COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 43 (2019)

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12765

Sticking to the Evidence? A Behavioral and Computational Case Study of Micro-Theory Change in the Domain of Magnetism

Elizabeth Bonawitz, Tomer D. Ullman, Sophie Bridgers, Alison Gopnik, Joshua B. Tenenbaum

Department of Psychology, Rutgers University – Newark

Received 26 April 2019; received in revised form 14 May 2019; accepted 21 May 2019

Abstract

Constructing an intuitive theory from data confronts learners with a "chicken-and-egg" problem: The laws can only be expressed in terms of the theory's core concepts, but these concepts are only meaningful in terms of the role they play in the theory's laws; how can a learner discover appropriate concepts and laws simultaneously, knowing neither to begin with? We explore how children can solve this chicken-and-egg problem in the domain of magnetism, drawing on perspectives from computational modeling and behavioral experiments. We present 4- and 5-year-olds with two different simplified magnet-learning tasks. Children appropriately constrain their beliefs to two hypotheses following ambiguous but informative evidence. Following a critical intervention, they learn the correct theory. In the second study, children infer the correct number of categories given no information about the possible causal laws. Children's hypotheses in these tasks are explained as rational inferences within a Bayesian computational framework.

Keywords: Theory change; Cognitive development; Causal learning; Bayesian inference; Magnetism

1. Introduction

Consider the first time a child plays with magnets with no discernable markings such as "north" and "south," red or black. Sometimes one side of an object attracts another, but other sides repel. More, she discovers that these strange objects are attracted to other things, like the refrigerator door, but neither stick to nor repel from superficially similar things, like the cupboard door. What theory can explain these interactions? Inferring the

*Correspondence should be sent to Elizabeth Bonawitz, Department of Psychology, 334 Smith Hall, Rutgers University – Newark, Newark, NJ 07102. E-mail: elizabeth.bonawitz@rutgers.edu

correct theory purely from such observations is a difficult challenge. The child must simultaneously discover how to conceptualize the objects (How many types of sides are there? Which side of each object belongs to each type? How many types of objects are there?) and how to predict their interactions in terms of causal laws defined on these concepts (Do sides of types X and Y repel or attract? Do magnets stick only to other magnets, or also to metals?). Our goal in this paper is to study empirically how children's beliefs might evolve through a process of theory discovery and to understand computationally how they can converge quickly on a novel but veridical system of concepts and causal laws.

The difficulty of discovering the laws of magnetism exemplifies a general difficulty with theory learning: There are often no clear or robust perceptual differences that mark the boundary between the ontological categories necessary to formulate causal laws. You might know that there are two categories (e.g., North, South), but mere inspection would not elucidate which sides are which and what the causal relation is between them. The very concept of a magnetic pole is abstract and theory-internal. Moreover, you may not know the number of categories in advance. You might not initially know that the magnets, the refrigerator door, and the cupboard door all belong to different categories, with different causal roles. We view the challenge of jointly inferring the correct categorical sorting and the causal laws operating over those categories, as the fundamental "chicken-and-egg" problem of theory discovery (Carey, 2009; Quine, 1960).

The chicken-and-egg problem is pervasive for both scientific and intuitive theories: Consider the relation between the concepts "gene," "allele," "recessive," "dominant," and the laws of inheritance in classical genetics, or the relation between concepts of "belief," "goal," and "intention," and the principle of rational action, in intuitive theories of mind. In each case, if you were given the theory's core concepts, it seems clear how you might infer the laws explaining the data, but it is not so clear how the concepts and laws can be constructed together.

Developmentalists have suggested an analogy between cognitive mechanisms of theory change in science and childhood (Chi, 1992; Thagard, 1988; Vosniadou & Brewer, 1992), and there is extensive evidence that children construct intuitive theories in the preschool years (Carey, 1985; Gopnik & Meltzoff, 1997; Keil, 1989; Murphy & Medin, 1985; Wellman & Gelman, 1992). Demonstrating whether and how preschool children can solve the chicken-and-egg problem remains an important empirical challenge.

We ask two related questions: first, can children solve the problem of jointly inferring causal law and category membership in practice; second, how might an intelligent inference engine solve this chicken-and-egg problem in principle? We show that the inference engine solution accounts for the inductive leaps made by children when presented with ambiguous data.

There is a long history in development, exploring children's early causal and scientific reasoning (e.g., Bullock et al., 1982; Bruner, Goodnow, & Austin, 1956; Carey, 1985; Chinn & Brewer, 1998; Piaget, 1930; Schauble, 1990; Shultz, 1982). Most studies of children's causal learning have proceeded by giving the children all the data at once and seeing if they make meaningful inferences from that data. These studies reveal important

clues about children's learning, but they do not necessarily capture critical components of conceptual change including the inference, and restructuring of ontological kinds. To get at these questions of representational change over time, we use a new "mini-microgenetic" method to explore this learning (Bonawitz et al., 2011). We give children evidence in stages and test their beliefs at each stage. This approach provides a more precise way to track the transitions from one belief to the next, following each introduction of new evidence.

We focus on magnetism because it can illustrate the chicken-and-egg problem in children's theory learning, yet it is fairly straightforward to frame. Magnetism can be adapted easily for simplified discovery experiments with preschool-aged children. Moreover, magnetism is a domain that has been extensively studied in the history of science, which may provide a skeletal analogy to children's theory discovery. Our studies are inspired by this historical development.

1.1. A brief history of magnetism

Cognitive historians have explored how scientists form coherent theories from their experience and have connected theory change in science with models of cognition (Gentner, 2002; Gentner et al., 1997; Gruber & Barrett, 1974; Kuhn, 1962; Wiser & Carey, 1983). The phenomena of magnetism provide an excellent case study in both the history of scientific discovery (e.g., see Nersessian, 1992) and the parallel possibilities for how children might come to their intuitive theories.

In 1269, Peregrinus first discovered the two poles of a lodestone and the repulsion between them. He proposed that the spontaneous attraction of lodestones to distinctive geographic directions could be the basis for classifying the stones' poles, but he got the theory wrong—he held that poles of the same type attract (Keithley, 1999).

It took almost another 350 years before William Gilbert reversed this misconception. He posited that Earth was also in fact a large magnet, and hence, the interactions between the poles of Earth and any lodestone should be the same as between any two lodestones (Gilbert, 1600/1958). Then, a critical experiment tested whether poles of the same type attract or repel: bringing together the sides of two lodestones that were both attracted to Earth's north. The fact that these two sides will repel, rather than attract, showed that they should in fact be of the same type as Earth's South Pole. The correct law of magnetic poles dictates that north and south poles attract each other, while poles of the same type repel.

1.2. Toward empirical exploration of the chicken-and-egg problem

Our experiments are inspired by the history of magnetism. In our first study, children are given the initial number of categories but must still infer which objects belong to which categories and what the causal laws are between these categories. First, we test children's beliefs after they observe a limited sample of interactions, analogous to seeing how a set of lodestones interacts with a single reference magnet (e.g., Earth). These

interactions are informative about the correct theory but still ambiguous in a crucial way. Second, we see how children's beliefs change after they observe a critical intervention that should determine a single correct theory. We also look at children's beliefs before they observe any evidence, to reveal the breadth of the theories they entertain and to test for any initial biases. In our second study, we ask whether preschoolers can infer the correct number of categories even without information about the possible causal laws.

We present a model that is abstract enough to generate a broad space of possible causal theories. This generative model uses a probabilistic context-free grammar that provides a broad language for defining multiple possible theories, and an objective syntax for scoring them. We show that the space of theories is vast, but that (a) search over this space is possible, and (b) rational learners are in principle capable of solving the chicken-and-egg problem—discovering the "best" theories—if they can appropriately integrate several pieces of evidence. While we specifically rely on a probabilistic Horn-clause grammar (i.e., a set of rules for building logical expressions that include predicates, their attributes, an relations), we mean it to serve as a representative of the more general approach to theory induction that relies on hierarchical probabilistic program induction (e.g., Lake, Salakhutdinov, & Tenenbaum, 2015; Piantadosi, Tenenbaum, & Goodman, 2016). That is, rather than arguing that Horn clauses are necessarily a component of cognition—we see them as a useful approach toward understanding the chicken-and-egg problem laid out here because it can capture ways to model a probabilistic language of thought.

While previous studies have demonstrated impressive parallels between children's causal learning and ideal Bayesian analyses (e.g., for reviews, see Gopnik, 2012; Gopnik & Bonawitz, 2015; Gopnik & Wellman, 2012; Xu & Kushnir, 2012), they have barely begun to address the hard problems of theory construction. Previous work has not explored how children solve the chicken-and-egg problem of jointly discovering a theory's core concepts and causal laws. In previous work, children are either given the causal laws and required to infer only the correct categories for objects (e.g., Gopnik & Sobel, 2001; Sobel, Tenenbaum, & Gopnik, 2004) or given all (or most) of the key category distinctions for objects and required to infer only the causal laws (Lucas, Bridgers, Griffiths, & Gopnik, 2014; Schulz, Goodman, Tenenbaum, & Jenkins, 2008). Other studies of theory change have examined how older school-aged children gradually revise rule strategies or beliefs following evidence (e.g., Siegler, 1996). Our central questions remain unaddressed: Can children in experimental theory-learning tasks solve the joint inference problem when neither kinds nor causal laws are known, and can this inference be explained as a form of Bayesian computation?

Although we used magnetism as a model for our studies, we constructed a new set of magnet-like objects that follow slightly different and simpler laws. We also included a condition in which the laws were the opposite of the laws of actual magnets: like-sides attract instead of repel. We did this for two reasons. First, real-world theories of magnetism can be quite complex. Second, some preschool-aged children may already have some familiarity with magnets and the true theory of magnetic poles, so we wanted to include a condition that could serve as a "knowledge control" to ensure that our approach was assessing real learning rather than a mere demonstration of prior knowledge.

