanistic, process level, exploring cognitively realistic learning algorithms. Put in terms of Marr's (1982) three levels of analysis, previous Bayesian accounts of theory acquisition have concentrated on the level of computational theory, while here we move to the algorithmic level of analysis, with the aim of giving a more plausible, practical and experimentally fertile view of developmental processes within the Bayesian paradigm.

Our work here aims to explain two challenges of theory acquisition in algorithmic terms. First is the problem of making learning work – getting the world right as reliably as children do. As any scientist knows, the 'how' of theory construction and revision is nontrivial. The process is often slow, painful, a matter of starts and stops, random fits and bursts, missteps and retreats, punctuated by occasional moments of great insight, progress and satisfaction – the flashes of 'Aha!' and 'Eureka!'. And as any parent knows, children's cognitive development often seems to have much the same character. Different children make their way to adult-like intuitive theories at very different paces. Transitions between candidate theories often appear somewhat random and unpredictable at a local level, prone to backtracking or "two steps forward, one step back" (Siegler & Chen, 1998). Yet in core domains of knowledge, and over long time scales, theory acquisition is remarkably successful and consistent. Different children (at least within a common cultural context of shared experience) tend to converge on the same knowledge structures, knowledge that is much closer to a veridical account of the world's causal structure than the infant's starting point, and they follow predictable trajectories along the way (Carey, 2009; Gopnik & Meltzoff, 1997; Wellman, Fang, & Peterson, 2011).

Our first contribution is an existence proof to show how this kind of learning could work – a model of how a search process with slow, fitful and often frustrating stochastic dynamics can still reliably get the world right, in part because of these dynamics, not simply in spite of them. The process may not look very algorithmic, in the sense of familiar deterministic algorithms such as those for long division, finding square roots, or sorting a list, or what cognitive scientists typically think of as a "learning algorithm", such as the backpropagation algorithm for training neural networks. Our model is based on a *Monte Carlo* algorithm, which makes a series of randomized (but not entirely random) choices as part of its execution. These choices guide how the learner explores the space of theories to find those that best explain the observed data – influenced by, but not determined by, the data and the learner's current knowledge state. We show that such a Monte Carlo exploratory search yields learning results and dynamics qualitatively similar to what we see in children's theory construction, for several illustrative cases.

Our second challenge is to address what could be called the "hard problem" of theory learning – learning a system of concepts that cannot be simply expressed as functions of observable sense data or previously available concepts – knowledge that is not simply an extension or addition to what was known before but that represents a fundamentally new way to think. Developmental psychologists, most notably Carey (2009), have long viewed this problem of conceptual change or theory change as one of the central explanatory challenges in cognitive development. To illustrate, consider the concepts of "magnet" or "magnetic object" or "magnetic pole" in our scenario above, for a child first learning about them. There is no way to observe an object on its own and decide if it falls under any of these concepts. There is no way to define or describe either "magnet" or "magnetic object" in purely sensory terms (terms that do not themselves refer to the laws and concepts of magnetism), nor to tell the difference between a "north" and a "south" magnetic pole from perception alone. How then could these notions arise? They could be introduced in the context of explanatory laws in a theory of magnetism, such as "Two objects will interact if both are magnetic and at least one is a magnet," or "Magnets have two poles, one of each type, and opposite types attract while like types repel." If we could independently identify the magnets and the magnetic objects, or the two poles of each magnetic object and their types, these laws would generate predictions that could be tested on observable data. But only in virtue of these laws' predictions can magnets, magnetic objects, or magnetic poles even be identified or made meaningful. And how could one even formulate or understand one of these laws without already having the relevant concepts?

Theory learning thus presents children with a difficult joint inference task – a "chicken-and-egg" problem – of discovering two kinds of new knowledge, new concepts and new laws, which can only be made sense of in terms of each other. The laws are defined over the concepts, but the concepts only get their meaning from the roles they play in the laws. If learners do not begin with either the appropriate concepts or the appropriate laws, how can they end up acquiring both successfully? This is also essentially the challenge that philosophers have long studied of grounding meaning in *conceptual role* or *inferential role* semantics (Block, 1986; Field, 1977; Fodor & Lepore, 1991; Harman, 1975, 1982). Traditional approaches to concept learning in psychology do not address this problem, nor do they even attempt to (Bruner, Goodnow, & Austin, 1956; Rogers & McClelland, 2004; Smith & Medin, 1981). The elusiveness of a satisfying solution has led some scholars, most famously Fodor, to a radical skepticism on the prospects for learning genuinely new concepts and a view that most concepts must be innate in some nontrivial way (Fodor, 1975, 1980). Carey (2009) has proposed a set of informal "bootstrapping" mechanisms for how human learners could solve this problem, but no formal model of bootstrapping exists for theory learning or concept learning in the context of acquiring novel theories.

We argue that the chicken-and-egg problem can be solved by a rational learner but must be addressed in algorithmic terms to be truly satisfying: a purely computational-level analysis will always fail for the Fodorian skeptic, and will fail to make contact with the crux of the bootstrapping problem as Carey (2009) frames it, since for the ideal learner the entire space of possible theories, laws and concepts is in a sense already available from the start. An algorithmic implementation of that same ideal learning process can, however, introduce genuinely new concepts and laws in response to observed data. It can provide a concrete solution to the problem of how new concepts can be learned and can acquire meaning in a theory of inferential role semantics. Specifically, we show how a Monte Carlo search process defined over a hierarchically structured Bayesian model can effectively introduce new concepts as blank placeholders in the context of positing a new candidate explanatory law or extending an existing law. The new concept is not expressed in terms of pre-existing concepts or observable data; rather it is posited as part of a candidate explanation, together with pre-existing concepts, for observed data. In testing the candidate law's explanatory power, the new concepts are given a concrete interpretation specifying which entities they are most likely to apply to, assuming the law holds. If the new or modified law turns out to be useful – that is, if it leads to an improved account of the learner's observations, relative to their current theory – the law will tend to be retained, and with it, the new concept and its most likely concrete grounding.

The rest of the article is organized as follows. We first present a nontechnical overview of the "what" and "how" of our approach to theory learning and contrast it with the most well-known alternatives for modeling cognitive development based on connectionism and other emergentist paradigms. We then describe our approach more technically, culminating in a Markov Chain Monte Carlo (MCMC) search algorithm for exploring the space of candidate theories based on proposing random changes to a theory and accepting probabilistically those changes that tend to improve the theory. We highlight two features that make the dynamics of learning more efficient and reliable, as well as more cognitively plausible – a prior that proposes new theoretical laws drawn from *law templates*, biasing the search toward laws that express canonical patterns of explanation useful across many domains, and a process of *annealing* the search that reduces the amount of random exploration over time. We study the algorithm's behavior on two case studies of theory learning inspired by everyday cognitive domains – the taxonomic organization of object categories and properties and a simplified version of magnetism. Finally, we explore the dynamics of learning that arise from the interaction between computational-level and algorithmic-level considerations – how theories change both as a function of the quantity and quality of the learner's observations and as a function of the time course of the annealing-guided search process, which suggests promising directions for future experimental research on children's learning.

1. A nontechnical overview

A proposal for *what* children learn and a proposal for *how* they learn it may be logically independent in some sense, but the two are mutually constraining. Richer, more structured accounts of the form and content of children's knowledge tend to pose harder learning challenges, requiring learning algorithms that are more sophisticated and more costly to execute. As we explain, our focus on explaining the origins of children's intuitive theories leads us to adopt relatively rich abstract forms of knowledge representations, compared to alternative approaches to modeling cognitive development, such as connectionism. This leaves us with relatively harder learning challenges – connectionists might argue, prohibitively large. But we see these challenges as inevitable. Sooner or later, computational models of development must face them. Perhaps for the first time, we can now begin to see what

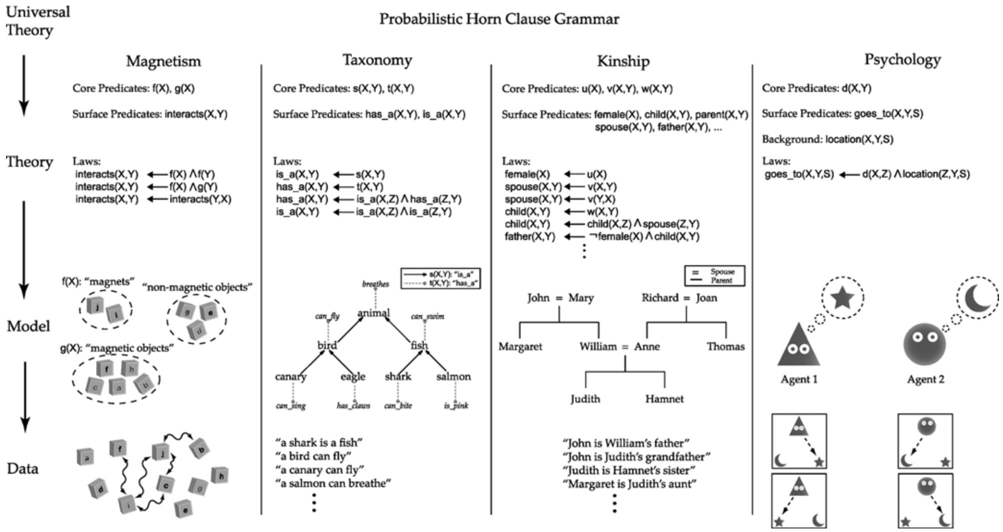
Universal Theory
Probabilistic Horn Clause Grammar
Theory
Model
Data
Magnetism
Core Predicates: f(X), g(X)
Surface Predicates: interacts(X,Y)
Laws:
interacts(X,Y) ← f(X) ∧ f(Y)
interacts(X,Y) ← f(X) ∧ g(Y)
interacts(X,Y) ← interacts(Y,X)
f(X): "magnets"
"non-magnetic objects"
g(X): "magnetic objects"
Taxonomy
Core Predicates: s(X,Y), t(X,Y)
Surface Predicates: has_a(X,Y), is_a(X,Y)
Laws:
is_a(X,Y) ← s(X,Y)
has_a(X,Y) ← t(X,Y)
has_a(X,Y) ← is_a(X,Z) ∧ has_a(Z,Y)
is_a(X,Y) ← is_a(X,Z) ∧ is_a(Z,Y)
s(X,Y): "is_a"
t(X,Y): "has_a"
breathes
animal
can_fly
can_swim
bird
fish
canary
eagle
shark
salmon
can_sing
has_claws
can_bite
is_pink
"a shark is a fish"
"a bird can fly"
"a canary can fly"
"a salmon can breathe"
Kinship
Core Predicates: u(X), v(X,Y), w(X,Y)
Surface Predicates: female(X), child(X,Y), parent(X,Y) spouse(X,Y), father(X,Y), ...
Laws:
female(X) ← u(X)
spouse(X,Y) ← v(X,Y)
spouse(X,Y) ← v(Y,X)
child(X,Y) ← w(X,Y)
child(X,Y) ← child(X,Z) ∧ spouse(Z,Y)
father(X,Y) ← ¬female(X) ∧ child(X,Y)
= Spouse
— Parent
John = Mary
Richard = Joan
Margaret
William = Anne
Thomas
Judith
Hamnet
"John is William's father"
"John is Judith's grandfather"
"Judith is Hamnet's sister"
"Margaret is Judith's aunt"
Psychology
Core Predicates: d(X,Y)
Surface Predicates: goes_to(X,Y,S)
Background: location(X,Y,S)
Laws:
goes_to(X,Y,S) ← d(X,Z) ∧ location(Z,Y,S)
Agent 1
Agent 2

Fig. 1. A hierarchical Bayesian framework for theory acquisition. Each level generates the space of possibilities for the level below, providing constraints for inference. Four examples of possible domain theories are given in separate columns, while the rows correspond to different levels of the hierarchy.

their solution might look like, by bringing together recent ideas for modeling the form and content of theories as probabilistic generative models over hierarchies of symbolic representations (Goodman, Ullman, & Tenenbaum, 2011; Katz, Goodman, Kersting, Kemp, & Tenenbaum, 2008; Kemp, Goodman, & Tenenbaum, 2008a) with tools for modeling the dynamics of learning as exploratory search based on stochastic Monte Carlo algorithms.

