Sticking to the Evidence? A computational and behavioral case study of micro-theory change in the domain of magnetism

Elizabeth Bonawitz^{a}**, Tomer Ullman**^{b}**, Alison Gopnik**^{a}**, Josh Tenenbaum**^{b}
(liz_b@berkeley.edu), (tomeru@mit.edu), (gopnik@berkeley.edu), (jbt@mit.edu)
^{a}**Department of Psychology, University of California, Berkeley**
^{b}**Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology**

Abstract—**An intuitive theory is a system of abstract concepts and laws relating those concepts that together provide a framework for explaining some domain of phenomena. Constructing an intuitive theory based on observing the world, as in building a scientific 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 is a learner to discover appropriate concepts and laws simultaneously, knowing neither to begin with? Even knowing the number of categories in a theory does not resolve this problem: without knowing how individuals should be sorted (which categories each belongs to), a the causal relationships between categories cannot be resolved. We explore how children can solve this chicken-and-egg problem in the domain of magnetism, drawing on perspectives from history of science, computational modeling, and behavioral experiments. We present preschoolers with a simplified magnet learning task and show how our empirical results can be explained as rational inferences within a Bayesian computational framework.**

Keywords: Theory Change; Cognitive Development; Models

I. Introduction

When children acquire or revise an intuitive theory, they must solve a fundamental "chicken-and-egg" or joint inference problem that is also at the heart of the scientific change: identifying both the causal laws and the types of things in the world over which the laws are defined. The history of scientists' attempts to understand magnetism shows the problem well. Peregrinus in 1269 was first to describe systematically the two types of poles of a lodestone, and to classify them based on their repulsive forces and their attraction to Earth's north and south directions, but he got the causal laws crucially wrong: he proposed that lodestone poles of the same type attract [1]. It took almost 350 years before William Gilbert reversed this mistake, proposing that the Earth was also a giant magnet, and hence the interactions between the poles of Earth and any lodestone should be the same as between any two lodestones [2]. Then a critical experiment tested whether poles of the same type in fact attract, or repel, revealing that poles of the same type repel, while opposites attract.

In part, what makes the theory of magnetism so difficult to discover is that the naturally occurring objects have no clear and robust perceptual differences that mark the boundary between ontological kinds and the application of causal laws. One could know that there are two categories (e.g. North, South), but mere inspection does not elucidate which sides are which and what the causal relation is between them. We view the challenge of jointly inferring the correct categorical sorting and the causal laws operating over those categories, neither of which are typically observable directly, as the fundamental "chicken-and-egg" problem of theory discovery [3], [4].

How scientists or children form coherent and useful theories from their experience has been explored by cognitive historians of science [5], who connect theory change in science with models of cognition [6]–[9], and developmentalists [10]–[13], who suggest an analogy between the cognitive mechanisms of theory change in science and childhood. Yet there has been relatively little systematic study of how children, or scientists, solve our chicken-and-egg problem. Some studies have looked at how school-aged children might learn theories [14]–[16], and several studies show that preschoolers can infer non-obvious categories from causal data [17], [18]. But to our knowledge this is the first study to investigate two related problems: first, how might an intelligent inference engine solve this chicken-and-egg problem in principle; second, can children solve the problem of jointly inferring causal law and category membership in practice.

We present a generative model for defining the problem involving a probabilistic context-free Horn Clause 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 rational learners are in principle capable of solving the chicken-and-egg problem, discovering the 'best' theories – if they can appropriately integrate a number of pieces of evidence. Our experimental results show that despite the ambiguity of the data and the difficulty of the problem, children's inferences in fact follow the qualitative predictions of this analysis. Children restrict their beliefs to hypotheses consistent with the data they observe, but maintain appropriate uncertainty as long as the data are ambiguous, making predictions in proportion to posterior probabilities. Finally, after they observe the critical intervention, they mostly converge on the single theory consistent with all the data – which is either the true theory of magnetic poles or an 'antimagnetism' variant in an alternative experimental condition.

II. Modeling the chicken-and-egg problem

Rational models let us understand how challenging learning problems may be solved in principle. Bayesian analyses, in particular, have provided compelling computational theories of causal learning and simple theory change in children and adults [17]–[23]. Kemp, Tenenbaum, Griffiths and colleagues [20], [24] have presented a Bayesian solution to the chicken-and-egg problem. They proposed a probabilistic generative model for relational data (e.g., Object 17 attracts Object 4) defined in terms of latent classes and probabilistic laws for how pairs of object classes tend to relate (e.g., Object 17 is in class A, Object 4 is in class B, and objects in class A tend to attract objects in class B). They used this model to discover theories on real-world data sets and also to describe how adults learned simple theories in laboratory settings. [25] 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 are inspired by this work, but simplified for work with children.