2. Experiment 1a: Inferring category membership and causal laws

We designed a task loosely based on Gilbert's historical experiments. Children were shown a set of six identical, unlabeled, magnetic blocks that contained *either* a north pole or a south pole on one face of the block. All other sides, including the reverse side of the block, were inert and never participated in any interactions. We labeled two additional blocks, "Yellow" and "Blue." These labeled blocks were placed at opposite ends of a linear frame, analogous to Earth's geographically referenced magnetic field. Children were told that blocks of particular colors might push against other blocks or stick to other blocks but that we did not know exactly how they worked. That is, some information is given to help constrain the space of plausible theories, but the chicken-and-egg remains: inferring which of the unlabeled blocks are Yellow or Blue along with the laws of how Yellow and Blue blocks causally interact.

Next, each of the six blocks was systematically bumped into the two labeled (Yellow and Blue) blocks, to see whether it pushed or stuck, and children were asked to infer the color of each unlabeled block. Following this sorting stage, children were asked to describe how pairs of blocks will interact as a function of their color.

At this stage, many hypotheses about the causal laws between Yellow and Blue blocks are ruled out. However, ambiguity remains. The correct causal law depends on the correct sorting. Specifically, under the correct law of magnetism, one might sort all blocks that *repel* to the labeled Yellow block and *stick* to the labeled Blue as "Yellow" (and vice versa for the blocks that behave the opposite way)—following the rule that objects of the same type repel and objects of different types attract.

However, a second possibility given these data is that objects of the same type *attract* and objects of different types *repel*. Under this hypothesis, one might sort all blocks that *stick* to the labeled Yellow block and *repel* to the labeled Blue as "Yellow" (and vice versa for the blocks that behave the opposite way). We refer to these two hypotheses as the "correct-magnet hypothesis" and the "reverse-magnet" hypothesis, respectively (see Fig. 1, "SPS" and "PSP").

The fact that these two different possible theories remain after this initial evidence helps illustrate how this task goes beyond a simple classification task in which children need only sort objects into categories. Not only must children infer which blocks belong to which of the two groups and identify the correct color label for these two groups; they must also infer the rules about how objects within (and between) these groups interact. The data are perfectly consistent with two different classification schemes, and therefore two different theories about the causal rules that guide interactions between those schemes.

Next, children observe a crucial piece of data, which is analogous to Gilbert's insightful experiment: a single interaction between either two blocks sorted as the same color or two blocks sorted as different colors. Together with the earlier evidence, this last piece of data simultaneously disambiguates the color of all the blocks and the causal laws relating the colors. The single "disambiguating" intervention does not provide sufficient information on its own to infer the correct law. To make this inductive leap, children must be

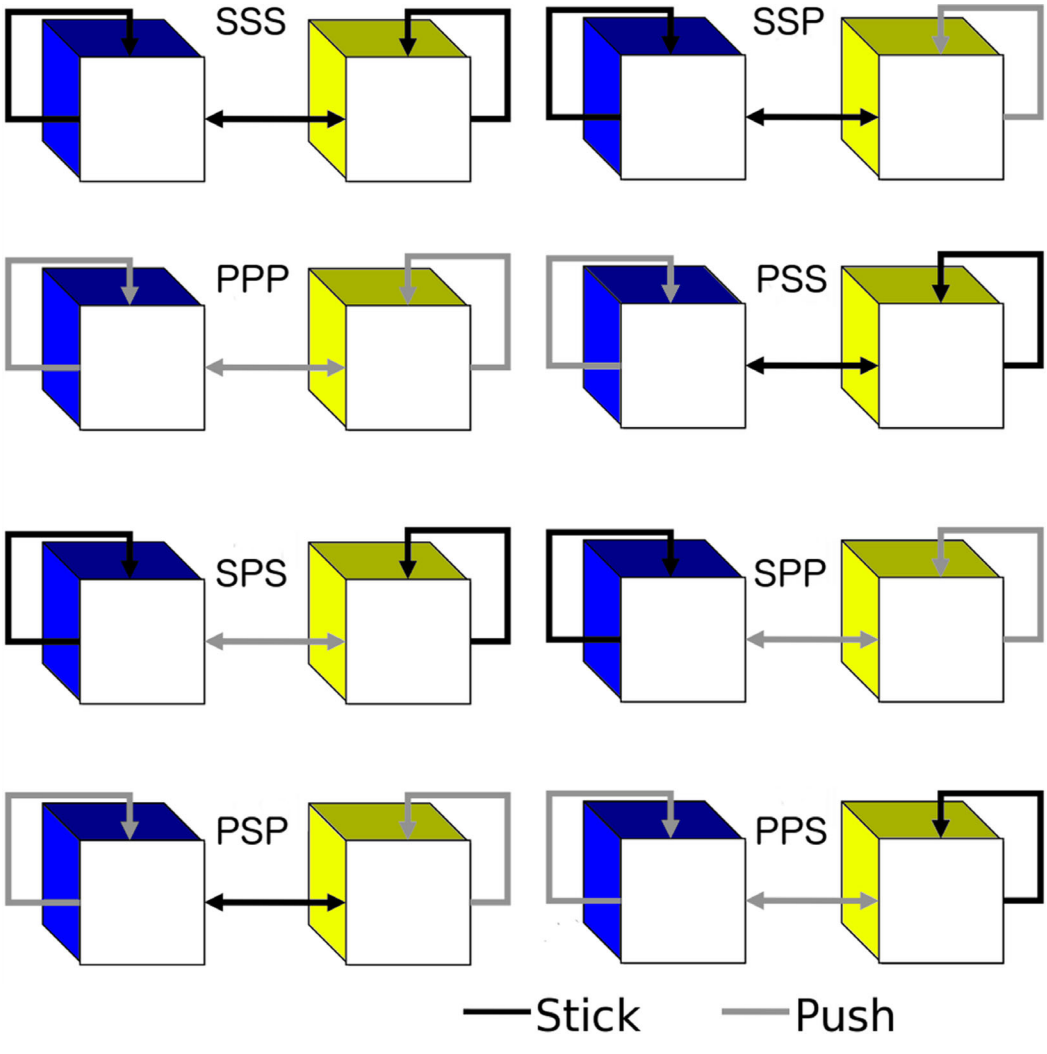


Fig. 1. Eight possible hypotheses for the causal laws in our simplified magnets example. For simplicity, we adopt a naming scheme in which the "stick" relation (represented by the dark black lines) is labeled as "S" and the "push" relation (represented by the lighter gray lines) is labeled as "P." Thus, the hypothesis in the upper left corner labeled SSS specifies that blue sticks to blue; blue sticks to yellow; and yellow sticks to yellow. The hypothesis in the lower right corner, labeled PPS, specifies that blue repels blue, blue repels yellow, and yellow sticks to yellow. Note that the correct hypothesis for magnets is PSP, but the reverse hypothesis (SPS) is also consistent with almost all of the data children observe in our experiment.

able to simultaneously integrate the informative (but ambiguous) evidence in the first trials with the disambiguating evidence in the last trial.

This scenario allows us to examine children's solutions to the chicken-and-egg problem following different stages of evidence. First, we can see what children do when they get

informative but still ambiguous data. We can explore how they update their beliefs and manage uncertainty over hypotheses. At this point, a rational learner should retain two hypotheses, the correct-magnet hypothesis and the reverse-magnet hypothesis, ruling out all others. Second, we can see what children do when they see the results of a "crucial experiment." At this point, they should rationally converge on a single best hypothesis.

2.1. Methods

2.1.1. Participants

Seventy-eight 4- and 5-year-olds were recruited from an urban area science museum (M = 59 months, range = 47–73 months). Approximately half[1] of the participants were female and a range of ethnicities proportional to urban populations in the U.S. Northeast were represented.

2.1.2. Design

All children participated in the ambiguous evidence phase, which involved a sorting task and a theory prediction task. After collecting data from 28 children, we added the disambiguating evidence phase that included a disambiguating evidence event and second theory prediction task. Thus, the following 50 children participated in both the initial ambiguous evidence phase as well as the disambiguating evidence phase. These final 50 children were assigned to either the *Magnet Consistent* condition (n = 30) or the *Magnet Inconsistent* condition (n = 20). One child in the *Magnet Inconsistent* condition and two children from the *Magnet Consistent* condition were dropped because they failed to complete the experiment.

2.1.3. Procedure

Ambiguous evidence phase: Children were shown a 4 in × 2 in ×8 in stand covered in felt that had a Yellow block (1.5 in^3) at one end and a Blue block (1.5 in^3) at the other and were told that Yellow and Blue blocks might push or stick to one another or to a block of the same color. The child's job was to help figure out how the blocks worked. For the sorting task, the experimenter brought out the six identical red blocks and told children, "See these blocks? They lost their Yellow and Blue covers, so we need your help figuring out which blocks are Yellow and which are Blue." The experimenter then picked up the first block and showed that it pushed against the [yellow, blue] labeled block and stuck to the [blue, yellow] block, and then asked, "What color do you think this block should be?" The experimenter then followed the same procedure with the remaining five blocks, selecting the next block in a pseudorandom order[2]. After children generated a response, the experimenter placed the block to one side of the table or the other (depending on the child's label, the experimenter sorted all the blocks that the child labeled as "Yellow" together and all the blocks that the child labeled as "Blue" in a separate pile).

The ambiguous theory prediction task followed the sorting task; children were asked: "What do you think would happen if two yellow blocks bumped together, would they push

or stick to each other? What if two blue blocks bumped together, would they push or stick? How about if a yellow and blue block bumped together, would they push or stick?"