1.1. The 'What': Modeling the form and content of children's theories as hierarchical probabilistic models over structured representations

As a form of abstract knowledge, an intuitive theory is similar to the grammar of a language (Tenenbaum, Griffiths, & Niyogi, 2007): The concepts and laws of the theory can be used to generate explanations and predictions for an infinite (though constrained) set of phenomena in the theory's domain. We follow a long tradition in cognitive science and artificial intelligence of representing such knowledge in terms of compositional symbol systems, specifically predicate logic that can express a wide range of possible laws and concepts (Fodor & Pylyshyn, 1988; Fodor, 1975; Russell & Norvig, 2009). Embedding this symbolic description language in a hierarchical probabilistic generative model lets us bring to bear the powerful inductive learning machinery of Bayesian inference, at multiple levels of abstraction (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Tenenbaum et al., 2011).

Fig. 1 illustrates this framework. We assume a domain of cognition is given, comprising one or more systems of entities and their relations, each of which gives rise to some observed data. The learner's task is to build a theory of the domain – a set of abstract concepts and explanatory laws that explain the observed data for each system in that domain. The learner is assumed to have a hypothesis space of possible theories generated by (and constrained by) some "Universal Theory". We formalize this Universal Theory as a probabilistic generative grammar, essentially a probabilistic version of a language of thought (Fodor, 1975). Within this universal language, the learner constructs a specific theory that can be thought of as a more specific language for explaining the phenomena of the given domain.

In principle, an ideal learner should consider all possible theories expressible in the language of thought and weigh them against each other in light of observed evidence. In practice, there are infinitely many candidate theories and it will be impossible to explicitly consider even a small fraction of them. Explaining how a learner proposes specific candidate theories for evaluation is a task for our algorithmic-level account to follow.

Candidate theories are evaluated using Bayes' rule to assess how likely they are to have generated the observed data. Bayes' rule scores theories based on the product of their prior probabilities and their likelihoods. The prior reflects the probability of generating the laws and concepts of a theory a priori from the generative grammar, independent of any data to be explained. The likelihood measures the probability of generating the observed data given the theory, independent of the theory's plausibility. Occam's-razor-like considerations emerge naturally from a Bayesian analysis. The prior will be highest for the simplest theories, whose laws can be generated with the fewest number of a priori stipulations, while the likelihood will be highest for theories whose laws allow a domain to be described accurately and compactly, generating the observed data with a spare set of minimal facts.

The fit of a theory to data cannot be evaluated directly. Its laws express the abstract principles underlying a domain but no specific expectations about what is true or false. One level below the theory in the hierarchical framework, the learner posits a *logical model* of each observed system in the domain. The logical model ("model" for short) specifies what is true of the entities in a particular system in ways consistent with and constrained by the theory's abstract laws. Each model can be thought of as one particular concrete instantiation of the abstract theory. It generates a probability distribution over possible observations for the corresponding system, and it can be scored directly in terms of how well those predictions fit the actual data observed.

As a concrete example of this framework, consider again the child learning about the domain of magnetism. She might begin by playing with a few pieces of metal and notice that some of the objects interact, exerting strange pulling or pushing forces on each other. She could describe the data directly, as "Object *a* interacts with object *j*", "Object *i* interacts with object *j*", and so on. Or she could form a simple theory, in terms of abstract concepts such as *magnet*, *magnetic object* and *non-magnetic* object, and laws such as "*Magnets* interact with other *magnets*", "*Magnets* interact with *magnetic objects*", and "Interactions are symmetric". It is important to note that terms like magnet convey no actual information about the object – they are simply labels. Systems in this domain correspond to specific subsets of objects, such as the set of objects *a*, . . ., *i* in Fig. 1. A model of a system specifies the minimal facts needed to apply the abstract theory to the system – in this case which objects are magnetic, which are magnets, and which are non-magnetic. From these core facts the laws of the theory determine all other facts. In our example, this means all pairwise interactions between objects – objects *i* and *j*, being magnets, should interact, but *i* and *e* should not, because the latter is non-magnetic. The facts generate the data observed by the learner via a noisy sampling process, e.g., observing a random subset of the object pairs that interact and occasionally misperceiving an object's identity or the nature of an interaction.

While the abstract concepts in this simplified magnetism theory are attributes of objects, more complex relations are possible. Consider for example a theory of taxonomy, as in Collins and Quillian's (1969) classic model of semantic memory as an inheritance hierarchy. Here the abstract concepts are *is_a* relations between categories and *has_a* relations between categories and properties. The theory underlying taxonomy has two basic laws: "The *is_a* relation is transitive" and "The *has_a* relation inherits down *is_a* relations" (laws 3 and 4 in the "Taxonomy" column of Fig. 1). A system consists of a specific set of categories and properties, such as *salmon*, *eagle*, *breathes*, *can_fly*, and so on. A model specifies the minimal *is_a* and *has_a* relations, typically corresponding to a tree of *is_a* relations between categories with properties attached by *has_a* relations at the broadest category they hold for (e.g., "A canary is a bird"). The laws then determine that properties inherit down chains of *is_a* relations to generate other observable facts (e.g., "A canary can breathe").

The analogy between learning a theory for a domain and learning a grammar for a natural language thus extends down through all levels of Fig. 1 hierarchy. A logical model for a system of observed entities and relations can be thought of as a parse of that system under the grammar of the theory, just as the theory itself can be thought of as a parse of a whole domain under the grammar of the universal theory. In our hierarchical Bayesian framework, theory learning is the problem of searching jointly for the theory of a domain and models of each observed system in that domain that together best parse all the observed data.[1]

Previous applications of grammar-based hierarchical Bayesian models have shown how, given sufficient evidence and a suitable theory grammar, an ideal Bayesian learner can identify appropriate theories in domains such as causality (Goodman et al., 2011; Griffiths et al., 2010), kinship and other social structures (Kemp et al., 2008a), and intuitive biology (Tenenbaum et al., 2007). While our focus here is the algorithmic level – the dynamics of how learners can search through a space of theories – we have found that endowing our theory grammars with one innovation greatly improves their algorithmic tractability. We make the grammar more likely to generate theories with useful laws by equipping it with law templates, or forms of laws that capture canonical patterns of coherent explanation arising in many domains. For example, law templates might suggest explanations for when an observed relation $r(X,Y)$ holds between entities X and Y (e.g., X attracts Y, X activates Y, X has Y) in terms of latent attributes of the objects, $f(X)$ and $g(Y)$, or in terms of some other relation $s(X,Y)$ that holds between them, or some combination thereof – perhaps $r(X,Y)$ holds if $f(X)$ and $s(X,Y)$ are both true. Explanatory chains introducing novel objects are also included among the templates. Perhaps $r(X,Y)$ holds if there exists a Z such that $s(X,Z)$ and $s(Z,Y)$ hold. As we explain, making these templates explicit in the grammar makes learning both more cognitively plausible and much faster.

The most familiar computational alternative to structured Bayesian accounts of cognitive development are connectionist models and other *emergentist* approaches (McClelland et al., 2010). Instead of representing children's abstract knowledge in terms of explicit symbol systems, these approaches attribute abstract knowledge to children only implicitly as an 'emergent' phenomenon that arises in a graded fashion from interactions among more concrete, lower-level non-symbolic elements – often inspired loosely by neuroscience. *Dynamical systems* models view the nervous system as a complex adaptive system evolving on multiple timescales, with emergent behavior in its dynamics. *Connectionist* models view children's knowledge as embedded in the strengths of connections between many neuron-like processing units and treat development as the tuning of these strengths via some experience-dependent adjustment rule. Connectionists typically deny that the basic units of traditional knowledge representation – objects, concepts, predicates, relations, propositions, rules and other symbolic abstractions – are appropriate for characterizing children's understanding of the world, except insofar as they emerge as approximate higher-level descriptions for the behavior dictated by a network's weights.

While emergentist models have been well-received in some areas of development, such as the study of motor and action systems (McClelland et al., 2010), emergentist models of the structure and origins of abstract knowledge (Rogers & McClelland, 2004) have not been widely embraced by developmentalists (Carey, 2009; Gopnik & Meltzoff, 1997). There is every reason to believe that explicit symbolic structure is just as important for children's intuitive theories as for scientists' more formal theories – that children, like scientists, cannot adequately represent the underlying structure of a domain such as physics, psychology or biology simply with a matrix of weights in a network that maps a given set of inputs to a given set of outputs. Children require explicit representations of abstract concepts and laws in order to talk about their knowledge in natural language and to change and grow their knowledge through talking with others; to reason causally in order to plan for the future, explain the past, or imagine hypothetical situations; to apply their knowledge in novel settings to solve problems that they have never before encountered; and to compose abstractions recursively.

Despite these limitations, connectionist models have been appealing to developmentalists who emphasize the processes and dynamics of learning more than the nature of children's knowledge representations (McClelland et al., 2010; Shultz, 2003). This appeal may come from the fact that when we turn from the 'what' to the 'how' of children's learning, connectionist models have a decided advantage. Learning in connectionist systems appears much better suited to practical algorithmic formulation, and much more tractable, relative to structured probabilistic models or any explicitly

[1] The idea of hierarchical Bayesian grammar induction, where the prior on grammars is itself generated by a grammar (or "grammar grammar"), dates back at least to the seminal work of Feldman and colleagues (Feldman et al., 1969).

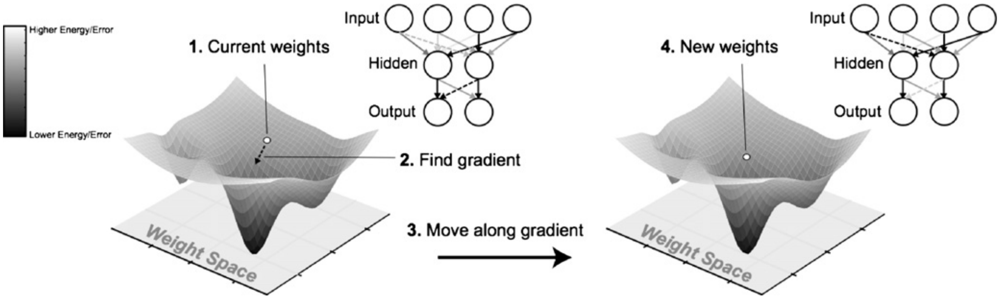

Fig. 2. A hypothetical neural network and a weight space spanning the possible values of two particular connections. Steps 1–4 show the sequence of a learning algorithm in such a space – the calculation of a gradient and the move to a lower point, corresponding to a shift in the network's connection weights and a smaller error on the output.

symbolic approach. As we explain below, making the 'how' of learning plausible and tractable may be the biggest challenge facing the structured probabilistic approach.