A. A simplified magnetism problem

Children were shown a set of six identical, unlabeled blocks that contained either a north pole *or* a south pole on one face of the block. All other sides, were inert and never participated in any interactions. We labeled two additional blocks, one of each type, 'Yellow' and 'Blue'. These labeled blocks were placed at opposite ends of a linear frame. 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. This constrained the space of plausible theories, but children still had to infer which unlabeled blocks were Yellow or Blue at the same time that they inferred how Yellow and Blue blocks causally interact. Next, each of the six blocks is systematically bumped into the two labeled (Yellow and Blue) blocks, to see whether they push or stick, and children are asked to infer the color of each unlabeled block. Children are then asked to describe how blocks will interact as a function of their color. Finally a crucial intervention is observed: a single interaction between either two blocks sorted as the same color or two blocks sorted as different colors. When combined with the earlier evidence, this last piece of data simultaneously disambiguates the color of all the blocks and the causal laws relating the colors. This scenario allows us to examine children's solutions to the chicken-and-egg problem as they get more evidence: in the presence of informative but still ambiguous data and then following data from a critical test. From a Bayesian perspective these stages correspond to the probability distribution given ambiguous evidence and then convergence to a single correct hypothesis given the crucial data. We turn to the question of priors in Experiment 2.

B. Hierarchical Theory Structure

We start with the problem of defining the space of all possible theories. Our model generates this space using the probabilistic context-free Horn Clause Grammar (PHCG) described in [25]; PHCG both defines the space of possible theories (made up of Horn Clauses) and defines the prior probability distribution over these theories (see [20], [25], [26] for applications to psychological theories). Following [25] we define two types of predicates which the Horn Clauses can consist of: "core" ($\mathsf{yellow}(\mathsf{X})$ and $\mathsf{blue}(\mathsf{X})$) and "surface" ($\mathsf{sticks}(\mathsf{X},\mathsf{Y})$ and $\mathsf{repels}(\mathsf{X},\mathsf{Y})$). Theory laws are combined with an assignment of the core predicates, (e.g. the law $\mathsf{repels}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{blue}(\mathsf{X}) \wedge \mathsf{blue}(\mathsf{Y})$ can be combined with the core predicate assignment $\mathsf{blue}(\mathsf{object1})$, $\mathsf{blue}(\mathsf{object2})$ to predict that the observed values of the surface predicate $\mathsf{repels}(\mathsf{X},\mathsf{Y})$ should be $\mathsf{repels}(\mathsf{object1},\mathsf{object2})$ and $\mathsf{repels}(\mathsf{object2},\mathsf{object1})$). The specific core predicate assignments are models; each theory can allow a large number of possible models. The meaning of the predicates themselves derives from their extension, in combination with the laws, which captures the principle chicken-and-egg problem described here.

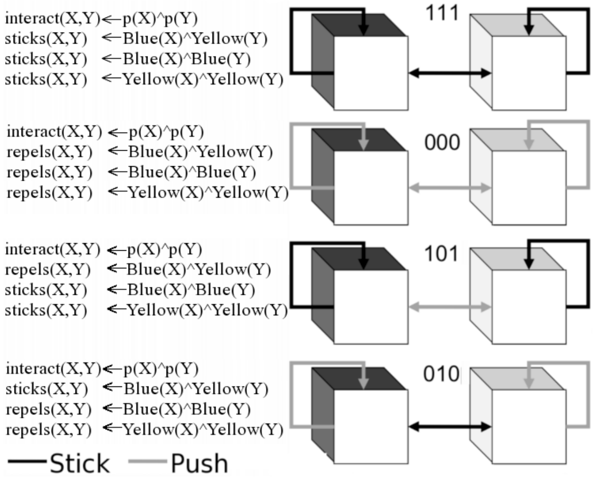
interact(X,Y)←p(X)^p(Y)
sticks(X,Y) ←Blue(X)^Yellow(Y)
sticks(X,Y) ←Blue(X)^Blue(Y)
sticks(X,Y) ←Yellow(X)^Yellow(Y)
111
interact(X,Y) ←p(X)^p(Y)
repels(X,Y) ←Blue(X)^Yellow(Y)
repels(X,Y) ←Blue(X)^Blue(Y)
repels(X,Y) ←Yellow(X)^Yellow(Y)
000
interact(X,Y) ←p(X)^p(Y)
repels(X,Y) ←Blue(X)^Yellow(Y)
sticks(X,Y) ←Blue(X)^Blue(Y)
sticks(X,Y) ←Yellow(X)^Yellow(Y)
101
interact(X,Y)←p(X)^p(Y)
sticks(X,Y) ←Blue(X)^Yellow(Y)
repels(X,Y) ←Blue(X)^Blue(Y)
repels(X,Y) ←Yellow(X)^Yellow(Y)
010
Stick
Push