Magnet consistent condition: Following the ambiguous evidence phase, those children who went on to complete the disambiguating evidence phase were told, "Okay let's see what would happen if we took two blocks and bumped them together." All children observed just one interaction: approximately half of the children observed two blocks from the *same* pile (sorted by the children), and the other children observed two blocks from the *different* piles interact (blue–yellow). During the interaction, the experimenter also described what was happening (e.g., "these two blocks stick!"). Children were then asked all three theory prediction questions again, which, critically, included the other two unobserved interactions.

Magnet inconsistent condition: The *Inconsistent* condition was identical to the *Consistent* condition with one exception: rather than the blocks behaving as predicted by actual magnetism, the experimenter manipulated the blocks so that the reverse result was "observed"—if two blocks were taken from the same pile, the experimenter pushed the blocks together (against their natural repelling force) such that the blocks appeared to stick, and if two blocks were taken from different piles, the experimenter retracted the blocks as they came close together so that they appeared to repel. At the end of both conditions, children were asked whether they knew what a magnet was and whether they thought these blocks were like magnets[3].

2.2. Results and discussion of Experiment 1a

Responses were coded by a research assistant and all responses uniquely and unambiguously fell into one of two groups ("Yellow" or "Blue" during the sorting task; or "Stick" or "Push" in the theory tasks). A portion (~40%) of the responses was also coded by the first author; reliability was 100%. There was no effect of sorting order or age on responding. For visual ease in our figures, we report children's theories following a simple binary coding scheme, where P = push/repels and S = sticks/attracts, and the ordering is given by [blue-blue; blue-yellow; yellow-yellow]. Thus, the correct-magnet theory would be written as (PSP) and the reverse-magnet theory would be written as (SPS). See Fig. 1.

2.2.1. Sorting

We coded whether children sorted the unlabeled blocks according to a magnet-consistent rule or a magnet-inconsistent rule. Most children (84%) sorted at least five of six blocks according to one of the two patterns, our criterion for success, and of these 84%, most sorted all six blocks consistently (also 84%). Of the 84% of children who sorted according to one of the two rules, almost all sorted according to the magnet-inconsistent rule (94%), only a handful sorted according to the magnet-consistent rule—a point we return to in the modeling results.

2.2.2. Ambiguous evidence theory prediction

Following the observation of ambiguous evidence, most children generated only the two theories that were consistent with the ambiguous evidence, the correct-magnet theory (PSP) and the reverse-magnet theory (SPS). Children generated both of these hypotheses above chance (correct-magnet: one-tailed Binomial test (n = 15/75), $p < .043$; reverse-magnet: one-tailed, Binomial test (n = 35/75), $p < .0001$). (See Fig. 2, column 1.).

2.2.3. Final theory prediction

Children in both conditions also learned from the final intervention trial, generating significantly different (and evidence-consistent) responses (*Magnet Consistent moving from 6/28 PSP responses to 13/28*: Fisher's exact one-tailed test (28), $p = .045$; *Magnet Inconsistent moving from 9/19 SPS responses to 15/19*: Fisher's exact one-tailed test (19), $p = .046$). That is, even though preschoolers observed just one of the three interactions, the single observation was sufficient to inform their predictions about the other two interactions; children were more likely to generate the correct-magnet theory in the *Magnet consistent* condition and children were more likely to generate the reverse-magnet theory in the *Magnet Inconsistent* condition, $\chi^2(2, N = 47) = 21.71$, $p < .0001$ (Fig. 2, columns 2 and 3).

3. Experiment 1b: Establishing the priors on children's theories

Although children certainly learned following the single disambiguating intervention in the final learning phase, it is possible that the children's pattern of responses in the earlier

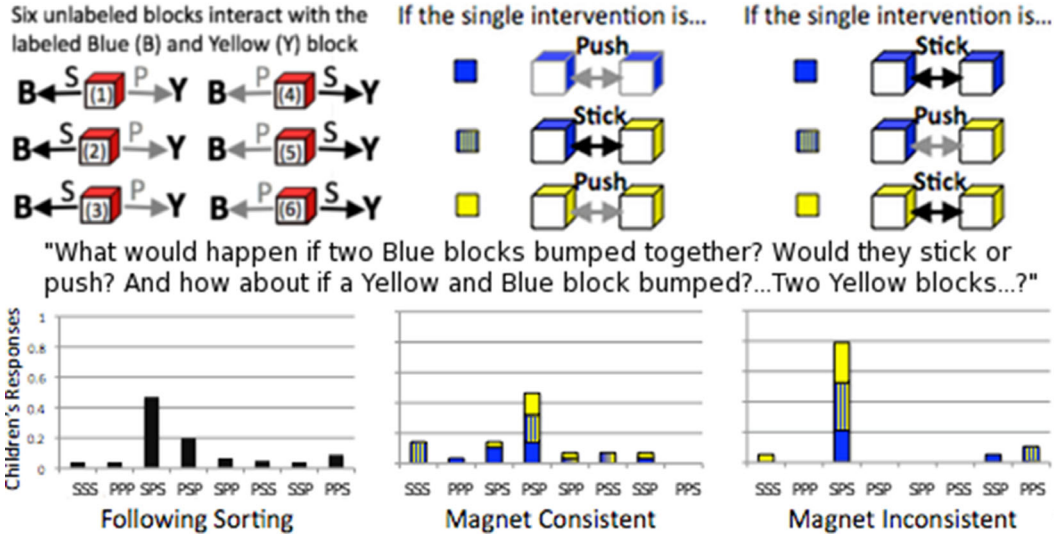


Fig. 2. Children's predictions from Experiment 1a, after seeing the six unlabeled blocks interact with the Yellow and Blue blocks (left); after observing the final disambiguating intervention that is magnet consistent (center) or after observing the final disambiguating intervention that is magnet inconsistent (right).

ambiguous evidence phase reflected a prior bias to follow a response pattern that matched the magnet-inconsistent and magnet-consistent rules, rather than genuinely learning from the evidence in this initial phase. We can rule out this deflationary account by asking whether, prior to seeing any evidence, children entertain a range of theories or only entertain the magnet consistent and inconsistent theories.

3.1. Methods

3.1.1. Participants

We tested a new group ($N = 20$) of 4- and 5-year-olds recruited from an urban area daycare ($M = 53$ months, range = 47–62 months). ‘ were female and a range of ethnicities proportional to Northern California urban populations were represented.

3.1.2. Procedure

“Push,” “Stick” Language Check. Children were introduced to four items (a toy hippo, toy onion, toy ball, and toy eggplant). Children were told that sometimes things push and some other things stick. The experimenter manipulated the hippo and the onion and acted as if the objects repelled, by rapidly pushing away the items from each other when they drew close together, and described the items as “pushing.” The experimenter then manipulated the ball and eggplant and acted as if the objects attracted, by rapidly moving the objects toward each other as they drew closer with difficulty pulling them apart, and described the items as “sticking.” This manipulation ensured that children understood what “stick” and “push” meant, as well as provided children with the same number of stick and push relations observed in Experiment 1a.

Children were then shown two yellow and two blue blocks and were asked (in random order) the theory question: “What do you think would happen if two yellow blocks bumped together, would they push or stick to each other? What if two blue blocks bumped together, would they push or stick? How about if a yellow and blue block bumped together?” Children’s responses for each of the pairs were recorded.

3.2. Results and discussion of Experiment 1b

Preschoolers entertained a variety of theories. Overall, children favored the “sticking” property in their responses, a point we will return to in the modeling results. However, the distribution of theories given by the responses did not differ from chance (Chi-square goodness of fit, $\chi^2(7, N = 20) = 10.4$, $p = .167$). Contrasting the distribution of hypotheses to Experiment 1a revealed significant differences between Experiments (Pearson Chi-square, $\chi^2(7, N = 94) = 19.9$, $p = .006$), suggesting that children’s responding in the ambiguous evidence phase of Experiment 1a reflected genuine learning from the initially ambiguous but still informative evidence (see Fig. 5, left bottom panel).

We note two potential population differences between Experiment 1a and Experiment 1b, which might call into question the applicability of this control. The first is that children in Experiment 1a were tested in a children’s center at a science museum while the

children in Experiment 1b were tested in a daycare setting. However, despite different testing locations, we believe these samples to be relatively well matched. Participants recruited at the daycare came from a similar socioeconomic status to those tested in the science museum. Specifically, daycare participants were drawn from a middle-class pool (typically academic parents on a university campus), which matches to samples drawn from the science museum setting.

A second difference is that the age of participants in Experiment 1b was significantly younger than the mean age of participants in Experiment 1a. One might be concerned that perhaps the younger participants in Experiment 1a were actually "noisy" performers matching the "no bias" younger children in Experiment 1b. It could be that older children carry the positive result in Experiment 1a.

To test for this possibility, we looked at the age-matched sample of participants from Experiment 1a ($N = 32$, $M = 53.5$ months) to see whether these children performed "noisy" sorts (as would be predicted under the alternative explanation). Critically, we found that this group of age-matched children also performed as well as the originally reported, complete sample (sorting with 81.3% consistency). Furthermore, this subset of younger children did not significantly differ in performance from the older subset ($n = 44$) of children, two-tailed Fisher's exact test, $p = .75$. Finally, contrasting the distribution of hypotheses between Experiment 1b and the matched-to-age subset of Experiment 1a revealed significant differences between groups, Pearson's $\chi^2(14, n = 52) = 15.36$, $p = .03$. This suggests that even controlling for age, the performance of participants in Experiment 1a could not be explained by prior bias.

Thus, taken together, our empirical results suggest that in practice, even preschool-aged children can solve the "chicken-and-egg" problem at least in simplified contexts.

