Developmental differences in learning the forms of causal relationships

Chris Lucas (clucas@berkeley.edu)
Alison Gopnik (gopnik@berkeley.edu)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California
Berkeley, CA 94720, USA

Abstract

Children learn causal relationships quickly, and make far-reaching causal inferences on the basis of what they see. In order to be such efficient learners, they must bring abstract knowledge to bear on their problems. This paper addresses children's ability to acquire that knowledge. We present evidence that children can learn about the abstract properties of causal relationships using only a handful of events, and – consistent with a hierarchical Bayesian model of casual inference – children can be more sensitive to evidence than adults.

Introduction

Recent work suggests that children are skilled at inferring specific causal relationships from patterns of data (Gopnik et al., 2004; Sobel, Tenenbaum, & Gopnik, 2004). For example, they can infer which blocks will activate a machine based on the contingencies between the blocks and the machine's activation. But an additional question is whether children can infer more abstract causal principles from patterns in data, and use those principles to shape their subsequent predictions. For example, can a child infer that a particular type of machine activates reliably, or requires only a single cause to activate? Will those abstract discoveries bias the child's interpretations of new data?

Developmental data suggest that children do have broad inductive biases. For example, in language learning the shape bias and the mutual exclusivity principle influence more specific inferences about word meaning (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002; Markman & Wachtel, 1988). However there is debate about whether these biases are the result of innate constraints or are themselves the product of learning (Elman et al., 1996; Leslie, 1994). Recent formal work on hierarchical Bayesian models suggests that, at least in principle, the shape bias may itself be learned as a result of normative inferences from patterns of data (Kemp, Perfors, & Tenenbaum, 2007). Similar high-level biases apply to causal learning, and we know that children can learn about causal types (Schulz, Goodman, Tenenbaum, & Jenkins, 2008), and the plausibility of cross-domain relationships (Schulz, Bonawitz, & Griffiths, 2007). In this paper, we explore whether children can learn abstract principles about the forms of causal relationships themselves.

The hierarchical Bayesian approach suggests that the nature of inductive biases may change as evidence accumulates. Absent evidence, a learner without strong built-in biases should assign similar probabilities to a wide range of hypotheses. As data accumulate, the abstract hypotheses consistent with those data become more probable, and the learner discounts any hypotheses that fit the current data but are less compatible with past experience. If this is correct, then we might expect to see different patterns of inductive bias in adults and children. In particular, children might rely less on past experience and more on present evidence than adults. This is a possibility that has not previously been explored in the causal learning literature, and one that we examine through head-to-head (or prior-to-prior) comparison of children and adults in a causal learning task that requires making an abstract generalization about the nature of causal relationships.

We test the high-level generalizations made by children and adults by contrasting two abstract "overhypotheses" (Goodman, 1955; Kemp et al., 2007) about how a causal system works. One is a noisy-OR model, in which each object has a certain independent probability of bringing about an effect. This model is pervasive in the literature on adult causal inference (e.g., Cheng, 1997; Griffiths & Tenenbaum, 2005). The other is an AND model in which individual causes are unable to produce an effect, but multiple causes in conjunction can produce an effect. We provided children and adults with evidence for either an AND or OR relationship and then examined how this evidence biased their judgment of a novel, ambiguous pattern of evidence. Would seeing several instances of a machine activated by a conjunction of causes lead them to assume that this would be the case for a new set of blocks? By comparing how children and adults respond to data that support these different overhypotheses, we can examine first whether children are capable of forming appropriate abstract generalizations, and second whether they are more willing to make these generalizations than adults.

The plan of the paper is as follows. First, we consider how an ideal Bayesian learner can gather evidence for overhypotheses relevant to causal induction. We then discuss the specific overhypotheses about the functional form of causal relationships that we contrast in this paper, together with a method that can be used to diagnose whether learners infer these overhypotheses from data. We go on to use this method to compare the abstract generalizations of children and adults in a causal learning task, finding support for the hypothesis that children are more willing to adopt a novel overhypothesis than adults. We close by discussing the implications of these results.