1.2. *The 'How': Modeling the dynamics of children's theory learning as stochastic (Monte Carlo) exploratory search*

It is helpful to imagine the problem children face in learning as that of moving over a "knowledge landscape", where each point represents a possible state of knowledge and the height of that point reflects the value of that knowledge-state – how well it allows a child to explain, predict, and act on the world. Such a picture highlights differences between our approach to cognitive development and connectionist and emergentist alternatives, and the much more serious 'how' challenge that confronts structured probabilistic models.

Viewed in landscape terms (Fig. 2), connectionist models typically posit that children's knowledge landscape is continuous and smooth, and this matters greatly for the mechanisms and dynamics of learning. Learning consists of traversing a high-dimensional real-valued "weight space", where each dimension corresponds to the strength of one connection in a neural network. Fig. 2 depicts only a two-dimensional slice of the much higher dimensional landscape corresponding to a three-layer network. The height of the landscape assigned to each point in weight space – each joint setting of all the network's weights – measures how well the network explains observed data in terms of an error or energy function, such as a sum-of-squared-error expression. The topology of these landscapes is simple and uniform. At any point of the space, one can always move along any dimension independently of every other, and changing one parameter has no effect on any other. The geometry is also straightforward. Neighboring states, separated by small changes in the weights or parameters, typically yield networks with very similar input-output functionality. Thus a small move in any direction typically leads to only a small rise or fall in the error or energy function.

This geometry directly translates into the dynamics of learning. Weight-adjustment rules (e.g., the Hebb rule, the Delta Rule, Backpropagation; Mcclelland & Rumelhart, 1986) can be seen as implementing gradient descent – descending the error or energy landscape by taking small steps along the steepest direction. It can be proven that this dynamic reliably takes the network to a local minimum of error, or a locally best fitting state of knowledge. In certain cases, particularly of interest in contemporary machine learning systems (Bishop, 2006), the error landscape can be designed to have a geometric property known as convexity, ensuring that any local minimum is also a global minimum and thus that the best possible learning end-state can be achieved using only local weight-adjustment rules based on gradient descent. Thus learning becomes essentially a matter of "rolling downhill", and is just as simple. Even when there are multiple distinct local minima, connectionist learning can still draw on a powerful toolkit of optimization methods that exploit the fact that the landscape is continuous and smooth to make learning relatively fast, reliable and automatic.

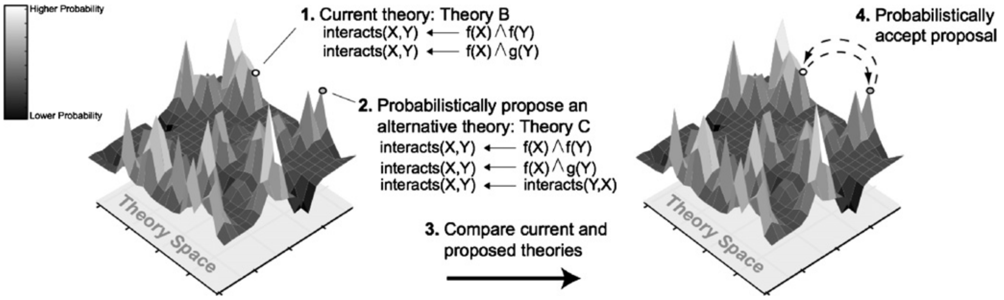
Higher Probability
Lower Probability
1. Current theory: Theory B
interacts(X,Y) ← f(X) ∧ f(Y)
interacts(X,Y) ← f(X) ∧ g(Y)
2. Probabilistically propose an alternative theory: Theory C
interacts(X,Y) ← f(X) ∧ f(Y)
interacts(X,Y) ← f(X) ∧ g(Y)
interacts(X,Y) ← interacts(Y,X)
3. Compare current and proposed theories
4. Probabilistically accept proposal
Theory Space
Theory Space

Fig. 3. Schematic representation of the learning landscape within the domain of simple magnetism. Steps 1–4 illustrate the algorithmic process in this framework. The actual space of theories is discrete, multidimensional and not necessarily locally connected.

Now consider the landscape of theory learning from the perspective of our structured Bayesian approach, and it becomes clear how much more difficult the problem is (Fig. 3). Each point on the landscape now represents a candidate domain theory expressed in terms of one or more laws in first-order logic and one or more abstract concepts indicated by a blank predicate, e.g., $f(X)$, $g(X)$. Two possibilities for a simple theory of magnetism are shown, labeled Theory B and Theory C. The height of the surface at a given point represents how well the theory is supported by the observed data, i.e., the Bayesian posterior probability. (In contrast to Fig. 2, where “lower is better”, here “higher is better”, and the goal is to seek out maxima of the landscape, not minima.) Unlike the weight space shown in Fig. 2, this portrait of a “theory space” as two-dimensional is only metaphorical; it is not simply a lower-dimensional slice of a higher-dimensional space. The space of theories in a language of thought is infinite and combinatorially structured with a neighborhood structure impossible to visualize faithfully on a page.

We can imagine an ideal Bayesian learner who computes the full posterior probability distribution over all possible theories, that is, who grasps this entire landscape and assesses its height at all points in parallel, conditioned on any given observed data set. But this is clearly unrealistic as a starting point for algorithmic accounts of children’s learning, or any practical learning system with limited processing resources. Intuition suggests that children may simultaneously consider no more than a handful of candidate theories in their active thought, and developmentalists typically speak of the child’s current theory as if, as in connectionist models, the learner’s knowledge state corresponds to just a single point on the landscape rather than the whole surface or posterior distribution. For the ideal learner, the entire hypothesis space is already defined and the learner’s task is merely to reshuffle probability over that space in response to evidence. However, the actual child must construct an abstract theory, piece by piece, generalizing from experience.

Considering how a learner could move around this landscape in search of the best theory, we see that most of the appealing properties of connectionist knowledge landscapes – the features that support efficient learning algorithms – are not present here. The geometry of the landscape is far from smooth. A small change in one of the concepts or laws of a theory will often lead to a drastic rise or fall in its plausibility, leading to a proliferation of isolated local maxima. There is typically no local information (such as a gradient) diagnostic of the most valuable directions in which to move. The landscape is even more irregular in ways that are not easily visualized. There is no uniform topology or neighborhood structure. The number and nature of variants that can be proposed by making local changes to the learner’s current hypothesis vary greatly over the space, depending on the form of that hypothesis. Often changing one aspect of a theory requires others to be changed simultaneously in order to preserve coherence. For instance, if we posit a new abstract concept, such as the notion of a *magnet*, or if we remove a conceptual distinction (such as the distinction between *magnets* and *magnetic objects*), one or more laws of the theory will need to be added, removed or redefined. Artificial intelligence has a long history of treating learning in terms of search through a discrete space of symbolic descriptions, and a wide variety of search algorithms have been proposed to solve problems such as rule discovery, concept learning and generalization, scientific law discovery, and causal learning (Bradshaw, Langley, & Simon, 1983; Mitchell, 1982; Newell & Simon, 1976; Pearl, 2000; Spirtes, Glymour, & Scheines, 2001). For some of these problems, there exist systematic search algorithms that can be as fast and reliable as the gradient-based optimization methods used in connectionist learning (Mitchell, 1982; Pearl, 2000; Spirtes et al., 2001). But for problems like scientific discovery (Bradshaw et al., 1983), or our formulation of children's theory learning, the best known search algorithms are not like this. Much like child learners, we suggest, these algorithms are slow, unreliable, and unsystematic (indeed often random), but with enough patience they can be expected to converge on veridical theories.

The specific search algorithm we describe is based on widely used methods in statistics and artificial intelligence for approximating intractable Bayesian inferences, known as Markov Chain Monte Carlo (MCMC). MCMC algorithms have been proposed as models for the short-timescale dynamics of perceptual inferences in the brain (Gershman, Vul, & Tenenbaum, 2009; Moreno-Bote, Knill, & Pouget, 2011; Sundareswara & Schrater, 2008), but they are also well-suited to understanding the much longer-term dynamics of learning. The MCMC algorithm addresses the two main challenges we identified earlier – explaining how children can reliably converge on veridical theories, given their constrained cognitive resources and a learning dynamic that often appears more random than systematic, and explaining how children can solve the hard "chicken-and-egg" inference problem of jointly learning new concepts and new laws defined in terms of those concepts.

The heart of MCMC theory learning is an iterative loop of several basic steps, shown in Fig. 3. The learner begins at some point in the theory landscape (e.g., theory B or C in Fig. 3). The learner then proposes a possible move to a different theory, based on modifying the current theory's form – adding/deleting a law or set of laws, changing parts of a law or introducing a new concept, and so on. The proposed and current theories are compared based on evaluating (approximately) how well they explain the observed data (i.e., comparing the relative heights of these two points on the theory landscape). If the proposed theory scores higher, the learner accepts it and moves to this new location. If the proposal scores lower, the learner may still accept it or reject it (staying at the same location), with probability proportional to the relative scores of the two theories. These steps are then repeated with a new proposal based on the new current location.

From the standpoint of MCMC, randomness is not a problem but rather an essential tool for exploring the theory landscape. Because MCMC algorithms consider only one hypothesis at a time and propose local modifications to it, and there are no generally available signals (analogous to the error gradient in connectionist learning) for how to choose the best modification of the current hypothesis from an infinite number of possible variations, the best learners can do is to explore variant theories chosen in a randomized but hopefully intelligent fashion. Our algorithm proposes variants to the current hypothesis by replacing a randomly chosen part of the theory with another random draw from the probabilistic generative grammar for theories (that is, the prior theories). This process could in principle propose any theory as a variant on any other, but it is naturally biased toward candidates most similar to the current hypothesis, as well as those that are a priori simpler and more readily generated by the grammar's templates for coherent laws. The use of law templates is crucial in focusing the random proposal mechanism on the most promising candidates. Without templates, all laws proposed could still have been generated from a more general grammar, but they would be much less likely a priori; learners would end up wasting most of their computational effort considering simple but useless candidate laws. The templates make it likely that any random proposal is at least a plausibly useful explanation, not just a syntactically well-formed expression in the language of thought.

The decision of whether to accept or reject a proposed theory change is also made in a randomized but intelligently biased fashion. If a proposed change improves the theory's account of the data, it is always accepted, but sometimes a change that makes the theory worse could also be accepted. This probabilistic acceptance rule helps keep the leaner from becoming trapped for too long in poor local maxima of the theory landscape (Gilks & Spiegelhalter, 1996).

Although we use MCMC as a search algorithm, aiming to find the best theory, the algorithm's proper function is not to find a single optimal theory but rather to visit all theories with probability proportional to their posterior probability. We can interpolate between MCMC as a posterior inference technique and MCMC as a search algorithm by *annealing* – or starting with more stochastic (or noisy) search moves and "lowering the temperature," making the search more deterministic over time (Kirkpatrick, Gelatt, & Vecchi, 1983; Spall, 2003). This greatly improves convergence to the true theory. Such an algorithm can begin with little or no knowledge of a domain and, given enough time and sufficient data, reliably converge on the correct theory or at least some approximation thereof, corresponding to a small set of abstract predicates and laws.

Annealing is also responsible for giving the MCMC search algorithm some of its psychologically plausible dynamics. It gives rise to an early high-temperature exploration period characterized by a large number of proposed theories, most of which are far from veridical. As we see in young children, new theories are quick to be adopted and just as quick to be discarded. As the temperature is decreased, partially correct theories become more entrenched, it becomes rarer for learners to propose and accept large changes to their theories, and the variance between different theory learners decreases. As with older children, rational learners at the later stages of an annealed MCMC search tend to mostly agree on what is right, even if their theories are not perfect. Without annealing, MCMC dynamics at a constant temperature could result in a learner who is either too conservative (at low temperature) or too aggressive (at high temperature) in pursuing new hypotheses – that is, a learner who is prone to converge too early on a less-than-ideal theory or to never converge at all.