4. Experiment 2: Inferring the number of categories

In Experiment 1, the total number of categories (two) was given to children, although they still had to infer which blocks belonged to each category and what the causal laws were between categories. In our second experiment, we were interested in whether children could correctly infer the total number of categories, even without knowing about the causal laws or how to sort the objects into categories. This question is particularly interesting because some modeling work on the "blessing of abstraction" (Goodman et al., 2011) suggests the possibility that learners might actually be able to infer this higher order abstract information even before they learn all the details of the causal laws themselves.

Answering this question required stimuli that could behave more "flexibly" than our magnetic blocks from Experiment 1, so rather than objects that stick and repel, we created blocks that could light up, depending on our experimental needs. In particular, we could predetermine the number of categories (two vs. three), the sorting of objects into those categories, and the causal rules (e.g., whether objects of certain categories should light objects from other categories). Critically, children were not aware of how many

categories of objects we had predefined but simply observed whether pairs of blocks would light up or not based on these predetermined factors.

The particular pattern of lighting was designed to parallel the real behavior of magnets, even though the objects themselves were nonmagnetic and responded causally by "lighting." Note that in Experiment 1, the relevant two categories (yellow and blue) paralleled the north and south poles of a magnet. However, magnetism also involves a different three-category classification of objects, into magnets, (ferromagnetic) metals, and inert objects. These three categories all interact in different ways, and the goal of this experiment was to explore children's ability to capture this categorical sorting in an analogous causal context.

Thus, the "lighting-up" pattern of our constructed blocks paralleled the interactions of actual magnets, metals, and inert objects. Thus, if two blocks were (secretly) labeled as "magnets," both would light (just as two magnets exert a force on one another); if one was a "magnet" and one was a "metal," both would light (because magnets exert a force on metals); but two "metals" would not light each other, and any block interacting with an inert block would not light. After the evidence, children were asked to sort the blocks into one, two, or three groups.

By varying the total number of categories (e.g., whether there were only two categories—just magnets and inerts, or three—magnets, metals, and inerts), we could investigate whether children could learn this higher order information from data alone. Furthermore, by constructing an analogous causal system using the lighting response, we could test whether children could learn this theory, without actual expectations or biases that could derive from experience with real magnets.

4.1. Methods

4.1.1. Participants

Twenty-eight 4-, 5-, and 6-year-olds were recruited from bay area preschools and children's science museums (M = 62 months, range = 49-80 months) and were randomly assigned to either a *3-Category* condition or a *2-Category* condition. Approximately half of the participants were female and a range of ethnicities proportional to urban populations in Northern California were represented. An additional six children were tested but dropped in the *3-Category* condition (two children did not finish the experiment because the session ran over the allotted testing time, the blocks malfunctioned for one child, and the experimenter made methodological errors—deviating from script—for three children). An additional six children were tested but dropped in the *2-Category* condition (two children self-terminated the experiment before completion, the blocks malfunctioned for two children, the experimenter made an error for one child, and one child's session was interrupted by a classmate).

4.1.2. Stimuli

A large cardboard equilateral triangle divided into four smaller triangles (each a different color) was used as a sorting mat.

Sorting warm-up task: Three sets of seven objects were used in the sorting warm-up task designed to get children familiar with a sorting task in general and also to give children practice grouping items by an intrinsic property (rather than a superficial property such as color). The number of different correct groups was varied to give children experience with the task structure without biasing them to infer that objects should always be grouped into a specific set-size. For the 1-grouping set: Seven small wooden blocks with magnets attached to the bottom so that they all could be used to pick up a metal washer[4]; 2-grouping set: Seven plastic eggs that were either filled with sand or were empty inside so that some made noise and others did not when shaken; 3-grouping set: Seven plastic balls filled with pennies so that they were light, medium, or heavy in weight. All of the objects were decorated with a unique identifying color and pattern (e.g., blue with yellow stripes, or green with black polka dots). There was no relationship between the colors and the object properties.

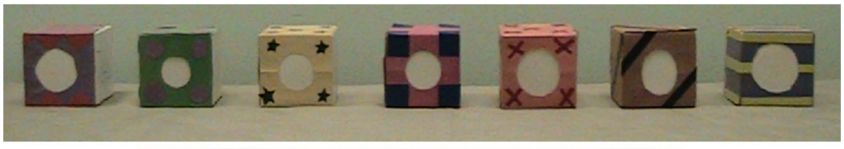
Sorting test trials: Seven 2.5" by 2.5" blocks, each with a LED light inside and decorated with a unique identifying color and pattern that did not correlate with their activation group, were used in the sorting test trials. (See Fig. 3.) When the experimenter bumped two of these blocks together, she could surreptitiously activate the lights with a switch located at the back of the block so that it appeared to the children that the bumping action was causing the blocks to light up. Memory cards with pictures of each block pair either lighting up or not lighting up were used to help children remember the pattern of activation.

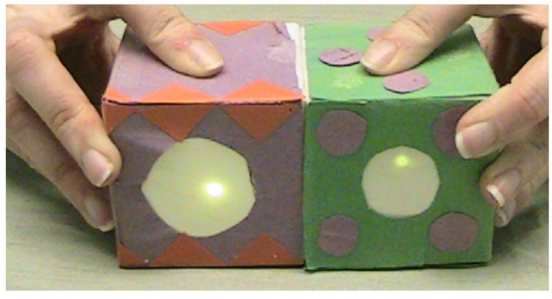
4.1.3. Procedure

Sorting warm-up task: Children were asked to help figure out how different objects worked and which objects belonged together so that the experimenter could later teach her friend about the objects. The experimenter took out the sorting map and told children to sort objects into one, two, or three groups based on what *kind* of object they were and *not* based on their appearance. Feedback was given when needed so that children's final sorts reflected the appropriate number of groups for each set. The order of presentation of the training sets was randomized across participants.

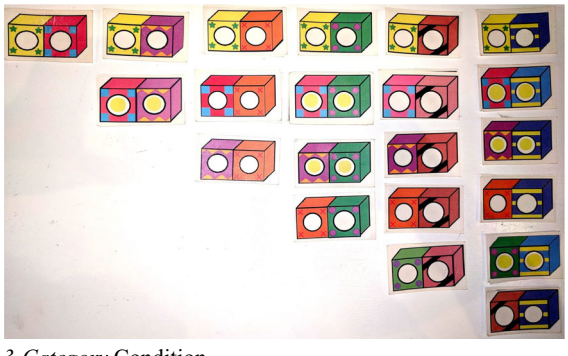
Sorting test trials: After the warm-up task, the experimenter aligned the seven test blocks in a pseudorandom order. The experimenter explained that some of these blocks lit up when bumped together and others did not. Children were asked to help figure out how these blocks worked and which blocks belonged together. Children observed as the experimenter proceeded to bump each block together with all of the other blocks for a total of 21 interactions. After each bump, the experimenter remarked, "See that? Both blocks [lit up, did not light up]" and placed a memory card on the table depicting what had happened. (See Fig. 3 for details on stimuli and the pattern of interactions as given by condition.) After bumping each of the blocks together, the experimenter asked children to identify the pairs of blocks that had lit up. Children were encouraged to refer to the memory cards for assistance.

(a)





(b)



(c) *3-Category* Condition

	Magnet	Metal 1	Metal 2	Metal 3	Inert 1	Inert 1	Inert 3
Magnet		Light	Light	Light	Not lit	Not lit	Not lit
Metal 1			Not lit	Not lit	Not lit	Not lit	Not lit
Metal 2				Not lit	Not lit	Not lit	Not lit
Metal 3					Not lit	Not lit	Not lit
Inert 1						Not lit	Not lit
Inert 2							Not lit
Inert 3							

(c) *2-Category* Condition

	Magnet 1	Magnet 2	Magnet 3	Magnet 4	Inert 1	Inert 2	Inert 3
Magnet 1		Light	Light	Light	Not lit	Not lit	Not lit
Magnet 2			Light	Light	Not lit	Not lit	Not lit
Magnet 3				Light	Not lit	Not lit	Not lit
Magnet 4					Not lit	Not lit	Not lit
Inert 1						Not lit	Not lit
Inert 2							Not lit
Inert 3							

Fig. 3. Procedures and design of Experiment 2. (a) Depiction of the blocks and of two blocks lighting. (b) Example placement of memory cards for the *2-Category* Condition. (c) Tables of Activations: *3-Category Condition* activations: one "Magnet" Block; three "Metal" Blocks; three "Inert" Blocks; *2-Category Condition* activations: four "Magnet" Blocks; three "Inert" Blocks.

Next, the experimenter asked children if they thought there were one, two, or three kinds of blocks. Children's responses were recorded. Finally, children were asked to sort the blocks into groups, placing blocks that were the same kind of block together. As with the training objects, children were instructed to group the test blocks based on what happened when they bumped together and not based on how they looked. The experimenter initiated the sorting process by putting one block onto the sorting mat and then asking children whether a second block: "Is just like this one (pointing to the first block) and should go in the same pile, or do you think it's different and should go in a new pile?" The experimenter placed the block where children indicated and then repeated with the remaining blocks until all blocks were sorted into one, two, or three groups. After all of the blocks were sorted, the experimenter asked children if they wanted to change any of the groupings and allowed them to move blocks around accordingly. Children were allowed to refer to the memory cards while sorting the blocks.

4.2. Results and discussion of Experiment 2

Children were more likely to say there should be three groups in the *3-Category* condition ($M = 2.64$ groups; number of children sorting into three groups = 9/14) than they were in the *2-Category* condition ($M = 2.07$ groups; number of children sorting into two groups = 11/14), $t(26) = 3.1$, $p < .01$, two-tailed Fisher's exact test, $p_a = .018$; $p_b = .012$. Children were more likely to select three categories in the *3-Category* condition than predicted by chance (chance = .33 because children could sort into 1, 2, or 3 groups; two-tailed Binomial test, $p = .03$) and were more likely to select two categories in the *2-Category* condition than predicted by chance (two-tailed Binomial test, $p = .001$). See Table 1.