Causal overhypotheses

Children can identify causes using only a handful of observations (Gopnik et al., 2004), but the extent to which they learn

about the abstract properties of causal relationships remains largely unexplored. From a Bayesian standpoint, learning about causal structure requires having *a priori* beliefs – or priors – about what items are plausible causes, and expectations about how a given causal structure leads to different observable events. These expectations can be expressed formally using a *likelihood* function, which specifies the probability of observing a particular set of events based on the underlying causal structure.

Most work on probabilistic models of causal learning has assumed a specific kind of likelihood function. This likelihood function is based on causes and effects interacting in a "noisy-OR" manner, each having an independent opportunity to produce the effect (Cheng, 1997; Griffiths & Tenenbaum, 2005; Glymour, 1998). More precisely, a noisy-OR relationship implies that the probability that an effect E occurs given the presence of a set of causes $C_1, \ldots, C_N$ is

$$P(E|C_1, ..., C_N) = 1 - \prod_{i=1}^{N} (1 - w_i) \tag{1}$$

where w_i is the probability that C_i generates the effect in the absence of other causes.

Despite the popularity of the noisy-OR in models of causal learning, other kinds of causal relationships are clearly possible. For instance, a noisy-OR model cannot describe an AND relationship, where an effect only occurs when multiple causes are present. This might be the case in an electrical circuit where multiple switches are wired in series, and a light only turns on when all of the switches are flipped. It is important, then, for models of causal inference to accommodate flexible beliefs about the forms relationships can take. Formalizing inferences about the form of a relationship is straightforward, using an expanded likelihood function, $P(E|C_1, ..., C_N, F)$, where F captures information about the form of the causal relationship. For example, F could indicate that the relationship has a noisy-OR form, but another value of F might indicate that a causal relationship has an AND form.

Learning the form of a causal relationship and generalizing that discovery when reasoning about other causal relationships requires inference at multiple levels of abstraction. This kind of inference, in which lessons from one context can be carried forward for future learning, is easily captured by using a hierarchical Bayesian model (Tenenbaum, Griffiths, & Kemp, 2006; Kemp et al., 2007). A learner's abstract beliefs, or overhypotheses, determine the probabilities of more-concrete hypotheses, each encoding specific causal structures and the form a relationship takes. These hypotheses, in turn, determine the likelihood of different patterns of events.

Formally, we can imagine an inference involving variables at three levels: the observed data D, hypotheses about the causal structure underlying those data H, and overhypotheses (or a "theory", as in Griffiths & Tenenbaum, 2009) T representing generalizations relevant to evaluating those hypotheses (see Figure 1). Bayes' rule then specifies how the events

T : Theory/overhypothesis

H : Hypothesis

D : Data

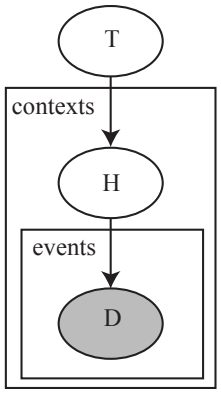

Figure 1: The structure of a hierarchical Bayesian model.

a learner sees (D) should change the learner's beliefs, both about the casual system at hand (H), and about the higher-level properties of that kind of system (T). Formally, we have

$$p(T|D) = \frac{p(D|T)\,p(T)}{p(D)} \tag{2}$$

where $p(T)$ is the prior probability of the overhypothesis T, $p(T|D)$ is the posterior probability, and $p(D)$ is obtained by summing the numerator over all overhypotheses T. The probability of the data given an overhypothesis is obtained by summing over all hypotheses consistent with that overhypothesis,

$$p(D|T) = \int p(D|H)\,p(H|T)\,dH, \tag{3}$$

and can be interpreted as an average of the probability of the observed data under those hypotheses weighted by the extent to which each hypothesis is consistent with the overhypothesis.