Fig. 1: Four of the eight possible hypotheses for the causal laws in our simplified magnets example. The 'stick' relation (represented by the dark black lines) is labeled as'1' and the 'push' relation (represented by the lighter grey lines) is labeled as '0'. Thus, a hypothesis 001 specifies that blue repels blue, blue repels yellow, and yellow sticks to yellow. Note that the correct hypothesis for magnets is 010, but the reverse hypothesis (101) is also consistent with almost all of the data children observe in our experiment. Next to each theory is an example of the laws in the PHCG.

C. The prior on theories

The prior probability of a given theory is the product of the probabilities of the choices involved in generating that theory, as each production rule has a probability associated with it. Such a prior favors overall simpler theories with fewer laws and fewer predicates in each law, as these require fewer productions to derive. The prior probability of specific theories also depends on the particular set of production probabilities; we consider two. The first (*Generic Prior*) assumes a uniform distribution giving a prior which favors *reuse* of the same terminal symbols over choosing many different terminal

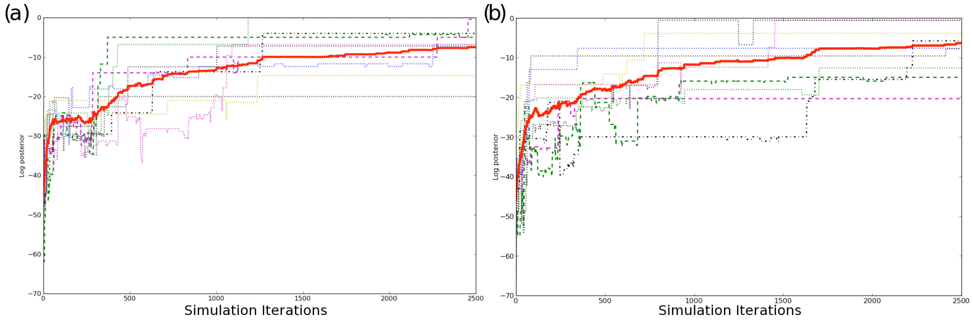

Fig. 2: Log posterior score of the theories for different runs of the simulation, conditioned on seeing ambiguous evidence. The solid line shows the mean score across all runs, while dashed lines show representative individual runs: (a) *Generic prior* (b) *Stick Bias*.

symbols[1]. In the second (*Stick bias*) the production probability of the *sticks* predicate, c_s, is higher than the production probability of the *repels* predicate, c_r (we will use c_r = 1-c_s). This means that under this prior, a Horn Clause is more likely a-priori to imply 'stick' than 'repel'[2].

D. Searching the space of possible theories

Our main analysis will focus on a restricted set of 8 possible theories, which describe all unique combinations of object properties and their relations. That is, each theory in the set represents a canonical example of a class of theories which are are equivalent in their extension. Given our experimental setup, the theories are indistinguishable within each class, but distinguishable between each class (see Figure 1). However, given the potentially infinite space of theories generated by the Horn Clause Grammar, it becomes prudent to first ask how a learner might actually search through and discover the correct theories. As a practical way to search, we adopt a grammar-based Metropolis-Hasting (MH) algorithm [25], [28], which begins with a specific theory t, and then uses the PHCG to propose random changes. Using the algorithm specifications outlined in [25] we ran several simulations to see which theories are sampled given interaction data between unlabeled blocks and the yellow and blue blocks that does not include the crucial intervention. Having run for both models 20 simulations for 2500 steps, both the *Generic Prior* and *Stick Bias* model found two possible alternative theories that scored highest in the search: One theory can be described as 'Like attracts like' (expecting blue to attract blue, yellow to attract yellow, and yellow and blue to repel, labeled "101" theory using terminology we'll introduce below); the second theory is 'Like repels like' (expecting two blues to repel, two yellows to repel, and yellow and blue to attract, "010" theory). (See Figure 2). Both theories are consistent with all the observed data and are intuitively simple, showing that stochastic search can indeed be used to find reasonable theories[3].