Although children were able to infer the correct number of groups by condition, children had difficulty accurately sorting the blocks into appropriate groups. For both conditions, we computed a Rand index, which provides a measure of similarity between how data are clustered (e.g., a measure of accuracy) that can be applied when class labels are not used. This is given by considering the pairs of elements that were correctly grouped (and not grouped) together versus the pairs that were incorrectly grouped (or not grouped). See Rand (1971) for details. We also computed a Rand index score for 15 artificial samples for each condition, assuming children were responding completely at chance. Comparing children's Rand index scores for each condition to the index scores obtained by samples of random guessing revealed that children in the *2-Category* condition sorted the blocks significantly better than the chance sample ($t(27) = 3.72$, $p < .001$); however, children in the *3-Category* condition were not significantly better than the chance sample ($t(27) = .97$, $p = .341$).

Although children were able to infer that blocks should be sorted into three categories in the *3-Category* condition, they had some difficulty sorting the blocks appropriately. This result is consistent with the "Blessing of Abstraction" (Goodman et al., 2001). The demands of simultaneously tracking the specific category membership of seven unique blocks may have also been too cognitively demanding for our preschoolers.

Table 1
Number of children by condition sorting the blocks into 1, 2, or 3 groups

Condition	1 Group	2 Groups	3 Groups
3-Category	0	5	**9**
2-Category	1	**11**	2

Bolded Ns highlight the majority sort type by condition.

Results of Experiments 1 and 2 suggest that children can solve the chicken-and-egg problem, at least in somewhat simplified contexts. We now turn to an investigation of how such learning may be solvable *in principle* following a generative Bayesian framework and comparing results to our specific models.

5. Models of theory change

There is a rich tradition in Cognitive Science of understanding theory change in physical domains by building models of this process, and in particular in building frameworks to search for algorithms (e.g., see Langley, 1981). Here, we focus on Bayesian models, as they can help explain how children's learning dynamics relate to the behavior of an ideal learner, who updates prior beliefs rationally in light of observed statistical evidence. They can also explain how challenging learning problems may be solved in principle. While there has been growing interest in neurally inspired models, these approaches cannot capture the key component of theory change discussed here, namely the search process that can model step-by-step changes in beliefs as each new evidence is acquired. Thus, we turn to a stochastic search approach that can operate over a logical representation of rules to explore the degree to which the model follows the same overall dynamics as observed in our empirical studies.

6. Solving the chicken-and-egg problem in a Bayesian framework

Kemp, Tenenbaum, Griffiths, and colleagues have presented a Bayesian account for how simple theories can be acquired (Kemp et al., 2006, 2010). Their probabilistic generative model for relational data applies to the chicken-and-egg problem. Building on this framing, Ullman et al. (2012) proposed a Bayesian grammar-based model of theory acquisition that simultaneously learns logical laws and the extension of the concepts related by these laws. Our experiments were partly inspired by this work, and our modeling is based mainly on Ullman, Goodman, and Tenenbaum (Ullman et al., 2012). In this section, we describe our approach to solving the chicken-and-egg problem. We give a general overview of the model in minimal technical terms, and then explain how the model relates to the specific learning tasks presented here around which the analysis is structured. The full formal details of the model are detailed in the Appendix S1, and a complete technical characterization of a more general model is to be found in Ullman et al. (2012).

6.1. Overview of knowledge representation as a hierarchical probabilistic structure

While the proposal that children and adults represent knowledge as "intuitive theories" is not new, it was only in recent years that this notion began to be accessible to computational modeling, through the synthesis of structured symbolic logic and probabilistic inference methods (see Tenenbaum, Kemp, Griffiths, & Goodman, 2011 for a review, as well as Gerstenberg & Tenenbaum, 1600/1958 for a recent probabilistic programming treatment). Following this approach, we assume that learning agents can represent knowledge in terms of compositional symbols, and that intuitive theories are sets of laws governing concepts, organized in a predicate logic. As an example of such a law, consider "objects of type *Magnet* attract objects of type *Metal*." An encapsulated set of laws and concepts that applies to a particular domain is termed a *theory* of that domain. As a relevant example, the set of concepts Magnet, Metal, and Inert, together with the laws for attraction and repulsion, form the intuitive theory of the domain of simplified magnetism.

We formalize concepts in an intuitive theory as logical predicates, such as Magnet(X). We formalize laws in a theory as logical clauses, which relate the predicates that represent concepts. We use a particular logical clausal form known as a Horn clause (Horn, 1951). Such Horn clauses always take the form of a conjunction of predicates that imply one other predicate, $q \leftarrow (p_1 \wedge p_2 \wedge \ldots \wedge p_n)$. In our case, an example of a law in an intuitive theory in Horn clause form would be attracts(X, Y) ← Magnet(X) ∧ Metal(Y), expressing the law "If X is a Magnet and Y is a Metal, then X will attract Y." Horn clause logic has been highly influential in the field of logic-based programming and logic-based AI, forming the basis for the Prolog language (Kowalski, 1979). Horn clauses have also proven useful in capturing intuitive psychological theories and causal relations (e.g., Katz et al., 2008; Kemp et al., 2010; Ullman et al., 2012)[5].

A domain theory is a set of predicates and laws relating those predicates in Horn clause form. However, a domain theory does not generate data directly. We must first assign objects to concepts. For example, consider the law "Attracts(X, Y) ← Magnet(X) ∧ Metal(Y)." Suppose we specify that A is a magnet and B is a metal, our law predicts that A and B attract. However, if instead we specify that C is a magnet instead of A, we will predict that C and B will attract. Different data can be produced by the same law, depending on object assignment. By assignment of object to predicate, we mean the specification of an object for which a predicate evaluates to *true* when applied to that object. A complete assignment of objects to predicates is called a *model*. So, even if a learner has the right *theory*, he or she still needs to propose the correct *model*.

At the theory level, the learner can generate new theories by using a probabilistic generative grammar. In the same way that an English grammar can generate new sentences and be used to parse observed sentences, the grammar for theories generates laws and predicates, and eventually (through assignment of objects to predicates) observable data. The space of all possible theories that can be generated by this grammar is termed the *universal theory space*, and for any reasonably complex grammar, it is very large or even infinite.

Not all theories are equally likely to be generated by a grammar, and the probability of a particular theory being generated a priori of any data is assigned a prior probability. The prior probability assigned to a theory as a whole is determined by the multiplication of the individual probabilities of producing the specific laws and predicates that make up the theory. That is, the production rules that specify the grammar and generate the laws and predicates that make up a theory are each associated with the probabilities of generating the different items within the production rule. In this paper, we consider two overall priors over theories. The first is a *Generic Prior*, which assumes a uniform distribution over the production probabilities within production rules. We also consider a *Stick Bias* model, in which the production probability of the *Sticks* predicate is higher than the production probability of the *Repels* predicate. We consider this model because a stick bias is consistent with the history of early magnetism theories and seems intuitively plausible if only because the attractive powers of magnets are more salient than their repulsive properties. Frequencies of naturally observable magnetic phenomena are also biased in this direction: A magnet can generate attractive forces not only with other magnets but with iron and other metals; in contrast, only two magnets can generate a repulsive force.

To recap, we consider a structured hierarchical framework, with each level probabilistically generating the level below it (see Fig. 4). At the topmost level, is the space of possible theories (all possible combinations of laws and predicates expressible in the logical form chosen). Below that are theories for given domains (e.g., simplified magnetism). Beneath that still are models (that is, within the theory of simplified magnetism, specifying object 1 is a Magnet, object 2 is a Metal, and so on), which finally generate and predict data patterns of objects in the world.

Using this structure, the learning problem for children and adults can now be clearly stated: Conditioned on a set of observable data, find the theory (set of Horn clauses and predicates) and model (assignment of objects to predicates) that best account for that data. Bayesian reasoning supplies the formal notion of "best account"—the desired theory should both be a priori more likely (provided by the prior probability, the grammar places on the universal theory space, often favoring simplicity—shorter and fewer rules) but also explain as much of the data as possible (provided by the likelihood).

The upshot of such a hierarchical framework is that a learner attempting to explain some observed data is faced with the dual problem of finding the right causal laws that govern the domain (theory), and the particular extension of the predicates over which the theory is defined (model). The meaning of the predicates themselves derives from their extension, in combination with the laws, which captures the chicken-and-egg problem (Carey, 2009; Quine, 1960).