On average, learners are consistently improving. On average, they are improving gradually. But individually, learners often get worse before they get better. Individually, they adopt theories in discrete jumps, signifying moments of fortuitous discovery. Such dynamics on the level of the individual learner are more in line with discovery processes in science and childhood than are the smoother dynamics of gradient descent on a typical connectionist energy landscape. Critics might reasonably complain that MCMC methods are slow and unreliable by comparison. But theory construction is a difficult, time-consuming, painful and frustrating business – in both science and children's cognition. We can no more expect the dynamics of children's learning to follow the much tamer dynamics of gradient learning algorithms than we could expect to replace scientists with a gradient-based learning machine and see the discoveries of new concepts and new scientific laws emerging automatically.[2] Currently we have no good alternative to symbolic representational machinery for capturing intuitive theories and no good alternative to stochastic search algorithms for finding good points in the landscape of these symbolic theories.

What of the "hard problem of theory learning," the challenge of jointly learning new laws and new concepts defined only in terms of each other? Our MCMC search unfolds in parallel over two levels of abstraction – an outer loop in the space of theories, defined by sets of abstract laws; and an inner loop in the space of explanations or models generated by the theory for a particular domain of data, defined by groundings of the theory's concepts on the specific entities of the domain. This two-level search lets us address the "chicken and egg" challenge by first proposing new laws or changes to existing laws of a theory in the outer search loop; these new laws can posit novel but 'blank' concepts of a certain form, whose meaning is then filled in the most plausible way on the next inner search loop. For example, the algorithm may posit a new rule never before considered, that objects of type f interact with objects type g, without yet specifying what these concepts mean; they are represented with blank predicates $f(X)$ and $g(X)$. The inner loop then searches for a reasonable assignment of objects to these classes – values for $f(X)$ and $g(X)$, for each object X – grounding them as magnets and magnetic objects, for example. If this law proves useful, it is likely to persist in the MCMC dynamics, and with it, the novel concepts that began as blank symbols f and g but have now effectively become what we call "magnets" and "magnetic objects."

In sum, we see many reasons to think that stochastic search in a language of thought, with candidate theories generated by a probabilistic generative grammar and scored against observations in a hierarchical Bayesian framework, provides a better account of children's theory acquisition than

[2] Not all connectionist architectures and learning procedures are confined to gradient-based methods operating on fixed parametric architectures. In particular the constructivist neural networks explored by Shultz (2003) and colleagues are motivated by considerations similar to ours, aiming to capture the dynamics of children's discovery with learning rules that implement exploratory search. These models are still limited in their representational power, however. They can only express knowledge whose form and content fits into the connections of a neural network and not the abstract concepts and laws that constitute an intuitive theory. We thus favor the more explicitly symbolic approach described here.

alternative computational paradigms for modeling development, such as connectionism. Yet there are also major gaps. Scientists and children alike are smarter, more active, more deliberate and driven explorers of both their theories and their experiences and experiments than are our MCMC algorithms (Schulz, 2012a). We now turn to a more technical treatment of our model but we return to these gaps in the general discussion below.

2. Formal framework

The hierarchical picture of knowledge shown in Fig. 1 provides the backbone for a multilevel probabilistic generative model. Conditional probability distributions that link knowledge at different levels of abstraction, supporting inference at any level(s) conditioned on knowledge or observations at other levels. For instance, given a domain theory *T* and a set of noisy, sparse observations *D*, a learner can infer the most likely model *M* and use that knowledge to predict other facts not yet directly observed (Katz et al., 2008; Kemp et al., 2008a). The theory *T* sets the hypothesis space and priors for the model *M*, while the data *D* determine the model's likelihood, and Bayes' rule combines these two factors into a model's posterior probability score,

$$P(M|D, T) \propto P(D|M)P(M|T). \quad (1)$$

If the theory *T* is unknown, the learner considers a hypothesis space of candidate theories generated by the higher-level universal theory (*U*) grammar. *U* defines a prior distribution over the space of possible theories, $P(T|U)$, and again the data *D* determine a likelihood function, with Bayes' rule assigning a posterior probability score to each theory,

$$P(T|D, U) \propto P(D|T)P(T|U). \quad (2)$$

Bayes' rule here captures the intuition of Occam's razor, that the theory which best explains a data set, i.e., has highest posterior probability $P(T|D, U)$, should balance between fitting the data well, as measured by the likelihood $P(D|T)$, and being simple or short to describe in our general language of thought, as measured by the prior $P(T|U)$. Probabilistic inference can operate in parallel across this hierarchical framework, propagating data-driven information upward and theory-based constraints downward to make optimal probabilistic inferences at all levels simultaneously.

2.1. A language for theories

Following (Katz et al., 2008) we choose to represent the laws in a theory as Horn clauses, logical expressions of the form $r \leftarrow (f \wedge g \wedge \ldots \wedge s \wedge t)$, where each term *r,f,g,s,t,...* is a predicate expressing an attribute or relation on entities in the domain, such as *f*(*X*) or *s*(*X,Y*). Horn clauses express logical implications – a set of conjunctive conditions under which *r* holds – but can also capture intuitive causal relations (Kemp, Goodman, & Tenenbaum, 2007) under the assumption that any propositions not generated by the theory are assumed to be false. The use of implicational clauses as a language for causal theories was explored by Feldman (2006).

While richer logical forms are possible, Horn clauses provide a convenient and tractable substrate for exploring the ideas of stochastic search over a space of theories. In our formulation, the Horn clauses contain two kinds of predicates, core and surface. Core predicates are those that cannot be reduced further using the theory's laws. Surface predicates are derived from other predicates, either surface or core, via the laws. Predicates may or may not be directly observable in the data. The core predicates can be seen as compressing the full model into the minimal bits necessary to specify all facts. A good theory is one that compresses a domain well, that explains as much of the observed data as possible using only the information specified in the core predicates. In our magnetism example, the core can be expressed in terms of two predicates *f*(*X*) and *g*(*X*). Based on an assignment of truth values to these core predicates, the learner can use the theory's laws such as $interacts(X, Y) \leftarrow f(X) \wedge g(Y)$ to derive values for the observable surface predicate *interacts*(*X,Y*). For *n* objects, there are $O(n^2)$ interactions that can be observed (between all pairs of objects), but these can be explained and predicted by specifying only $O(n)$ core predicate values (for each object, whether or not it is a magnet or is magnetic).

As another example of how a theory supports compression via its core predicates and abstract laws, consider the domain of kinship as shown in Fig. 1. A child learning this domain might capture it by core predicates, e.g., *parent*, *spouse*, and *gender*, and laws, e.g., each child has two parents of opposite gender who are each other's *spouse*; a male parent is a *father*; two individuals with the same parent are *siblings*; a female sibling is a *sister*. Systems in this domain correspond to families the child knows about. A system can then be compressed by specifying only the values of the core predicates, for example which members of a family are spouses, who is the parent of whom, and who is male or female. From this minimal set of facts and concepts all other facts about a particular family can be derived, predicting new relationships not directly observed.

In constructing a theory, the learner introduces abstract predicates via new laws, or new roles in existing laws, and thereby essentially creates new concepts. Notice that the core predicates in our magnetism theory need be represented only in purely abstract terms, $f(X)$ and $g(X)$, and initially they have only this bare abstract meaning. They acquire their meaning as concepts in virtue of the role they play in the theory's laws and the explanations they come to support for the observed data. This is the sense in which our framework allows introduction of genuinely new abstract concepts via their inferential or conceptual roles.

Entities may be typed and predicates restricted based on type constraints. For example, in the taxonomy theory shown in Fig. 1, $has\_a(X, Y)$ requires that X be a category and Y be a property, while $is\_a(X,Y)$ requires *that* X and Y both be categories. Forcing candidate models and theories to respect these type constraints provides the learner with another valuable and cognitively natural inductive bias.

Although our focus is on the acquisition of intuitive theories in general, much research has been concerned with young children's theories in a few core domains and their development early in life. Our horn-clause language is too limited to express the full richness of a two-year-old's intuitive physics or psychology, but it can represent simplified versions of them. For example, in Fig. 1 we show a fragment of a simple "desire psychology" theory, one hypothesized early stage in the development of intuitive psychology (Wellman & Woolley, 1990). This theory aims to explain agents' goal-directed actions, such as reaching for, moving toward or looking for various objects, in terms of basic but unobservable desires. In our language $desires(X,Y)$, or $d(X,Y)$ in Fig. 1, is a core predicate relating an agent X to an object Y. Desires are posited to explain observations of a surface predicate $goes\_to(X,Z,S)$. Agent X goes to (or reaches for or looks in) location Z in situation S. We also introduce background information in the form of an additional predicate $location(Y,Z,S)$ available to the child, specifying that object Y is in location Z in situation S. By positing which agents desire which objects, and a law that says effectively, "an agent will go to a certain location in a given situation if that location contains an object that the agent desires," a child can predict how agents will act in various situations, and explain why they do so.

2.2. The theory prior $P(T|U)$

We posit U knowledge in the form of a probabilistic context-free Horn clause grammar (PHCG) that generates the hypothesis space of possible Horn-clause theories, and a prior $P(T|U)$ over this space (Fig. 4). This grammar and the Monte Carlo algorithms we use to sample or search over the theory posterior $P(T|D,U)$ are based heavily on work by Goodman, Tenenbaum, Feldman, and Griffiths (2008), who introduced the approach for learning single rule-based concepts rather than the larger law-based theory structures we consider here. Given a set of possible predicates in the domain, the PHCG draws laws from a random construction process (Law) or from law templates (Tem) until the Stop symbol is reached and then grounds out these laws as horn clauses. The prior $P(T|U)$ is the product of the probabilities of choices made at each point in this derivation. Because all these probabilities are less than one, the prior favors simpler theories with shorter derivations. The precise probabilities of different laws in the grammar are treated as latent variables and integrated out, favoring re-use of the same predicates and law components within a theory (Goodman, Tenenbaum, et al., 2008).