Intuitively, this hierarchical Bayesian approach provides a way to explain how learners can form and use abstract generalizations about causal systems. For example, if a child sees events that are likely under an AND relationship, such as a machine activating only when pairs of causal objects are placed on it, then the probability of an overhypothesis predicting future AND relationships increases. This is because the best hypotheses for explaining the observed events are those that are most likely under this overhypothesis, so Equation 3 yields a high value. Incorporating this value into Equation 2, the posterior probability for that overhypothesis will increase.

As the evidence supporting a particular overhypothesis increases, it will be easier to learn about the structure and form of causal systems that are consistent with that overhypothesis. This comes with a cost: if a causal system has strange or rare abstract properties, such as an unlikely functional form, much more evidence will be necessary to learn about it. The implication is that adults, who have seen a great deal of evidence, should find it very easy to learn about the structure and form of causal relationships that have typical properties. Conversely, children, with their limited experience, should be more sensitive to evidence when learning about relationships

that have unusual properties. In the following section, we discuss an experimental design for testing this idea.

The functional form of causal relationships

If children update their abstract beliefs about causal systems in a manner consistent with Bayesian inference, then the events they see should influence their judgments about different sets of events and prospective causes. To test this hypothesis, we used an experiment with two phases, each with a distinct set of objects. In the first phase, children saw a set of events designed to be likely under one of two abstract overhypotheses about the forms of causal relationships. In the second phase, they saw events where different beliefs about the form of the causal relationship should lead them to make different judgments about which objects are causes.

The specific evidence we provided to participants was very similar to that given to adults in Lucas and Griffiths (2009), where the task was to identify the blickets within a set of objects, knowing only that blickets have "blicketosity". Prospective blickets could be placed on a "blicketosity meter", causing it to either activate by lighting up and playing music or do nothing. People might entertain a variety of expectations about the relationship between the blickets and the machine, determining how they interpret different events. For example, if they think that two blickets are necessary to activate the machine, seeing a single object fail to activate it provides no information. At the same time, their expectations about the form of the relationship between blickets and the blicketosity meters can be shaped by the events they observe. For instance, seeing two objects fail to activate the machine separately but succeed together suggests that two blickets are necessary for activation.

We used events from two conditions from Experiment 2 of Lucas and Griffiths (2009). Since this experiment is closely related to the approach we take here, we will recapitulate the method and results. In the *AND*[1] condition of the experiment, participants saw a training block of events where objects labeled A, B, and C were placed sequentially on the machine, which failed to activate in all cases. Next, all pairs of objects were placed on the machine sequentially, with only A and B together causing activation. See Figure 2 for a summary of the events in the training and test blocks. Participants were then asked to rate the probability that A, B, and C were blickets on a 0-10 scale, with 0 indicating the object was definitely not a blicket, a 10 indicating it definitely was, and 5 indicating it was as likely to be a blicket at not.

After making these judgments, participants saw three new objects, D, E, and F, which they had never seen before, and a series of test events intended to be ambiguous, leading to different judgments about which of D, E, and F were blickets, depending on participants' expectations about the form of the relationship. If people expect that a single blicket suffices to activate the machine, they should believe then F is likely to be a blicket, while D and E are not. If, in contrast, people exploit the information provided by the training block so they conclude that two blickets are necessary to activate the machine, then they should think that objects D and F are blickets, and be uncertain about object E.

In the *OR* condition, participants saw a different set of events in the training block, which were chosen to indicate that an OR relationship applied (see Figure 2). Then they saw the same test events that the participants in the *AND* condition saw. Based on the training evidence, participants in this condition were predicted to say that only object F was a blicket.

As predicted, people in the *AND* condition assigned significantly higher probabilities to object D being a blicket, giving a mean score of 3.08 (SD=3.32), versus 0.23 (SD=0.99) in the *OR* condition. The mean rating was less than 5 in the *AND* condition, consistent with the idea that adults believe that disjunctive relationships are more probable, and could interpret the *AND* condition events in several ways, including as evidence for a noisy relationship in which the machine happened to fail to activate when a single, normally sufficient blicket was placed on it.