While finding the best theory and model is a clear characterization of the learning problem at the computational level, this procedure is not simple from an algorithmic point of view. Ullman et al. (2012) explored a rational algorithmic approach to this problem by considering Markov Chain Monte Carlo (MCMC) search methods, which consider one theory and model at a time and suggest changes to it by editing and resampling predicates, laws, and extensions. Following Goodman et al. (2008), our algorithm begins with a specific theory t, and then uses the probabilistic Horn-clause grammar (PHCG) to

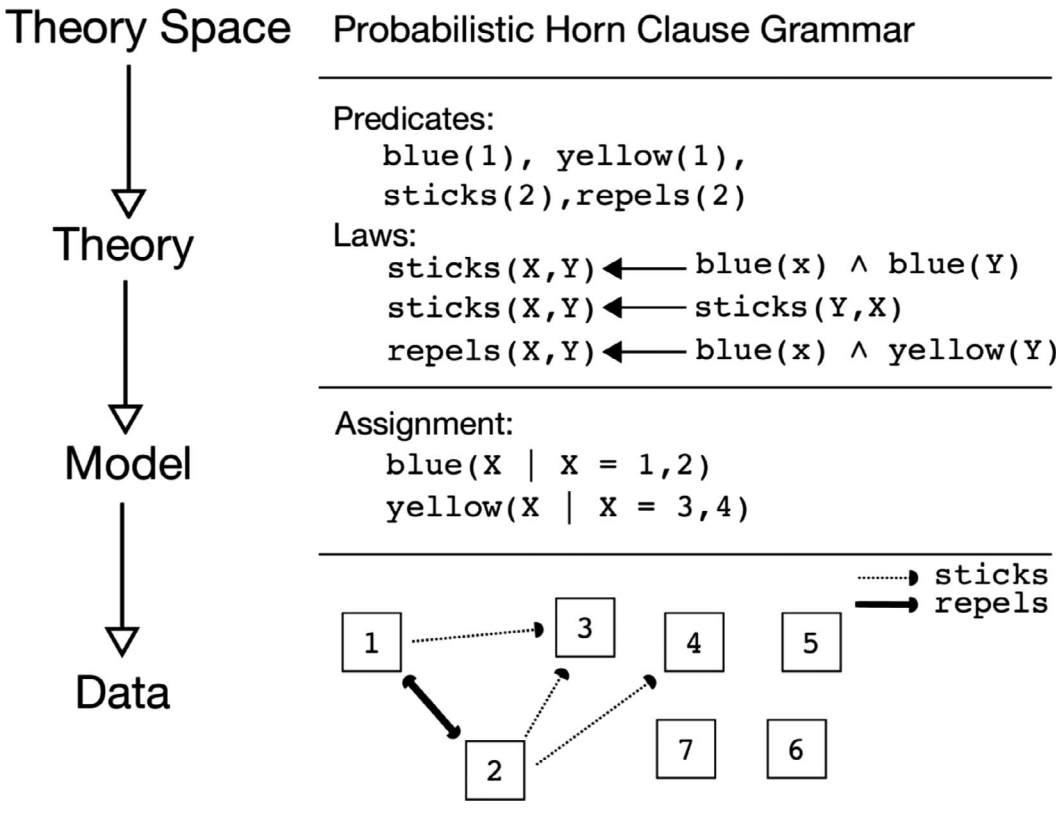


Fig. 4. Representation of the relationship between Theory Space, Theory, Model, and Data in the domain of Magnetism, using Probabilistic Horn-clause Grammars, Predicates/Laws, Assignments, and observed interactions.

propose random changes to the currently held theory by choosing a point along the theory derivation path and regenerating production choices from that point. This new proposed theory t' is probabilistically accepted or rejected, depending on how well the new theory explains the data compared to the currently held theory, as well as how much simpler or more complex it is. Ullman et al. (2012) suggested that this method could explain how practical learners, such as young children, can rationally approximate an ideal Bayesian analysis. This method allows a practical learner to search over a potentially infinite space of theories, holding on to one theory at a time and discarding it probabilistically as new, potentially better alternatives are considered.

The reader is referred to the Appendix S1 for a more in-detail account of the predicate logic, the priors induced by the grammar, and the search process.

6.2. Modeling results for Experiment 1

The *Generic Prior* and the *Stick Bias* modeling results for each phase of the experiment are presented in Fig. 5. Notably, following the ambiguous data, our simulations

were able to discover the SPS and PSP theories and rated these theories as best given the data. Given one observed interaction between two unlabeled blocks which should discriminate between hypotheses PSP and SPS, both models' posterior probabilities $P(h|d^*)$ increase strongly for the correct hypothesis (relative to the data) over the alternative, although the *Stick Bias* model still shows a slight asymmetry in favor of SPS, inheriting from its prior (Fig. 5, columns 3 & 4).

6.2.1. Priors and sorting

We compared children's empirical priors obtained in Experiment 1b to the priors in the models (although the precise profiles of these priors over all eight possible theories cannot be evaluated without a substantial *N*). The *Stick Prior* model predicts more sticking relations than repel relations; the number of children in the *Priors* Experiment 1b showing a bias to favor theories with more sticking relations was similarly higher than chance, two-tailed Binomial test, $p < .041$. The *Generic Prior* does not predict more sticking relations, but it does predict more reuse of "stick" or "repel," favoring the SSS and PPP theories over the others; children did not favor the SSS and PPP (2 of 20 children). Thus, children's responses in the *Priors* Experiment provide initial qualitative support for the *Stick Bias* model over the *Generic Prior* model. Additionally, almost all of the children sorted according to the magnet-inconsistent rule with only a handful sorting according to the magnet-consistent rule; this is also consistent with the *Stick Bias* prior, which favors the magnet-inconsistent rule over the magnet-consistent rule.

6.2.2. Ambiguous evidence

The distribution of children's responses correlated with the *Generic Prior* model ($r^2(73) = .86$), but very highly with the *Stick Bias* model ($r^2(73) = .95$); comparing the significance of the difference between these two correlation coefficients with Fisher two-tailed r-to-z transformation revealed significantly higher values for the *Stick Bias* model's fit to data (z (N's = 75) = −3.23, $p = 001$). This provides additional evidence that the *Stick Bias* model better captured children's intuitions. We also computed correlation for a variety of values on the *Stick Bias*, ranging from .6 to .9; results were robust, with all correlations of $r^2 > .95$. (See Fig. 5, Column 2.)

6.2.3. Final theories

We also compared how children's responses distributed across the theories compared to the model predictions. The distribution of responses in the *Magnet Consistent* condition and *Magnet Inconsistent* condition correlated very well with both models ($r^2(47) > .93$). As with the ambiguous evidence correlations, these correlations were robust across a range of values for the *Stick Bias* model ranging from .6 to .8 ($r^2(47) > .91$); values with the extreme bias of .9 correlated slightly worse ($r^2(47) > .84$) due to the over-favoring of the stick-rule. (See Fig. 5)

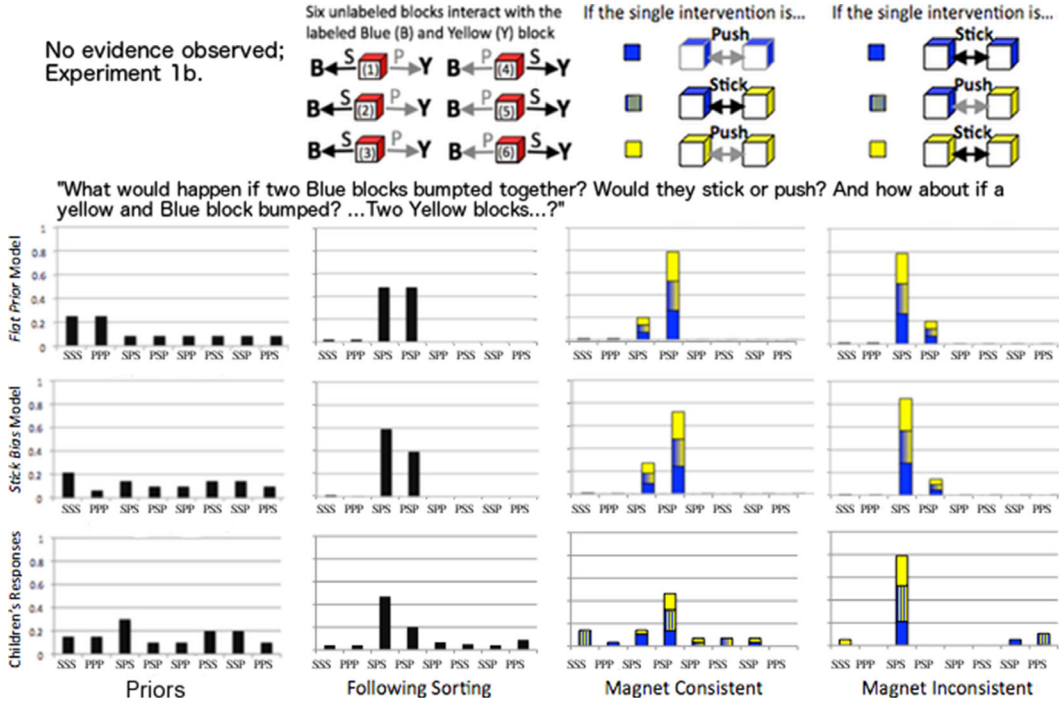


Fig. 5. Predictions of the *Generic Prior* and *Stick Bias* models compared to children's predictions from Experiment 1 before seeing the evidence (*Priors* condition, rightmost column 1); after seeing the six unlabeled blocks interact with the Yellow and Blue blocks (column 2); after observing the final disambiguating intervention that is magnet consistent (column 3) or after observing the final disambiguating intervention that is magnet inconsistent (leftmost column 4). The disambiguating intervention consisted of one of three possible interactions. Darker, blue portions of the bars in column 4 represent the portion of children selecting each theory after observing a blue–blue interaction; the blue-yellow graded portions of the bars represent the portion of children selecting each theory after observing a blue–yellow interaction; and the lighter yellow portions of the bars capture a yellow-yellow observation.

6.3. Modeling Experiment 2: Inferring the number of categories given ambiguous data

All of the theories considered have used the same number of underlying core predicates (meaning the same number of conceptual categories appeared in all theories). However, this is not an imposed requirement, and the actual number of conceptual categories used in the theory will depend on the observed data.

Ullman et al. (2012) theoretically investigated the results of parametrically varying the proportion of interactions between metals and magnets. They showed that when there is a sufficient number of metals and magnets, a rational model mostly discovers theories that use three categories (one theory being a version of the "correct" intuitive theory that makes use of all three categories—Inert, Magnet, and Metal). However, when few metals are present (or equivalently when there are few observable interactions with metals), we should expect the construction of a simplified theory that makes use of only two

conceptual categories—Inert, and a predicate that is a mixture of Metals and Magnets. This simplified theory predicts metals should interact with other metals, and their failures to do so are treated as tolerable outliers, as the "price" for using a new predicate and new causal laws is too high (in terms of a priori probability).