E. Likelihood evaluation

Given some observed data d on how the blocks interact, the Bayesian ideal learner updates the prior over theories $P(t)$ to a posterior degree of belief $P(t|d)$ via Bayes' rule. This requires the prior - described in the previous section - and the likelihood $P(d|t)$ for each theory t. This is the probability of observing the data d if that theory were true. However, the data cannot be computed directly from the theory without the additional assignment of the core predicates' truth values. For example, consider the law $\mathsf{repels(X,Y)} \leftarrow \mathsf{blue(X)} \wedge \mathsf{blue(Y)}$ cannot predict which objects would *repel* without first assigning which objects are *blue*. We refer to such an assignment of the core predicates as a model, m. Thus, the likelihood can be calculated by summing over the possible models for a given theory:

$$P(d|t) = \sum_i P(d|m_i,t)P(m_i|t) \tag{1}$$

Given a particular theory and model assignment, we can logically deduce the predicted observed data. As in previous Bayesian models of children's causal learning [19], we allow for a small probability ε of a noisy observation – due to noise in the demonstration, noise in the child's observation, etc. We set $\varepsilon = 0.2$ to capture the intuition that observed causal interactions generally conform to the true theory's predictions. The actual value of ε does not dramatically affect the results, as long as ε is relatively small.

For a given model m, we can compute the likelihood $P(d|m,t)$, as ε or $1-\varepsilon$ according to whether each of the interactions in d is consistent with the laws of t under the model m. To update beliefs in a candidate theory, t, following the final crucial interaction of two of the original unlabeled

[1] For example, the law $\mathsf{sticks(X,Y)} \leftarrow \mathsf{blue(X)} \wedge \mathsf{blue(Y)}$ would have a higher prior than the law $\mathsf{sticks(X,Y)} \leftarrow \mathsf{yellow(X)} \wedge \mathsf{blue(Y)}$, due to the reuse of the *blue* predicate. By virtue of the grammar favoring the re-use of previous predicates, this leads in turn to like-objects sticking more often. See [27] for the derivation of this 'syntactic prior' and technical details of its properties.

[2] We consider this model because a stick bias is consistent with the history of early magnetism theories, seems intuitively plausible if only because the attractive powers of magnets are more salient than their repulsive properties, and because 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.

[3] As noted, there are many variants of extensionally equivalent theories. For example, one theory might use the following laws to describe the 'repels' relation: $\mathsf{repels(X,Y)} \leftarrow \mathsf{blue(X)} \wedge \mathsf{blue(Y)}$ and $\mathsf{repels(X,Y)} \leftarrow \mathsf{yellow(X)} \wedge \mathsf{yellow(Y)}$, while another could use the law $\mathsf{repels(X,Y)} \leftarrow \mathsf{sticks(X,Z)} \wedge \mathsf{sticks(Z,Y)}$.

blocks, we let $\tilde{d} = d$ augmented with this one additional interaction and compute the likelihood with the updated $\tilde{d}$.

We now turn to the question of computing the probability of a model (or core predicate assignment) under a given theory, $P(m|t)$. Following [25], [26] we generate the model assignments by treating each core predicate value as a Bernoulli random variable with a conjugate beta prior. That is, for each core predicate i we draw $\theta_i \sim Beta(\alpha, \beta)$, where α and β are hyperparameters chosen to encourage sparsity. In order to calculate the likelihood in Equation 1 one needs to sum over all possible models allowable under the considered theory. This sum is generally computationally intractable, and so we approximate it by using an estimate of the maximum a-posteriori (MAP) model assignment $m*$. Thus we replace Equation 1 with:

$$P(d|t) = \sum_i P(d|m*, t) P(m*|t) \tag{2}$$

The MAP estimate $m*$ is found using a Gibbs sampler over the possible model assignment (see [25] for some of the technical details). As mentioned earlier, the process of finding an optimal assignment of the core predicates for a given theory parallels the problem facing children in the labeling phase of the experiment: Given how I think blocks work and the data observed so far, is this particular unlabeled block x_i an instance of *yellow*, or *blue*?