2.3. Law templates

We make the grammar more likely to generate useful laws by equipping it with templates, or canonical forms of laws that capture structure likely to be shared across many domains. While

Top level theory			
(S1)	S	$\Rightarrow$	(Law) $\wedge$ S
(S2)	S	$\Rightarrow$	(Tem) $\wedge$ S
(S3)	S	$\Rightarrow$	Stop
Random law generation			
(Law)	Law	$\Rightarrow$	($\mathrm{F}_{left} \leftarrow \mathrm{F}_{right} \wedge$ Add)
(Add1)	A	$\Rightarrow$	F $\wedge$ Add
(Add2)	A	$\Rightarrow$	Stop
Predicate generation			
(F_{left}1)	F_{left}	$\Rightarrow$	*surface*1()
$\vdots$			
($\mathrm{F}_{left}\alpha$)	F_{left}	$\Rightarrow$	*surface*α()
(F_{right}1)	F_{right}	$\Rightarrow$	*surface*1()
$\vdots$			
($\mathrm{F}_{right}\alpha$)	F_{right}	$\Rightarrow$	*surface*α()
($\mathrm{F}_{right}(\alpha+1)$)	F_{right}	$\Rightarrow$	*core*1()
$\vdots$			
($\mathrm{F}_{right}(\alpha+\beta)$)	F_{right}	$\Rightarrow$	*core*β()
Law templates			
(Tem1)	Tem	$\Rightarrow$	template1()
$\vdots$			
(Temγ)	Tem	$\Rightarrow$	templateγ()

Fig. 4. Production rules of the Probabilistic Horn Clause Grammar. *S* is the start symbol and Law, Add, *F* and Tem are non-terminals. α, β, and γ are the numbers of surface predicates, core predicates, and law templates, respectively.

it is possible for the PHCG to reach each of these law forms without the use of templates, their inclusion allows the most useful laws to be invented more readily. They can also serve as the basis for transfer learning across domains. For instance, instead of having to re-invent transitivity anew in every domain with some specific transitive predicates, a learner can recognize that the same transitivity template applies in several domains. It may be costly to invent transitivity for the first time, but its abstract form can then be readily re-used. The specific law templates used are described in Fig. 5. Each "$F(\cdot)$" symbol stands for a non-terminal representing a predicate of a certain arity. This non-terminal is later instantiated by a specific predicate. For example, the template $F(X, Y) \leftarrow F(X, Z) \wedge F(Z, Y)$ might be instantiated as $is\_a(X, Y) \leftarrow is\_a(X, Z) \wedge is\_a(Z, Y)$ (a familiar transitive law) or as $has\_a(X, Y) \leftarrow is\_a(X, Z) \wedge has\_a(Z, Y)$ (the other key law of taxonomy, which is like saying that "*has_a* is transitive over *is_a*"). This template could be instantiated differently in other domains, for example in kinship as $child(X, Y) \leftarrow child(X, Z) \wedge spouse(Z, Y)$, which states that the child-parent relationship is transitive over spouse.

$F(X,Y) \leftarrow F(X,Z) \wedge F(Z,Y)$	$F(X,Y) \leftarrow F(X) \wedge F(Y)$
$F(X,Y) \leftarrow F(Z,X) \wedge F(Z,Y)$	$F(X,Y) \leftarrow F(Y,X)$
$F(X,Y) \leftarrow F(X,Z) \wedge F(Y,Z)$	$F(X,Y) \leftarrow F(X,Y)$
$F(X,Y) \leftarrow F(Z,X) \wedge F(Y,Z)$	$F(X) \leftarrow F(X)$
$F(X,Y) \leftarrow F(X,Y) \wedge F(X)$	$F(X) \leftarrow F(X,Y) \wedge F(X)$
$F(X,Y) \leftarrow F(Y,X) \wedge F(X)$	$F(X) \leftarrow F(Y,X) \wedge F(X)$
$F(X,Y) \leftarrow F(X,Y) \wedge F(Y)$	$F(X) \leftarrow F(X,Y) \wedge F(Y)$
$F(X,Y) \leftarrow F(Y,X) \wedge F(Y)$	$F(X) \leftarrow F(Y,X) \wedge F(Y)$

Fig. 5. Possible templates for new laws introduced by the grammar. The leftmost *F* can be filled in by any surface predicate, the right *F* can be filled in by any surface or core predicate, and *X* and *Y* follow the type constraints.

2.4. The theory likelihood P(D|T)

An abstract theory makes predictions about the observed data in a domain only indirectly, via the models it generates. A theory typically generates many possible models. Even if a child has the correct theory and abstract concepts of magnetism, she could categorize a specific set of metal bars in many different ways, each of which would predict different interactions that could be observed as data. Expanding the theory likelihood,

$$P(D|T) = \sum_{M} P(D|M)P(M|T), \tag{3}$$

we see that theory *T* predicts data *D* well if it assigns high prior *P*(*M*|*T*) to models *M* that make the data probable under the observation process *P*(*D*|*M*).

The model prior *P*(*M*|*T*) reflects the intuition that a theory *T* explains some data well if it compresses well. If it requires few additional degrees of freedom beyond its abstract concepts and laws – that is, few specific and contingent facts about the system under observation, besides the theory's general prescriptions – to make its predictions. This intuition is captured by a prior that encourages the core predicates to be as sparse as possible, thereby penalizing theories that can only fit well by "overfitting" with many extra degrees of freedom. This sparseness assumption is reasonable as a starting point for many domains, given that core predicates are meant to explain and compress the data. Formally, we assume a conjugate beta prior on all binary facts in *M*, modeled as Bernoulli random variables that we integrate out analytically (Katz et al., 2008).

Finally, the model likelihood *P*(*D*|*M*,*T*) comes from assuming that we are observing randomly sampled facts (sampled with replacement, so the same fact could be observed on multiple occasions), which also encourages the model extension to be as small as possible. This provides a form of implicit negative evidence (Tenenbaum & Griffiths, 2001), useful as an inductive bias when only positive facts of a domain are observed.

2.5. Stochastic search in theory space: a grammar-based Monte Carlo algorithm

Following Goodman, Tenenbaum, et al. (2008), we use a grammar-based Metropolis-Hastings (MH) algorithm to sample theories from the posterior distribution over theories conditioned on data, $P(T|D,U)$. This algorithm is applicable to any grammatically structured theory space, such as the one generated by our PHCG; it is also a version of the Church MH inference algorithm (Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2008). The MH algorithm is essentially a Markov chain on the space of potential derivations from the grammar, where each step in the chain – each proposed change to the current theory – corresponds to regenerating some subtree of the derivation tree from the PHCG. For example, if our theory of magnetism includes the law *interacts* $(X, Y) \leftarrow f(X) \wedge g(Y)$, the MH procedure might propose to add or delete a predicate, e.g., *interacts* $(X, Y) \leftarrow f(X) \wedge g(Y) \wedge h(Y)$ or *interacts*$(X,Y) \leftarrow f(X)$, to change one predicate to an alternative of the same form, e.g., *interacts* $(X, Y) \leftarrow f(X) \wedge h(Y)$, or a different form if available, e.g., *interacts* $(X, Y) \leftarrow f(X) \wedge s(X, Y)$; to resample the law from a template, e.g., *interacts* $(X, Y) \leftarrow t(X, Z) \wedge t(Z, Y)$; or to add or delete a whole law.

These proposals are accepted with probability equal to the maximum of 1 and the MH acceptance ratio,

$$\frac{P(T'|D, U)}{P(T|D, U)} \cdot \frac{Q(T|T')}{Q(T'|T)}, \tag{4}$$

where T is the current theory, T' is the new proposed theory, and $Q(\cdot/\cdot)$ is the transition probability from one theory to the other, derived from the PHCG (Goodman, Tenenbaum, et al., 2008). To aid convergence we raise these acceptance ratios to a power greater than 1, which we increase very slightly after each MH step in a form of simulated annealing. Early on in learning, a learner is thus more likely to try out a new theory that appears worse than the current one, exploring candidate theories relatively freely. However, with time the learner becomes more conservative – increasingly likely to reject new theories unless they lead to an improved posterior probability.

While this MH algorithm could be viewed merely as a way to approximate the calculations necessary for a hierarchical Bayesian analysis, we suggest that it could also capture in a schematic form the dynamic processes of theory acquisition and change in young children. Stochastic proposals to add a new law or change a predicate within an existing law are consistent with some previous characterizations of children's theory learning dynamics (Siegler & Chen, 1998). These dynamics were previously proposed on purely descriptive grounds, but here they emerge as a consequence of a rational learning algorithm. Although the dynamics of an MH search might appear too random to an omniscient observer who knows the "true" target of learning, it would not be fair to call the algorithm sub-optimal, because it is the only known general-purpose approach for effectively searching a complex space of logical theories. Likewise, the annealing process that leads learning to look child-like in a certain sense – starting off with more variable, rapidly changing and adventurous theories, then becoming more conservative and less variable over time – also makes very good engineering sense. Annealing has proven to be useful in stochastic search problems across many scientific domains (Kirkpatrick et al., 1983) and is the only known method to ensure that a stochastic search converges to the globally optimal solution. It is plausible that some cognitive analog of annealing could be at work in children's learning.[3]

2.6. Approximating the theory score: an inner loop of MCMC

Computing the theory likelihood $P(D|T)$, necessary to compare alternative theories in Equation (4), requires a summation over all possible models consistent with the current theory (Equation (3)). Because this sum is typically very hard to evaluate exactly, we approximate $P(D|T)$ with

[3] Annealing could be implemented in a learning system without an explicit temperature parameter or cooling schedule, merely based on experience accumulating over time. Here for simplicity we have kept the learner's dataset fixed, but if the learner is exposed to increasing amounts of data over time and treats all data as independent samples from the model, this also acts to lower the effective temperature by creating larger ratios between likelihoods (and hence posterior probabilities) for a given pair of theories.

$P(D|M^*,T)P(M^*|T)$, where M^* is an estimate of the maximum a-posteriori (MAP) model inferred from the data, the most likely values of the core predicates. The MAP estimate M^* is obtained by running an inner sampling procedure over the values of the core predicates. As in (Katz et al., 2008) we use Gibbs sampling, a specialized form of Metropolis Hastings. The Gibbs sampler goes over each core predicate assignment in turn, keeping all other assignments fixed and proposes changes to the currently considered assignment. As a concrete example of how the Gibbs loop works, consider a learner who is proposing a theory that contains the law *interacts* $(X, Y) \leftarrow f(X) \wedge g(Y)$, i.e., objects for which core predicate *f* is true interact with objects for which core predicate *g* is true. The learner begins by randomly extending the core categories over the domain's objects, e.g., *f* might be posited to hold for objects 1, 4, and 7, while *g* holds for objects 2, 4, 6, and 8. (Either, both or none of the predicates may hold for any object, a priori.) The learner then considers the extension of predicate *f* and proposes removing object 1, scoring the new model (with all other assignments as before) on the observed data and accepting the proposed change probabilistically depending on the relative scores. The learner then considers objects 2, 3, and so on in turn, considering for each object whether predicate *f* should apply, before moving on to predicate *g*. (These object-predicate pairs are often best considered in random order on each sweep through the domain.) This process continues until a convergence criterion is reached. We anneal slightly on each Gibbs sweep to speed convergence and lock in the best solution. The Gibbs sampler over models generated by a given theory is thus an "inner loop" of sampling in our learning algorithm, operating within each step of an "outer loop" sampling at a higher level of abstract knowledge, the MH sampler over theories generated by *U* knowledge.

3. Case studies

We now explore the performance of this stochastic approach to theory learning in two case studies, using simulated data from the domains of taxonomy and magnetism introduced above. We examine the learning dynamics in each domain and make more explicit the possible parallels with human theory acquisition. The data presented to the learning algorithm for both of these case studies can be found at http://web.mit.edu/cocosci/tlss/tlss_data.pdf.