This theoretical result leads to a qualitative prediction regarding the use of two versus three categories, which is confirmed by the results of Experiment 2, as discussed next.

6.4. Qualitative modeling results for Experiment 2

It is not simple to translate the theoretical results of Ullman et al. (2012) into exact quantitative predictions about the proportion of children who would infer theories with two versus three categories, as the analysis relies on a number of free parameters that are hard to measure independently. Furthermore, the theoretical results originally considered 10 objects interacting, while only seven objects were used in the empirical studies reported here, to prevent further memory overload. We can overcome these difficulties by choosing two distinct cases along the spectrum considered by Ullman et al. (2012), and simplifying them such that they include fewer objects. *Case 1* in Ullman et al. originally considered three Magnets, one Metal, and six Inert objects. By removing three Inert objects, we make this case equivalent to the *2-Category* condition[6]. *Case X* in Ullman et al. originally considered one Magnet, seven Metals, and two Inert objects. By removing four Metal objects and adding an Inert object, we can make this case similar to the *3-Category* condition, without changing the underlying principles.

Given these simplifications, we can now draw a clear qualitative prediction—in the case of few metals (*Case 1/ 2-Category* condition), we would expect more children to infer a simplified theory with only two categories and fewer causal laws. As the number of metals increases (*Case X/ 3-Category* condition), we would expect more children to infer a theory with three categories.

7. General Discussion

We presented a case study in how preschool-aged children solve the chicken-and-egg problem of theory learning, jointly identifying causal laws and the hidden categories they are defined over. Although our research was not designed to capture the radical restructuring associated with some forms of conceptual change (e.g., Carey, 2009), our work goes beyond work that is simply looking at belief confirmation. Specifically, we tested the correspondence between children's inferences at three phases of discovery in a simplified magnet task and a Bayesian analysis of how ideal learners could solve this task. Our approach demonstrated how this problem can be solved in principle and provides a general framework for describing theory change: a broad language for defining a potentially infinite space of possible theories, an objective syntax for scoring those theories, and an algorithmic search framework for discovery.

In Experiment 1b, we asked whether the distribution over hypotheses that children consider reflects an initial weak bias favoring "attract" relations. We found preschoolers' initial responses could be explained qualitatively, assuming a rational model with a prior that favors the sticking relation. This fit was superior to the *Generic Prior* model. This finding echoes the early magnetism theories of Peregrinus, as well as early historical writings about magnets that are populated with numerous reference to attraction and making new iron objects "like a magnet," but have few references to the property of repelling. The parallels between the biases we found with preschoolers and those found in historical analysis highlight a difficulty in solving the chicken-and-egg problem in magnetism: How can a learner overcome a prior bias that supports incorrect theories?

After showing children interactions that were informative about the correct theory but still ambiguous in a crucial way (analogous to the naturally occurring evidence in the history of magnetism), we asked whether they would demonstrate an appropriate but focused uncertainty, restricting beliefs to only those hypotheses consistent with the data but effectively drawing on priors as well as observations to set their posterior degrees of belief. Results from Experiment 1a suggest that even preschool-aged children can respond rationally to such ambiguity, favoring the correct two hypotheses over the others initially considered prior to observing the data. However, only the *Stick Bias* model qualitatively captured children's mild preference for the "reverse-magnet" theory over the correct theory of magnetism, inheriting from its prior preference for hypotheses with more "stick" relations.

Finally, we asked whether, analogous to William Gilbert's classic studies, a single critical intervention between just two blocks could lead children to a strong belief in a single hypothesis, simultaneously disambiguating the hidden category identities of objects and the causal rules between categories. We found that even 4-year-old children were sensitive to this single intervention and were able to infer an appropriate causal law (the magnet rule following one kind of evidence, and the reverse-magnet rule following the other.). Even in the course of a short experiment, preschool-aged children were able to solve a simple version of the chicken-and-egg problem in a basically rational way—integrating multiple pieces of evidence across different phases of the experiment. We suggest that these same inference capacities help to drive theory change in the normal course of children's cognitive development.

In Experiment 2, we asked whether children could infer the total number of categories based on ambiguous data and no information about the causal laws. Although children had difficulty sorting the blocks in the *3-Category* condition, they were able to infer the correct number of categories for both conditions, suggesting that at some level, children were able to abstractly infer the total number of groups. Children's behavior was qualitatively predicted by a Bayesian model, which specified a rational response to the data.

The hierarchical model we presented here relies on stochastic search over logical laws and predicates that form an "intuitive theory" of a domain. The search for a compressed logical representation for explaining observed data has been studied mainly under the heading of Inductive Logical Programming (ILP, for a review see De Raedt & Kersting,). ILP has not been the main focus of the recent boom in Machine Learning, but it is

receiving increased attention due to its human-readability, ability to handle much smaller training sets, and generalizability (e.g., consider the difference between a program containing the few lines of code that generate all and only even numbers between 1 and 1000, and the neural network backend required to achieve the same output). In particular, hybrid approaches have recently been proposed in which the learned representation is made of logical laws and predicates, but the program is differentiable in that it is also simultaneously able to run forward chaining and thus is amenable to current gradient-search techniques (see in particular Evans & Grefenstette, 2018). While we do not mean to rule out gradient-based approaches to learning, these new hybrids still rely on prespecifying the possible predicates and their extensions. They are also more limited than a grammar-based approach, and in that sense have difficulty scaling up and handling the hard problem of search faced by children.

While many of the "moving parts" of our approach have been proposed before in isolation, we see the contribution of the formal framework here as made up of the sum of its parts, and its application to children's learning. That is to say: While hierarchical, probabilistic grammar- and logic-based models for knowledge representation have been proposed and studied for adult and machine intelligence (e.g., Kemp et al., 2010, Tenenbaum et al., 2011), and while the general approach of Bayesian, theory-based reasoning in children has also been proposed (e.g., Gopnik, 2012), the actual nuts-and-bolts implementation and investigation of intuitive theory acquisition and its dynamics in a microgenetic case with a grammar-based model that can allow for many potential theories in an open-ended way is an exciting current direction. With the growing interest in the Machine Learning community in building machines that learn like children (Lake, Ullman, Tenenbaum, & Gershman, 2017), we think it is useful to propose how that learning could work, in formal terms.

The strength of our model is in its general applicability to domains that require reasoning in the form of intuitive theories. However, we readily accept that the full mechanics may appear too cumbersome for any one given domain. For example, while our model is able to construct a stick-biased domain theory for simplified magnetism, surely a simpler model could have discovered the same without the need to propose logical rules from a context-free grammar over all Horn clauses. The argument for simpler models has merit, but only in a domain-limited fashion. By analogy, consider a model for predicting whether a tower of blocks is stable (cf. Battaglia et al., 2013). One computational proposal (the "mental physics-engine") is that people reconstruct the tower of blocks in their minds, and then run simplified noisy Newtonian physics to examine whether the reconstructed simulated blocks fall. Such a proposal can recapture people's judgment, but at what cost? A much simpler combination of features such as "how tall is the tower" can achieve similar results, and without the cumbersome overhead of having to represent 3D objects and run dynamical equations. However, the force of the richer representation approach is in generality. If one were to ask people *where* the blocks will fall, entirely new features must be constructed and trained, whereas the same physics-engine reconstruction can be run with a different query (see also Lake et al., 2017). This argument by analogy is not meant to convince that a richer representation for children's learning is

necessarily true, but to point to its potential. Further microgenetic investigations in other domains are necessary before the specter of "simpler models" can be ruled out.

How can children have learned the correct rule in the course of our short experiments when historically such theory change can take centuries (e.g., Carey, 2009)? While historical analogies can provide some insight into the difficulties and strategies of intuitive theory discovery (Hacking, 1993; Kitcher, 1988; Kuhn, 1982; Nersessian, 1992), there are several ways in which our study with children was simpler than most cases of theory change in science—and in particular, than the historical case that inspired it. In Experiment 1, for example, children are told in advance that there are just two kinds of objects, and they are shown the critical interventional data without having to come up with the intervention themselves. However, much research shows that the process of designing informative experiments and controlling for variables is difficult for children and adults alike (Klahr et al., 1993; Kuhn, 1989; Schauble, 1990). Indeed, this task could be viewed as requiring "hypothesis search and formation" without requiring "experimental search and design" as in the well-known cognitive models of dual search (Klahr & Dunbar, 1988). Children in our task did not have to be metacognitively aware of how to design informative interventions, although that is an inexorable part of the problem scientists face. Furthermore, the children in our populations grow up in a culture that teaches physical, causally mechanistic explanations, which may scaffold learning in these domains and help children recognize the role of uncertainty in theory-change (e.g., Metz, 2004). In contrast, scientists such as Gilbert were surrounded by magical-spiritual frameworks (Ferngren, 2002), which may have provided satisfactory, deterministic, alternative explanations for hidden forces, such as magnetism[7]. Nonetheless, given the myriad ways that learners can respond to anomalous data (Chinn & Brewer, 1998), it is particularly impressive that preschoolers in our task overwhelming accepted the data and made appropriate changes to their beliefs.

Despite these differences, however, we suggest that our results can take the "child as scientist" analogy to a new level of empirical richness and computational rigor. Children as young as 4 years old can respond reasonably in the face of theory ambiguity, choose approximately rationally between competing hypotheses, and rapidly learn from disambiguating evidence.