1) Modeling Results: Figure 3 shows the behavior of these Bayesian models, in both the *Generic Prior* and *Stick Bias* variants, and also given one observed interaction between two unlabeled blocks which should discriminate between hypotheses 010 and 101. Both models' posterior probabilities $P(h|\tilde{d})$ 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 101, inheriting from its prior (Figure 3, columns 2 & 3).

III. Preschoolers and the Chicken-and-egg Problem

The modeling demonstrates how a theory of magnetism can be learned in principle. After some evidence, a rational learner should retain two hypotheses, the correct-magnet hypothesis and the reverse-magnet hypothesis, ruling out all others. Do preschoolers do this or simply respond at chance? In the second step children get a single data point that, together with the initial evidence, supports a single theory. Do children correctly recognize the one theory that now uniquely explains all the data they have observed? Can they use the data to override their initial stick bias? To make this inductive leap, children must be able to simultaneously integrate the informative (but ambiguous) evidence in the first trial with the disambiguating evidence in the last trial. To our knowledge, no study has examined whether children can actually do this.

A. Methods

1) Participants: Seventy-eight four- and five-year-olds were recruited from an urban area science museum ($M = 59$ months, range = 47–73 months). Approximately half of the participants were female and a range of ethnicities proportional to urban populations were represented.

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 a disambiguating evidence phase which 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 self-terminated the experiment before completion.

3) Procedure: Ambiguous Evidence Phase. Children were shown a stand that had a single Yellow and Blue block placed at either end and told that Yellow and Blue blocks might push or stick to one another or might push or stick 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 additional 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?" 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 experimenter then followed the same procedure with the remaining five blocks, selecting the next block in pseudo random order.[4]

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. For those children who also completed the disambiguating evidence phase, following the ambiguous evidence phase, the experimenter said, "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* groups interact (blue-yellow). Children were then asked all three theory prediction questions again, which, critically, included the other two unobserved interactions.

[4] 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.

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'. 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 [5].

B. Results

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 ($\sim$ 40%) of the responses were coded by the first author; reliability was 100%. There was no effect of sorting order or age on responding.

1) Sorting: We coded whether children sorted the unlabeled blocks according to a 'likes-attract' rule or a 'likes-repel' rule. Most children (84%) sorted at least 5 of 6 blocks according to one of the two patterns, our criterion for success. In fact, most of those children (also 84%) sorted all six blocks consistently. Just 16% of the children sorted the blocks randomly. Almost all of the remaining children sorted according to the 'likes-attract' rule (94%), only a handful sorted according to the 'likes-repel' rule. This is consistent with the Stick Bias prior.

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 (010) and the reverse-magnet theory (101). Children generated both of these hypotheses above chance (correct-magnet: binomial (n = 15/75), $p < .05$; reverse-magnet: binomial $(n = 35/75), p < .0001$). The distribution of children's responses correlated marginally with the *Generic Prior* model ($r^2 = .86$), but very highly with the *Stick Bias*[6] model ($r^2 = .95$). (See Figure 3, Column 1).

3) Final theory prediction: Children in both conditions also learned from the final intervention trial, generating significantly different (and evidence-consistent) responses (*Magnet Consistent*: Fisher Exact $(28), p < .01$; *Magnet Inconsistent*: Fisher Exact $(19), p < .05$). 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.

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 ($r^2 > .93$)[7]. (See Figure 3, columns 2 & 3).