3.1. Taxonomy

As we saw earlier, the domain of taxonomy illustrates how a compressive knowledge representation is useful in capturing semantic data. How can such a powerful organizing principle itself be learned? Katz et al. (2008) showed that a Bayesian ideal observer can pick out the best theory of taxonomy given a small set of eight possible alternatives. Here we show that the theory of taxonomy can be learned in a more constructive way, via an MCMC search through our infinite grammar-generated hypothesis space. The theory to be learned takes the following form:

Two core predicates:	$s(X,Y)$ and $t(X,Y)$
Two observable predicates:	$is\_a(X,Y)$ and $has\_a(X,Y)$
Law 1:	$is\_a(X,Y) \leftarrow s(X,Y)$
Law 2:	$has\_a(X,Y) \leftarrow t(X,Y)$
Law 3:	$is\_a(X, Y) \leftarrow is\_a(X, Z) \wedge is\_a(Z, Y)$
Law 4:	$has\_a(X, Y) \leftarrow is\_a(X, Z) \wedge has\_a(Z, Y)$

These laws by themselves do not yet capture the complete knowledge representation we seek; we also need to instantiate the core predicates in a particular model. These laws allow many possible models for any given data sets. One of these models is the compressed tree representation (Fig. 1), which specifies only the minimal facts needed to derive the observed data from Laws 1 to 4. A different model could link explicitly all the $is\_a(X,Y)$ connections, for example drawing the links between salmon and animal, shark and animal and so on. Another model could link explicitly all the $has\_a(X,Y)$ connections. However, these latter two models would be much less sparse than the compressed tree representation, and thus would be disfavored relative to the compressed tree shown in Fig. 1, given how we have defined the model prior $P(M|T)$. In sum, in this framework, the organization of categories and properties into a tree-structured inheritance hierarchy comes about from a combination

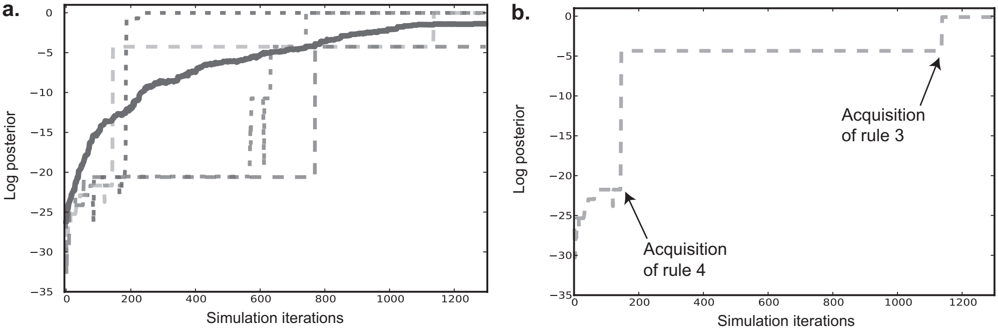

Fig. 6. Representative runs of theory learning in Taxonomy. (a) Dashed lines show different runs. Solid line is the average across all runs. (b) Highlighting a particular run, showing the acquisition of law 4, followed by the acquisition of law 3 and thus achieving the final correct theory.

of positing the appropriate abstract laws and core predicates together with a sparsity preference on the assignments of the core predicates' values.

Note also that the core predicates $s(X,Y)$ and $t(X,Y)$ acquire their meaning in part by their inferred extensions and in part by how they are related to the observed surface predicates. The surface predicates are assumed to be verbal labels that the learner observes and needs to account for. The link between these verbal labels and the core relations are given by Laws 1 and 2. While these links could in general also be learned, we follow Katz et al. (2008) in taking Laws 1 and 2 as given for this particular domain and asking whether a learner can theoretically discover Laws 3 and 4 – but now at the algorithmic level. We test learning for the same simple model of the taxonomy domain studied by Katz et al., using seven categories and seven properties in a balanced tree structure. We presented all facts from this model as observations to the learner, including both property statements (e.g., "An eagle has claws") and category membership statements (e.g., "An eagle is a bird").

We ran 60 simulations, each comprising 1300 iterations of the outer MH loop (i.e., moves in the space of theories). Four representative runs are shown in Fig. 6, as well as the average across all runs. Of 60 simulations, 52 found the correct theory within the given number of iterations, and 8 discovered a partial theory which included only Law 3 or Law 4.

It is striking that abstract structure can be learned effectively from very little data. Using simple local search, our learning algorithm is able to navigate an infinite space of potential theories and discover the true laws underlying the domain, even with relatively few observations. This is a version of the "blessing of abstraction" (Goodman et al., 2011; Tenenbaum et al., 2011), but one that is realized at the algorithmic level and not just the computational level of ideal learning.

Individual learning trajectories proceed in a characteristic pattern of stochastic leaps. Discovering the right laws gives the learner strong explanatory power. However, surrounding each "good" theory in the discrete hypothesis space are many syntactically similar but nonsensical or much less useful formulations. Moving from a good theory to a better one thus depends on proposing just the right changes to the current hypothesis. Because these changes are proposed randomly, the learner often stays with a particular theory for many iterations, rejecting many proposed alternatives that score worse or not significantly better than the current theory, until a new theory is proposed that is so much better it is almost surely accepted. This leads to the observed pattern of plateaus in the theory score, punctuated by sudden jumps upward and occasional jumps downward in probability.

While we do not want to suggest that people acquire theories only by making random changes to their mental structures, the probabilistic nature of proposals in a stochastic search algorithm could in part explain why individual human learning curves rarely proceed along a smooth path and can show broad individual variation. While individual learning trajectories may be discontinuous, on average learning appears smooth. Aggregating performance over all runs shows a smooth improvement of the theory's score that belies the discrete nature of individual learning, a point in common with the motivation for microgenetic studies of learning trajectories in individual children (Siegler & Crowley, 1991).

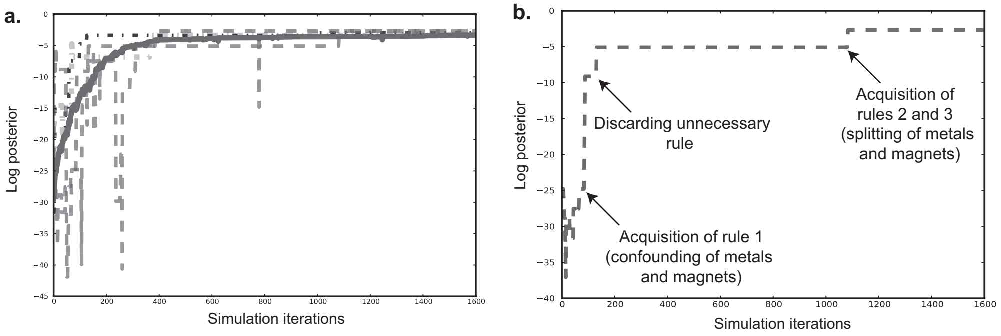

Fig. 7. Representative runs of theory learning in Magnetism. (a) Dashed lines how different runs. Solid line is the average across all runs. (b) Highlighting a particular run, showing the acquisition of Law 1 and the confounding of magnets and magnetic objects, to the acquisition of the final theory and the conceptual splitting of the two.

3.2. *Magnetism*

We now turn to the domain of magnetism, where we observe interesting intermediate stages and transitions corresponding to classic phenomena of conceptual change in childhood and early science (Carey, 2009). The simplified theory of magnetism to be learned includes two core predicates and three laws, as shown in Fig. 1 or Fig. 8 (“Theory C”).

The learner observed data drawn from a model with 10 objects: 3 magnets, 5 magnetic objects and 2 non-magnetic objects. The learner was given all true facts in this model, observing interactions between each magnet and every other object that was either a magnet or a magnetic object, but no other interactions. The learner was given none of the laws or core predicate structure to begin with; the entire theory had to be constructed by the search algorithm.

We ran 70 simulations, each comprising 1600 iterations of the outer MH loop sampling over candidate theories. The dynamics of representative runs are displayed in Fig. 7, as well as the average over all the runs. As in the domain of taxonomy, individual learners experienced radical jumps in their theories, while aggregating across runs learning appears to be much smoother. Out of 70 simulated learning runs, 50 found the correct theory or a minor logical variant which was equivalent to it; the rest discovered a partial theory. As we discuss in the next section, the most frequent partial theories provide a good first approximation of the domain, and are plausible intermediate points for learning.

The most interesting aspects of learning here were found in the transitions between distinct stages of learning, when novel core predicates are introduced and existing core predicates shift their meaning in response. Key transitions in children’s cognitive development may be marked by restructuring of concepts, as when one core concept differentiates into two (Carey, 1985). In the magnetism task there was no single order of concept acquisition that the algorithm always followed, but the most common trajectory (shown in Fig. 7b) involves learning Law 1 first, followed later by the acquisition of Laws 2 and 3. For a learner who knows only Law 1, the optimal setting of the core predicates is to use only one core predicate and to lump together magnets and magnetic objects, essentially not differentiating between them. Only when Laws 2 and 3 are learned does the learner also acquire a second core predicate that carves off the magnetic non-magnets from the magnets. This ‘conceptual splitting’ moment is marked in a representative run in Fig. 7b. Other less common learning trajectories showed similar but less dramatic conceptual restructurings.

4. Two sources of learning dynamics

To construct adult-level intuitive theories, children require time to ponder and exposure to sufficient evidence. For a child on the verge of grasping a new theory, either additional data or additional time to think can make the difference (Carey, 2009). The dynamics of learning typically follow an arc from simpler theories that only coarsely predict or encode experience to more complex theories that

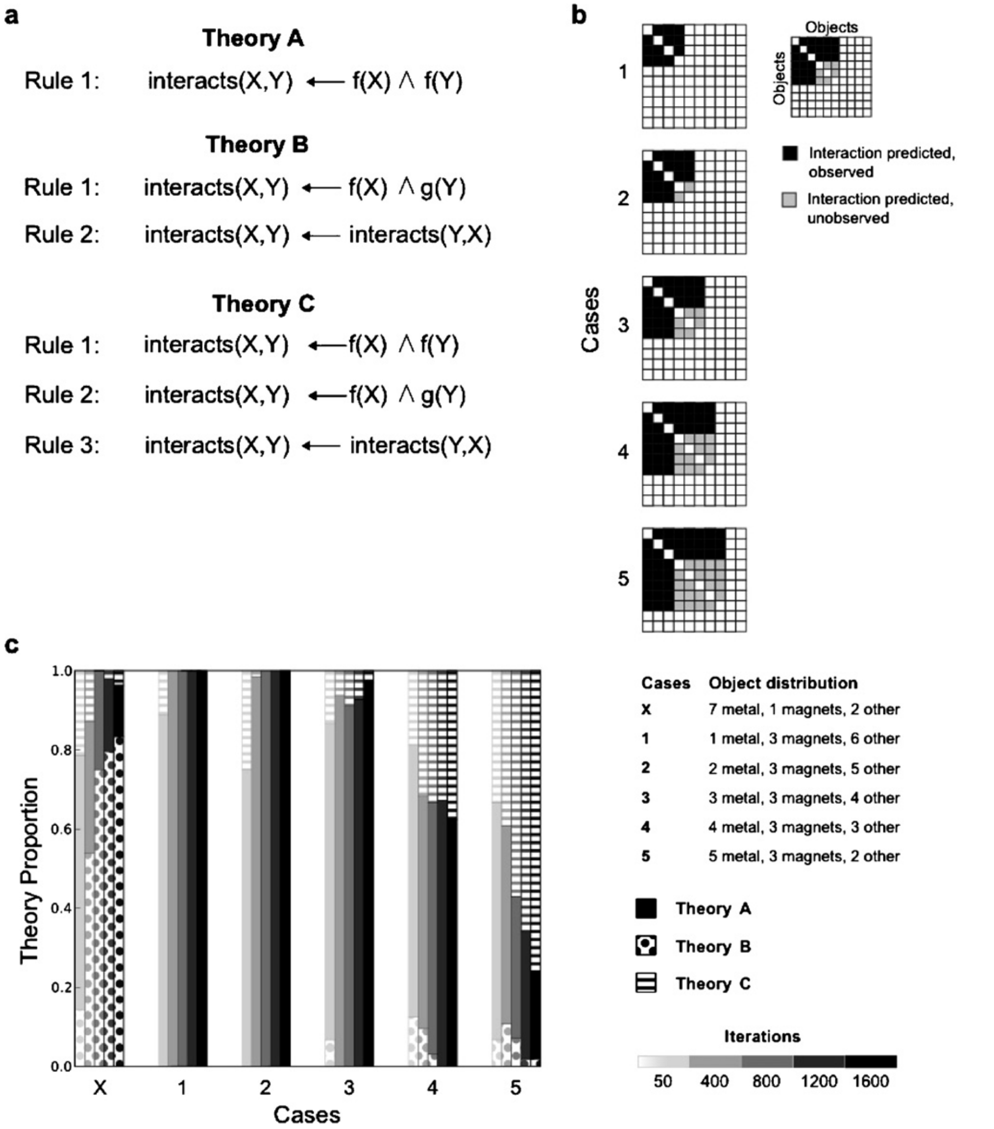

Fig. 8. Learning dynamics resulting from two different sources: (a) a formal description of theories A, B and C; (b) predicted and observed interactions given theory A for the different cases, showing the growing number of outliers as the number of magnetic non-magnet objects grows; (c) proportion of theories accepted by learner for different cases, during different points. More opaque bars correspond to later iterations.

more faithfully predict and encode it. Our case studies of theory learning in the domains of taxonomy and magnetism show this dynamic as a function of time elapsed in the search process, for a fixed data set. Previous Bayesian models of theory learning (Kemp & Tenenbaum, 2009) have emphasized the complementary perspective – how increasing amounts of data naturally drive an ideal learner to more complex but more predictive theories, independent of the dynamics of search or inference.