The focus of this work has been to demonstrate empirically a case where children can solve the chicken-and-egg problem and to formalize and test a computational-level model (Marr, 1982) for how ideal learning might deal with the ambiguity inherent in this problem. The Bayesian computations needed for theory learning—in particular, those needed to solve the chicken-and-egg problem in the presence of even slightly complicated patterns of data—are complex; it seems likely that children are not exactly implementing all the necessary Bayesian calculations, not even unconsciously or implicitly. This work has started to experimentally explore the connections between algorithmic and computational accounts (e.g., how learners might be carrying out approximate rational inference), via search algorithms such as MCMC, as proposed theoretically by Ullman et al. (2012). Other research has looked at whether learners sample a small subset of hypotheses to evaluate, rather than performing Bayesian inference over a potentially infinite space of hypotheses (Bonawitz et al., 2014;

Bonawitz, Denison, Griffiths, & Gopnik, 2011; Bonawitz & Griffiths, 2010, in review; Goodman et al., 2008; Vul et al., 2009; Vul & Pashler, 2008) and other studies suggest cases where children's causal inferences might also be accounted for by this approach (Bonawitz, Denison, Gopnik, & Griffiths, 2014; Denison et al., 2013). Future work is needed to explore precisely the algorithmic steps by which children might approximate rational Bayesian learners and the cases where children fail to approximate them.

This study is a promising starting point, demonstrating parallels between inferences made by children, those made in the history of science, and those made by a rational Bayesian learner: Children appropriately push away (even a priori likely) hypotheses that are not well supported by the data and stick with the evidence to rationally revise their beliefs.

Acknowledgments

We specially thank Catherine Clark, Sonia Spindt, and Stephanie Denison for data collection and helpful discussion, and Laura Schulz for generous support and feedback. We also thank participating museums, daycares, and families as well as Nicole Brooke, Annie Chen, Madeline Hanson, Jingan Li, Zi Lin, Kathie Pham, Jessica Ho, and Dhaya Ramarajan for additional data collection and Joseph Austerweil and Wolf Vanpaemel for feedback on an earlier draft of this manuscript. This research was supported in part by the James S. McDonnell Foundation Causal Learning Collaborative (AG, EB, JBT), the Paul E. Newton Career Development Chair (JBT), the AFOSR Contract FA9550-1-0075 (JBT), ARO MURI W911NF-08-1-0242 (JBT), NSF Award SES-1627971 (EB), the Jacobs Foundation (EB), and the American Psychological Foundation (EB).

Notes

1. The gender of a few participants was not recorded.
2. The first two trials were counterbalanced (blue, yellow); the remaining trials were randomly dictated by whichever type of unlabeled block the experimenter happened to grab.
3. While the majority of children stated that they had played with magnets previously, almost no children believed that these blocks were like magnets.
4. Although one of the familiarization tasks involved actual magnets, because the test blocks involved lighting we did not expect carryover from the initialization phase. Furthermore, if children somehow were primed by this warm-up phase to consider magnets, we note first that the prime was for one-category grouping (which did not match either experimental test phase) and second that children in both experimental conditions received the same warm-up, so any bias from this initial warmup could not explain differential performance between conditions.
5. We refer the reader to the Appendix S1 for more formal and technical details, but stress here that Horn clauses are one way for capturing causal structure in formal

predicate logic. This does not mean that we are committed to the mind using Horn clause representations, but that it is a useful simplified Language-of-Thought proposal, and we expect other approaches that consider theory learning as stochastic search over logical laws to follow a similar overall dynamics.

6. A matching 2-Category condition would technically include four magnetic block and 0 metals; the data produced by this 2-Category condition is identical to the Ullman et al. (2012) example of three magnetic blocks, one metal block, and three inert blocks.
7. We thank an anonymous reviewer on an earlier iteration of the research for drawing our attention to this explanation.

References

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*, 18327–18332. 201306572.

Bonawitz, E., Denison, S., Gopnik, G., & Griffiths, T. L. (2014). Win-stay, lose-shift: A simple sampling algorithm for approximating Bayesian inference. *Cognitive Psychology*, *74*, 35–65.

Bonawitz, E., Denison, S., Griffiths, T., & Gopnik, A. (2014). Probabilistic models, learning algorithms, response variability: Sampling in cognitive development. *Trends in Cognitive Science*, *18*, 497–500.

Bonawitz, E., & Griffiths, T. (2010). Deconfounding hypothesis generation and evaluation in Bayesian models. In T. Camtrabone & S. Ohlsson (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive* (pp. 2260–2265). Portland, OR: Cognitive Science Society.

Bonawitz, E. B., & Griffiths, T. L. (2010) (in review) Considering psychological mechanisms can change the interpretation of Bayesian models.

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

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), *The developmental psychology of time* (pp. 209–254). New York: Academic Press.

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

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

Chi, M. T. H. (1992). Conceptual change within and across ontological categories. Examples from learning and discovery in science. In N. N. Giere (Ed.), *Cognitive models of science. Minnesota Studies in the Philosophy of Science*, *15*, 129–186. Minneapolis: University of Minnesota Press.

Chinn, C. A., & Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *35*(6), 623–654.

De Raedt, L., & Kersting K. (2008). Probabilistic inductive logic programming. In L. De Raedt, P. Frasconi, K. Kersting, & S. Muggleton (Eds.), *Probabilistic inductive logic programming*. Lecture Notes in Computer Science, vol. 4911. Berlin: Springer.

Denison, S., Bonawitz, E. B., Gopnik, A., & Griffiths, T. L. (2013). Rational variability in children's causal inferences: The sampling hypothesis. *Cognition*, *126*, 285–300.

Evans, R., & Grefenstette, E. (2018). Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, *61*, 1–64.

Ferngren, G. B. (Ed.) (2002). *Science and religion: A historical introduction*. Baltimore: JHU Press.

Gentner, D. (2002). Analogy in scientific discovery: The case of Johannes Kepler. In L. Magnani & N. Nersessian (Eds.), *Model-based reasoning: Science, technology, values* (pp. 21–39). New York: Kluwer.

Gentner, D., Brem, S., Ferguson, R., Markman, A., Levidow, B., Wolff, P., & Forbus, K. (1997). Analogical reasoning and conceptual change: A case study of Johannes Kepler. *The Journal of the Learning Sciences*, *6*, 3–40.

Gilbert, W. (1600/1958). *De Magnete*. Mineola, NY: Dover.

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

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

Gopnik, A. (2012). Scientific thinking in young children. Theoretical advances, empirical research and policy implications. *Science*, *337*, 1623–1627.

Gopnik, A., & Bonawitz, E. (2015). Bayesian models of child development. *Wiley interdisciplinary reviews: cognitive science*, *6*(2), 75–86.

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

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory. *Psychological Bulletin*, *138*, 1085–1108.

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

Gruber, H., & Barrett, P. (1974). *Darwin on man: A psychological study of scientific creativity*. New York: Dutton.

Hacking, I. (1993). Working in a new world: The taxonomic solution. In P. Horwich & J. Thomson (Eds.), *World changes* (pp. 275–310). Cambridge, MA: MIT Press.

Horn, A. (1951). On sentences which are true of direct unions of algebras. *The Journal of Symbolic Logic*, *16*(1), 14–21.

Katz, Y., Goodman, N., Kersting, K., Kemp, C., & Tenenbaum, J. (2008). Modeling semantic cognition as logical dimensionality reduction. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.

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

Keithley, J. F. (1999). *The story of electrical and magnetic measurements: From 500 B.C. to the 1940s*. Hoboken, NJ: John Wiley and Sons.

Kemp, C., Tenenbaum, J., Griffiths, T., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In Y. Gil & R. J. Mooney (Eds.), *Proceedings of the National Conference on Artificial Intelligence* (vol. *3* p. 5). Menlo Park, CA: AAAI Press.

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

Kitcher, P. (1988). The child as parent of the scientist. *Mind and Language*, *3*, 217–228.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*(1), 1–48.

Klahr, D., Fay, A., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology*, *25*, 111–146.

Kowalski, R. (1979). Logic for problem solving (Vol. *7*). Ediciones Díaz de Santos.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, *96*, 674–689.

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Kuhn, T. S. (1982). *Commensurability, comparability, communicability*. East Lansing, MI: Philosophy of Science Association.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, E253. https://doi.org/10.1017/S0140525X16001837

Langley, P. (1981). Data-driven discovery of physical laws. *Cognitive Science*, *5*(1), 31–54.

Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284–299.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

Metz, K. E. (2004). Children's understanding of scientific inquiry: Their conceptualization of uncertainty in investigations of their own design. *Cognition and Instruction*, *22*(2), 219–290.

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

Nersessian, N. (1992). How do scientists think? Capturing the dynamics of conceptual change in science. In R. Giere & H. Feigle (Eds.), *Minnesota studies in the philosophy of science* (pp. 3–43). Minneapolis: University of Minnesota Press.

Piaget, J. (1930). *The child's conception of physical causality*. New York: Harcourt, Brace.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, *123*(4), 392.

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

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of American Statistical Association*, *66*, 846–850.

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology*, *49*(1), 31–57.

Schulz, L., Goodman, N., Tenenbaum, J., & Jenkins, C. (2008). Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data. *Cognition*, *109*, 211–233.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, *47*, (Serial No. 194).

Siegler, R. (1996). *Emerging minds: The process of change in children's thinking*. New York: Oxford University Press.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, *28*, 303–333.

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

Thagard, P. (1988). *Conceptual revolutions*. Princeton, NJ: Princeton University Press.

Ullman, T., Goodman, N., & Tenenbaum, J. (2012). Theory acquisition as stochastic search in the language of thought. *Cognitive Development*, *27*(4), 455–480.

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2009). One and done? optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*, 645–647.

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

Wiser, M., & Carey, S. (1983). When heat and temperature were one. In D. Gentner & A. Stevens (Eds.), *Mental models* (pp. 267–297). Hillsdale, NJ: Erlbaum.

Xu, F., & Kushnir, T. (Eds.) (2012). *Rational constructivism in cognitive development. Advances in child development and behavior* (Vol. *43*). Waltham, MA: Academic Press.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Supplementary Materials.