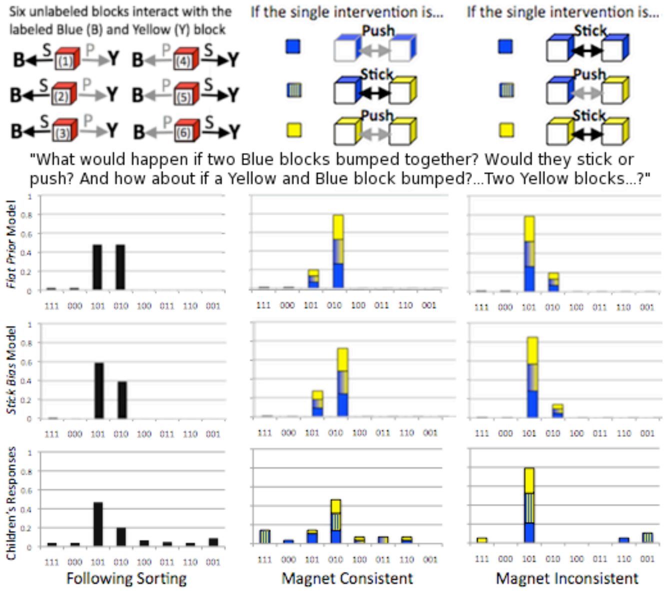

Fig. 3: Predictions of the *Generic Prior* and *Stick Bias* models compared to children's predictions from the Experiment, after seeing the six unlabeled blocks interact with the Yellow and Blue blocks (column 1); after observing the final disambiguating intervention that is magnet consistent (column 2) or after observing the final disambiguating intervention that is magnet inconsistent (column 3).

IV. Establishing the priors on children's theories

One might be concerned that, rather than learning from the initial evidence, children's responding on the ambiguous evidence task simply reflected strong prior beliefs for the 'likes-attract' and 'likes-repel' rules. We can rule out this deflationary account out by asking whether prior to seeing any evidence, children entertain a range of theories. We tested a new group (N = 20) of four- and five-year-olds in a *Priors* condition. Children were told that sometimes things push and some other things stick with four familiarization items, two items repelled (a toy hippo and 'onion') and two others attracted (a toy ball and eggplant). 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 1. 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?" The experimenter mimed the action with the sets of blocks (without touching the blocks) as each question was asked.

Preschoolers entertained a variety of theories. Contrasting the distribution of hypotheses to Experiment 1 revealed significant differences between Experiments, Fisher Exact, $p < .05$, suggesting that children's responding in Experiment 1 reflected genuine learning. We also compared children's empirical priors to the priors in the models (although the

[5] While, the majority of children stated that they had played with magnets previously, almost no children believed these blocks were like magnets.

[6] We also computed correlation for a variety of values on the *Stick Bias*, ranging from .6 to .9; results were robust, with all correlations $r^2 > .95$

[7] These correlations were robust across a range of values for the *Stick Bias* ranging from .6 to .8 ($r^2 > .91$); values with the extreme bias of .9 correlated slightly worse ($r^2 > .84$) due to the over-favoring of the stick-rule.

precise profiles of these priors over all 8 possible theories cannot be evaluated without a substantial n). The *Stick Prior* model predicts more sticking relations than repel relations; children in the *Priors* Experiment did the same, Binomial, $p < .05$. The *Generic Prior* does not predict more sticking relations, but does predict more re-use of "stick" or "repel", favoring the 111 and 000 theories over the others; children did not favor the 111 and 000. Thus, children's responses in the *Priors* Experiment provide additional qualitative support for the *Stick Bias* model over the *Generic Prior* model.

V. Discussion

Children's behavior was consistent with our rational model, particularly the Stick Bias model. Children showed a stick bias prior which echoes the early magnetism theories of Peregrinus and others which also showed a bias towards attraction, however both scientists and the children eventually overcame this incorrect bias as predicted by our model. Second, in both the history of science and in our task, learners received ambiguous but informative evidence. The children demonstrated an appropriate but focused uncertainty, rationally restricting beliefs to hypotheses that were consistent with the data but also drawing on both priors and observations to set posteriors. Finally, we asked whether, analogous to William Gilbert's classic studies, a single critical intervention between just two blocks could lead children to adopt a single hypothesis, simultaneously disambiguating the hidden categories of objects and the causal rules between categories. Children were able to infer the appropriate causal law (the magnet rule following one kind of evidence, and the reverse-magnet rule following the other.) Thus, even in the course of a short experiment, four-year-olds were able to solve a simple version of the chicken-and-egg problem by rationally 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 cognitive development.

How can children have learned the correct rule in the course of our short experiment when historically such theory change can take centuries [4]? While historical analogies can provide some insight into the difficulties and strategies of intuitive theory discovery, 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. Children are told in advance that there are just two kinds of objects and are shown the interventional data without having to come up with the intervention themselves: children in our task, unlike the scientists, did not have to be meta-cognitively aware of how to design informative interventions. Despite these differences, our results can take the "child as scientist" analogy to a new level of empirical richness and computational rigor. Even four-year-olds can work out the relation between the epistemic chickens and eggs.

References

[1] J. F. Keithley, *The Story of Electrical and Magnetic Measurements: From 500 B.C. to the 1940s*. Hoboken, NJ: John Wiley and Sons, 1999.