These two sources of learning dynamics are most naturally understood at different levels of analysis. Data-driven learning dynamics seem best explained at the computational level, where the ideal learner shifts probability mass between candidate theories as a function of the data observed. In contrast, time-driven dynamics (independent of amount of data observed) seem best approached at the algorithmic level, with models that emphasize how the learner's process of searching over a hypothesis space unfolds over time independent of the pace at which data accumulates.

Our modeling approach is well suited to studying both data-driven and time-driven dynamics and their interactions because of its focus on the interface between the computational and algorithmic levels of analysis. How does varying time and data affect our ideal learner? We provide the learner with several different data sets and examine how the learning dynamics unfold over time for each one of these sets. We again use the domain of simplified magnetism, and parametrically vary the number of magnetic objects across the data sets, as shown in Fig. 8. At the end of the simulation the learner almost always settled on one of three theories. We therefore focus here on these three theories (the formal laws of which appear in Fig. 8a):

Theory A: There is one class of interacting objects; objects in this class interact with other objects in this class.
Theory B: There are two classes of interacting objects; objects from one class interact with objects in the other class. These interactions are symmetric.
Theory C: There are two classes of interacting objects; objects from one class interact with objects in the other class. Also, objects in one of the classes interact with other objects in the same class. These interactions are symmetric.

Theories A, B and C were not given to the learner as some sort of limited hypothesis space. Rather, the number of possible theories the learner could consider in each case is potentially infinite, but practically it settles on one of these three or their logical equivalents. Many other theories were considered by the learner, but they do not figure significantly into the trajectory of learning. These theories are much less suitable (i.e., unnecessarily complex or poorly fitting) relative to neighboring knowledge states, so they tend to be proposed and accepted only in the early, more random stages of learning, and are quickly discarded. We could not find a way to group these other theories into cohesive or sensibly interpreted classes, and as they are only transient states of the learner, we removed them for purposes of analyzing learning curves and studied only the remaining proportions, renormalized.

When there are few magnetic objects that are not magnets, perhaps 1 or 2 (cases 1 and 2. Fig. 8), a partial theory such as theory A might suffice, where there is only one type of interacting object and one law. If there are two magnetic non-magnets in the domain, the partial theory will classify them as 'interacting' objects based on their behavior with the magnets, conflating them with the magnets. However, it will incorrectly predict the two magnetic non-magnets should interact with each other. Their failure to interact will be treated as an outlier by the learner who holds theory A. The full theory C can correctly predict this non-interaction, but it does so by positing more laws and types of objects, which has a lower prior probability. As the number of magnetic non-magnets increases, the number of 'outliers' in the data increases as well (see Fig. 8b). Theory A now predicts more and more incorrect interactions. There is a point at which these failures can no longer be ignored, and a qualitative shift to a new theory is preferred. In a completely different scenario, such as the extreme case of only 1 magnet (case X), we might expect the learner to not come up with magnet interactions laws, and settle instead on theory B.

For each one of these cases we ran 70 simulations for 1600 iterations. Fig. 8c displays the relative proportion of the outlined theories at the end of the iteration. Note the transition from case 1 to case 5: With a small number of non-magnet magnetic objects, the most frequently represented theory is theory A, which puts all magnetic objects (magnet or not) into a single class and treats the lack of interactions between two magnetic non-magnets as essentially noise. As the number of magnetic non-magnets increases, the lack of interactions between the different non-magnets can no longer be ignored and the full theory becomes more represented. Case X presents a special scenario in which there is only 1 magnet, and as expected theory B is the most represented there. The source of the difference between the proportion of theories learned in these different cases is the data the learner was exposed to. In contrast, within each case the learner undergoes a process of learning similar to that described in the case studies–adopting and discarding theories in a process or time-driven manner.

To summarize, theory acquisition can be both data-driven and process-driven. Only when the observed data provide a strong enough signal – as measured here by potential outliers under a simpler theory – is there sufficient inductive pressure for a Bayesian learner guided by simplicity priors to posit a more complex theory. Yet even with all the data in the world, a practical learning algorithm still requires sufficient time to think, to search through a challenging combinatorial space of candidate laws and novel concepts and construct a sequence of progressively higher scoring theories that will reliably converge on the highest scoring theory. The fact that both sufficient data and sufficient time are needed for proper theory learning fits with the experience of teachers and parents. Having laid out for a child all the data needed to solve a problem, grasp an explanation, or make a discovery, the child may still take surprisingly long to get it, or not get it before a parent or teacher runs short on patience.

5. Evidence from experiments with children

Our key result is that a simple, cognitively plausible, stochastic search algorithm, guided by an appropriate grammar and language for theories, is capable of solving the rather sophisticated joint inference problem of learning both the concepts and the laws of a new theory. Several lines of experimental work have shown that children and adults can indeed solve this joint inference problem in the course of acquiring new theories. Kemp, Tenenbaum, Niyogi, and Griffiths (2010) showed that adults were able to learn new causal relations, such as *objects of type A light up type B*, and to use these relations to categorize objects, for example *object 3 is of type A*. Lucas, Gopnik, and Griffiths (2010) had adults perform a task which required inferring the abstract functional form of the causal relations (do Blickets activate a meter via a noisy-OR function, a deterministic disjunctive function or a conjunctive function?).

While in these studies children were explicitly told that only one type of concept is involved, Schulz, Goodman, Tenenbaum, and Jenkins (2008) showed that young children can solve an even more challenging task. Given sparse evidence in the form of different blocks touching and making different noises, children correctly posited the existence of three different causal kinds underlying the observed relations. They had to both infer the abstract relations governing the behavior and posit how many concepts underlie these relations. These findings are consistent with our predictions. Bonawitz, Gopnik, Denison, and Griffiths (2012) showed a more quantitative correspondence between our model predictions and children's categorization judgments. Children were shown interactions in a domain of simplified magnetism, where several unlabeled blocks interacted with blue and yellow blocks, either attracting or repelling them. We also showed that the Monte Carlo search algorithm given here is capable of finding just the theories that children do, or theories that are behaviorally indistinguishable from them, and revising them appropriately.

Connectionist architectures could potentially solve aspects of the tasks described by Lucas et al. (2010). There are certainly networks capable of distinguishing between such functional forms, which may be seen as learning governing laws in a theory. Connectionist networks can also form new concepts – in the sense of clusters of data that behave similarly – via competitive learning. However, it has yet to be shown that a connectionist network can learn or represent the kinds of abstract knowledge that our approach does, and that children grasp in the other experiments cited above. Solving the joint inference problem of discovering a system of new concepts and laws that together explain a set of previously unexpected interactions or relations. This problem poses an intriguing open challenge for connectionist modelers.

We would also like more fine-grained tests of whether our Monte Carlo learning mechanism corresponds to the way children explore the space of theories. We are currently working with Bonawitz and colleagues to test general predictions of our approach, such as the tradeoff between data and time. In these experiments we ask children to categorize objects by type, in a domain similar to simplified magnetism. We vary the length of time children have to think about a fixed amount of evidence, and we ask whether the number of types that children find follows the time-dependent theory transitions of the algorithm.

More precision could come from microgenetic methods (Siegler & Crowley, 1991), which study developmental change by giving children the same task several times and inspecting the strategies used to solve the task at many intervals. Similar to how microgenetic studies keep a task fixed, we can observe how children play and experiment with a given set of objects, without introducing new objects or new interactions. As in classic microgenetic studies, we ask the children questions and encourage them to talk out loud about their hypotheses. We can classify and score the theories they uncover using our computational framework, and observe whether the pattern of theories abandoned, adopted and uncovered fits with the predictions of Monte Carlo search.

6. Conclusions

We have presented an algorithmic model of theory learning in a hierarchical Bayesian framework and explored its dynamics. We find encouraging the successful course of acquisition for several sample theories and the qualitative parallels with phenomena of human theory acquisition. These results suggest that previous "ideal learning" analyses of Bayesian theory acquisition can be approximately realized by a simple stochastic search algorithm and thus are likely well within the grasp of child learners. It is also encouraging to think that state-of-the-art Monte Carlo methods used in Bayesian statistics and artificial intelligence to approximate ideal solutions to inductive inference problems might also illuminate how children learn. At this intersection point between computational and algorithmic levels of analysis, we showed that theory change is expected to be both data-driven and process-driven. This is an important theoretical distinction, but the psychological reality of these two sources of learning dynamics and their interaction needs to be further studied in experiments with children and adults.

While the main contributions of this paper are in addressing the algorithmics of theory acquisition, the 'how', the introduction of law templates provides some insight regarding 'what' the structure of children's knowledge might be, and the coupling between how we answer 'what?' and 'how?' questions of learning. On an algorithmic level, we found such templates to be crucial in allowing learning to converge on a reasonable timescale. On a computational level, these templates can be seen as generalizing useful abstract knowledge across domains and providing high-level constraints that apply across all domain theories. The formal framework did not directly treat where such templates come from, but it is possible to imagine that they are themselves learned during the algorithmic acquisition process. An algorithmic grammar-based model can learn templates by abstracting successful rules from their particular domain instantiation. That is, if the model (or child) discovers a particularly useful rule involving a specific predicate such as "if *is_a*(*X*,*Y*) and *is_a*(*Y*,*Z*), then *is_a*(*X*,*Z*)", the specific predicate might be abstracted away to form the transitive template "if *F*(*X*,*Y*) and *F*(*Y*,*Z*), then *F*(*Y*,*Z*)". Learning this transitive template then allows its reuse in subsequent theory and represents a highly abstract level of knowledge.

The algorithm we have explored is only one particular instance of a more general proposal for how stochastic search operating over a hierarchically structured hypothesis space can account for theory acquisition. The specific theories considered here were only highly simplified versions of the knowledge children have about real-world domains. Part of the reason that actual concepts and theories are richer and more complex is that children have a much richer underlying language for representations. Horn clauses are expressive and suitable for capturing only some knowledge structures, and in particular certain kinds of causal relations. A potentially more suitable theory space would be built on a functional language, in which the laws are more similar to mathematical equations. Such a space would be harder to search through, but it would be much more expressive. A functional language of this sort would allow us to explore rich theories described in children, such as basic notions about objects and their interactions (Spelke, 1990), and the intuitive physics that guides object behavior (Baillargeon, 1994). Despite the need for a more expressive language, we expect the same basic phenomena found in the model domains considered here to be replicated in more complex models.

Relative to previous Bayesian models of cognitive development that focused on only the computational level of analysis, we have emphasized algorithmic-level implementations of a hierarchical Bayesian computational theory and the interplay between the computational and algorithmic levels. We have not discussed at all the level of neural implementation, but analogous stochastic-sampling ideas could plausibly be used to carry out Bayesian learning in the brain (Fiser, Berkes, Orbn, & Lengye, 2010). More generally, a "top-down" path to bridging levels of explanation in the study of mind and brain, starting with higher, more functional levels and moving down to lower, more mechanistic levels, appears most natural for Bayesian or other “reverse-engineering” approaches to cognitive modeling (Griffiths et al., 2010).

Other paradigms for cognitive modeling adopt different ways to navigate the same hierarchy. Connectionist approaches, for instance, start from hypothesized constraints on neural representations (e.g., distributed codes) and learning mechanisms (e.g., error-driven learning) and move up from there, to see what higher-level phenomena emerge (McClelland et al., 2010). While we agree that actual biological mechanisms will ultimately be a central feature of any account of cognitive development, we are skeptical that this is the best place to start (Griffiths et al., 2010). The details of how the brain might represent or learn knowledge such as the abstract theories we consider here remain largely unknown, making a bottom-up emergent alternative to our approach hard to contemplate. In contrast, while our top-down approach has yet to make contact with neural phenomena, it has yielded insights spanning levels. In moving from computational-level accounts to algorithms that explicitly (if approximately) implement the computational theory we see plainly how the basic representations of children’s theories could be acquired, and suggest explanations for otherwise puzzling features of the dynamics of learning in young children, as the consequences of efficient and effective algorithms for approximating the rational computational-level ideal of Bayesian learning. We hope that ultimately our top-down approach can be meaningfully extended from the algorithmic level to the level of implementation in the brain’s hardware.

What do the dynamics explored here tell us about the coupled challenges of learning the laws of a theory, and the invention of truly novel concepts, and the opposing views represented by Fodor and Carey? There is a sense in which, at the computational level, the learner already must begin the learning process with all the laws and concepts needed to represent a theory already accessible. Otherwise the necessary hypothesis spaces and probability distributions for Bayesian learning could not be defined. In this sense, Fodor’s skepticism on the prospects for learning or constructing truly novel concepts is justified. Learning cannot really involve the discovery of anything “new”, but merely the changing of one’s degree of belief in a theory, transporting probability mass from one part of the hypothesis space to another. However, on the algorithmic level, the level of active processing for any real-world learner, there is in fact genuine discovery of new concepts and laws. Our learning algorithm can begin with no explicitly represented knowledge in a given domain – no laws, no abstract concepts with any non-trivial extensions in the world – and acquire reasonable theories comprised of novel laws and concepts that are meaningfully grounded and predictively useful in that domain.

Our specific algorithm suggests an account of how new concepts derive their meanings. Initially, the concepts themselves are only blank predicates. The theory prior induces a non-arbitrary structure on the space of possible laws relating these predicates and in that sense can be said to contain a space of proto-meanings. The data are then fused with this structure in the prior to create a structured posterior: The concepts are naturally extended over the observed objects in those regions where the posterior has a high probability, and those are the areas in theory space toward which the learner will converge. This algorithmic process is, we suggest, an instance (albeit a very simple one) of Carey’s (2004, 2009) “bootstrapping” account of conceptual change, and a concrete computational implementation of concept learning under an inferential role semantics.

According to Carey’s account of the origins of new concepts, children first use symbols as placeholders for new concepts and learn the relations between them that will support later inferential roles. Richer meaning is then filled in on top of these placeholders and relations, using a “modeling process” involving a range of inductive inferences. The outer loop of our algorithm explains the first stage – why some symbolic structures are used rather than others and how their relations are created. The second stage parallels the inner loop of our algorithm, which attempts to find the likeliest and sparsest assignment of the core predicates, once their interactions have been fixed by the proposed theory. During our algorithmic learning process, new concepts may at times have only a vague meaning, especially when they are first proposed. Concepts that are fragmented can be unified, and concepts that are lumped together may be usefully dissociated, as learners move around theory space in ways similar to how new concepts are manipulated in both children’s and scientists’ theory change (Carey, 2009).

Despite our optimism, it is important to end by stressing that our models at best only begin to capture some aspects of how children acquire theories. We agree very much with the view of Schulz

(2012a) that the hardest aspects of the problem are as yet unaddressed by any computational account, that there are key senses in which children's learning is a kind of exploration much more intelligent and sophisticated than even a smart randomized search such as our grammar-based MCMC. How could our learning algorithms account for children's sense of curiosity, knowing when and where to look for new evidence? How do children come up with the proper interventions to unconfound concepts or properties? How can a learning algorithm know when it is on the right track, so to speak, or distinguish good bad ideas from bad bad ideas, which children seem able to do? How do pedagogy and learning from others interact with internal search dynamics? Are the ideas being taught simply accepted, or do they form the seed of a new search? How can algorithmic models go beyond the given evidence and actively explore, in the way children search for new data when appropriate? There is much toil left – much rewarding toil, we hope – until we can say reasonably that we have found a model of children's learning and believe it.

Acknowledgments

We wish to thank Laura Schulz, Liz Bonawitz, Alison Gopnik, Rebecca Saxe, Henry Wellman and Yarden Katz for helpful discussions. This work was funded by grants from the McDonnell Causal Learning Collaborative, ONR (N00014-09-0124), ARO (W911NF-08-1-0242) and an NSF Graduate Fellowship to the first author.

References

Baillargeon, R. (1994). How do infants learn about the physical world? *Current Directions in Psychological Science*, *3*, 133–140.

Bishop, C. M. (2006). *Pattern recognition and machine learning* (1st ed.). Springer.

Block, N. (1986). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy*, *10*, 615–678.

Bonawitz, E., Gopnik, A., Denison, S., & Griffiths, T. (2012). Rational randomness: The role of sampling in an algorithmic account of preschooler's causal learning. In F. Xu, & T. Kushnir (Eds.), *Rational constructivism in cognitive development*. Elsevier.

Bradshaw, G., Langley, P., & Simon, H. (1983). Studying scientific discovery by computer simulation. *Science*, *22*, 971–975.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.

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

Carey, S. (2004). Bootstrapping and the origin of concepts. *Daedalus*, *133*, 59–68.

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

Collins, A., & Quillian, M. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*, 240–247.

Feldman, J. (2006). An algebra of human concept learning. *Journal of Mathematical Psychology*, *50*, 339–368.

Feldman, J. A., Gips, J., Horning, J. J., & Reder, S. (1969). *Grammatical complexity and inference*. Technical Report Stanford University.

Field, H. (1977). Logic, meaning, and conceptual role. *Journal of Philosophy*, *74*, 379–409.

Fiser, J., Berkes, P., Orbn, G., & Lengye, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences*, *14*, 119–130.

Fodor, J., & Lepore, E. (1991). Why meaning (probably) isn't conceptual role. *Mind & Language*, *6*, 328–343.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Fodor, J. A. (1980). On the impossibility of acquiring 'more powerful' structures. In *Language and learning: The debate between Jean Piaget and Noam Chomsky*. Harvard University Press.

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

Gershman, S., Vul, E., & Tenenbaum, J. B. (2009). Perceptual multistability as Markov Chain Monte Carlo inference. *Advances in Neural Information Processing Systems*, *22*, 611–619.

Gilks, W., & Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC.

Goodman, N. D., Mansinghka, V. K., Roy, D. M., Bonawitz, K., & Tenenbaum, J. B. (2008). Church: A language for generative models. In *Uncertainty in artificial intelligence*

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*, 108–154.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*, 110–119.

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

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*, 357–364.

Harman, G. (1975). Meaning and semantics. In M. Munitz, & P. Unger (Eds.), *Semantics and philosophy*. New York: New York University Press.

Harman, G. (1982). Conceptual role semantics. *Notre Dame Journal of Formal Logic*, *23*, 242–257.

Katz, Y., Goodman, N. D., Kersting, K., Kemp, C., & Tenenbaum, J. B. (2008). Modeling semantic cognition as logical dimensionality reduction. In *Proceedings of the thirtieth annual conference of the cognitive science society*.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2007). Learning causal schemata. In *Proceedings of the twenty-ninth annual meeting of the cognitive science society*.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2008a). Learning and using relational theories. *Advances in Neural Information Processing Systems*, *20*, 753–760.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2008b). Theory acquisition and the language of thought. In *Proceedings of thirtieth annual meeting of the cognitive science society*.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*, 20–58.

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

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671–680.

Lucas, C. G., Gopnik, A., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. In *Proceedings of the 32nd annual conference of the cognitive science society*.

Marr, D. (1982). *Vision*. Freeman Publishers.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., et al. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*, 348–356.

Mcclelland, J. L., & Rumelhart, D. E. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 2: Psychological and biological models*. Cambridge, MA: MIT Press.

Mitchell, T. (1982). Generalization as search. *Artificial Intelligence*, *18*(2), 203–226.

Moreno-Bote, R., Knill, D., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, *108*, 12491–12496.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289–316.

Newell, A., & Simon, H. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19*, 113–126.

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

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

Russell, S., & Norvig, P. (2009). *Artificial Intelligence: A modern approach* (3rd ed.). Prentice Hall.

Schulz, L. E. (2012a). *Finding new facts; thinking new thoughts*. *Rational constructivism, Advances in child development and behavior* Elsevier.

Schulz, L. E. (2012b). The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in Cognitive Science*, *16*, 382–389.

Schulz, L. E., Goodman, N. D., Tenenbaum, J. B., & Jenkins, A. C. (2008). Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data. *Cognition*, *109*, 211–223.

Shultz, T. R. (2003). *Cognitive developmental psychology*. Cambridge, MA, USA: MIT Press.

Siegler, R., & Crowley, K. (1991). The microgenetic method. *American Psychologist*, *46*, 606–620.

Siegler, R. S., & Chen, Z. (1998). Developmental differences in rule learning: A microgenetic analysis. *Cognitive Psychology*, *36*, 273–310.

Smith, E., & Medin, D. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Spall, J. C. (2003). *Introduction to stochastic search and optimization: Estimation simulation and control*. John Wiley and Sons.

Spelke, E. (1990). Principles of object perception. *Cognitive Science*, *14*, 29–56.

Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, prediction and search* (2nd ed.). Cambridge, MA: MIT Press.

Sundareswara, R., & Schrater, P. R. (2008). Perceptual multistability predicted by search model for Bayesian decisions. *Journal of Vision*, *8*, 1–19.

Tenenbaum, J. B., & Griffiths, T. L. (2001). The rational basis of representativeness. In *Proceedings of the 23rd annual conference of the cognitive science society* (pp. 1036–1041).

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*, 309–318.

Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (2007). Intuitive theories as grammars for causal inference. In A. Gopnik, & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation*. Oxford: Oxford University Press.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279–1285.

Wellman, H. M., Fang, F., & Peterson, C. C. (2011). Sequential progressions in a theory-of-mind scale: Longitudinal perspectives. *Child Development*, *82*, 780–792.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*, 337–375.

Wellman, H. M., & Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, *35*, 245–275.