[2] W. Gilbert, *De Magnete*. Mineola, NY: Dover Publications, 1600/1958.

[3] W. V. O. Quine, *Word and object*. Cambridge, MA: MIT Press, 1960.

[4] S. Carey, *The Origin of Concepts*. New York: Oxford University Press, 2009.

[5] N. Nersessian, "How do scientists think? capturing the dynamics of conceptual change in science," in *Minnesota studies in the philosophy of science*, R. Giere and H. Feigle, Eds. Minneapolis: MN: University of Minnesota Press, 1992.

[6] D. Gentner, S. Brem, R. Ferguson, A. Markman, B. Levidow, P. Wolff, and K. Forbus, "Analogical reasoning and conceptual change: A case study of johannes kepler," *The Journal of the Learning Sciences*, vol. 6, no. 1, pp. 3–40, 1997.

[7] H. Gruber and P. Barrett, *Darwin on man: A psychological study of scientific creativity*. New York: Dutton, 1974.

[8] T. Kuhn, *The structure of scientific revolutions*. Chicago: University of Chicago Press, 1962.

[9] M. Wiser and S. Carey, "When heat and temperature were one," in *Mental models*, D. Gentner and A. Stevens, Eds. Hillsdale, NJ: Erlbaum, 1983.

[10] S. Carey, *Conceptual change in childhood*. Cambridge, MA: MIT Press, 1985.

[11] A. Gopnik and A. N. Meltzoff, *Words, thoughts, and theories*. Cambridge, MA: MIT Press, 1997.

[12] F. C. Keil, *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press, 1989.

[13] H. M. Wellman and S. A. Gelman, "Cognitive development: Foundational theories of core domains," *Annual Review of Psychology*, vol. 43, pp. 337–375, 1992.

[14] R. Siegler, *Emerging minds: The process of change in childrens thinking*. New York: Oxford University Press, 1996.

[15] C. Smith and C. Unger, "Whats in dots-per-box? conceptual bootstrapping with stripped down visual analogs," *Journal of the Learning Sciences*, vol. 6, no. 2, pp. 143–181, 1997.

[16] S. Vosniadou and W. Brewer, "Mental models of the earth: A study of conceptual change in childhood," *Cognitive Psychology*, vol. 24, pp. 535–585, 1992.

[17] D. M. Sobel, J. B. Tenenbaum, and A. Gopnik, "Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers," *Cognitive Science*, vol. 28, pp. 303–333, 2004.

[18] L. Schulz, N. Goodman, J. Tenenbaum, and C. Jenkins, "Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data," *Cognition*, vol. 109, no. 2, pp. 211–233, 2008.

[19] T. Griffiths and J. Tenenbaum, "Theory-based causal induction," *Psychological Review*, vol. 116, pp. 661–716, 2009.

[20] C. Kemp, J. Tenenbaum, S. Niyogi, and T. Griffiths, "A probabilistic model of theory formation," *Cognition*, vol. 114, pp. 165–196, 2010.

[21] T. Kushnir and A. Gopnik, "Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions," *Developmental Psychology*, vol. 44, pp. 186–196, 2007.

[22] C. Lucas, A. Gopnik, and T. Griffiths, "Developmental differences in learning the forms of causal relationships," in *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 2010.

[23] L. E. Schulz, E. B. Bonawitz, and T. L. Griffiths, "Can being scared make your tummy ache? naive theories, ambiguous evidence, and preschoolers' causal inferences," *Developmental Psychology*, vol. 43, pp. 1124–1139, 2007.

[24] C. Kemp, J. Tenenbaum, T. Griffiths, T. Yamada, and N. Ueda, "Learning systems of concepts with an infinite relational model," in *Proceedings of the National Conference on Artificial Intelligence*, vol. 21, no. 1. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006, p. 381.

[25] T. Ullman, N. Goodman, and J. Tenenbaum, "Theory acquisition as stochastic search," in *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 2010.

[26] Y. Katz, N. Goodman, K. Kersting, C. Kemp, and J. Tenenbaum, "Modeling semantic cognition as logical dimensionality reduction," in *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 2008.

[27] N. Goodman, J. Tenenbaum, J. Feldman, and T. L. Griffiths, "A rational analysis of rule-based concept learning," *Cognitive Science*, vol. 32:1, pp. 108–154, 2008.

[28] N. Goodman, T. Ullman, and J. Tenenbaum, "Learning a theory of causality," *Psych. Review*, in press.