In summary, Lucas and Griffiths (2009) showed that people's inferences about causal structure are driven by their beliefs about the probable forms of causal relationships, which are in turn influenced by events they have seen in the past. The specific pattern of judgments is consistent with the predictions of a hierarchical Bayesian model given priors reflecting a strong bias in favor of disjunctive (OR) and deterministic relationships. Such priors are also consistent with adults' performance in other experiments (Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2006). This prior could be chiefly due to adults' experiences revealing that OR relationships are more common, or an innate bias. By comparing the judgments of 4-year-old children to those of adults, we aim to answer that question and better understand the origins of the abstract knowledge that drives efficient causal inference.

Causal overhypotheses in children and adults

We used the experimental design from Lucas and Griffiths (2009) to explore two questions about the use of causal overhypotheses by children and adults. The first question was whether children, like adults, can use events to update their knowledge about the likely forms of causal relationships, and apply that knowledge to learn the causal structure behind new and ambiguous sets of events. The second question was whether children are more or less sensitive to evidence supporting such high-level generalizations, as opposed to their prior beliefs.

If children are more likely than adults to call objects D and E blickets in the *AND* condition, we can conclude that much of the bias we see in adults is due to learning during and after childhood, including, for instance, experience with machines to which OR relationships apply. If children's judgments are

[1] Lucas and Griffiths labeled their conditions *conjunctive* and *disjunctive* rather than AND and OR, to highlight a hypothesis space that included a wide range of functional forms, including AND and OR as special cases. We use AND and OR here for the sake of simplicity.

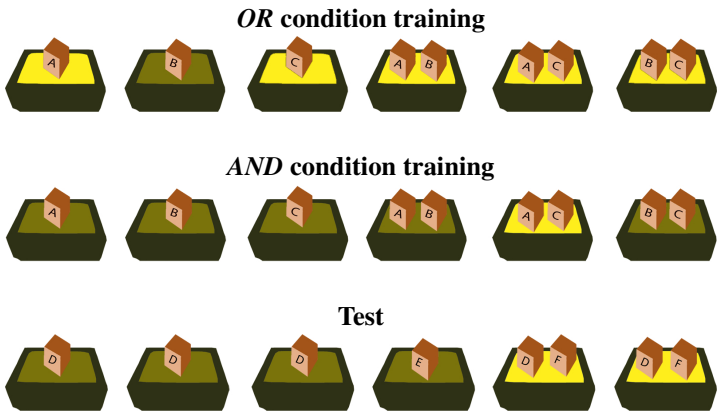

Figure 2: Evidence presented to participants in the two training phases, as well as the subsequent test phase which all participants saw. Events are given as a set of prospective causes and the presence or absence of an effect. The bright-paneled machines represent events in which the effect occurs and the dark-paneled machines represent events in which the effect does not occur.

indistinguishable from adults', we have evidence that learning about the forms of causal relationships occurs early, or plays a minor role in driving our expectations. Finally, if there is no effect of training evidence on test-block judgments, we should question the applicability of the model used by Lucas and Griffiths (2009) to causal inference in children.

We can generate more detailed predictions by speculating about the priors that children bring to the problem of identifying blickets. It seems unlikely that children are constrained to a small set of discrete overhypotheses – it is more natural to suppose that they consider a space of possibilities that includes both OR and AND relationships as special cases. Following Lucas and Griffiths (2009), we use a sigmoid family of likelihood functions, where the probability of the machine's activation given that n blickets are present is

$$P(\text{effect}|N_{\text{blickets}} = n) = \frac{1}{1+\exp\{-g(n-b)\}}. \quad (4)$$

The overhypotheses determine the probability of different values of the gain g and the bias b. The gain specifies how many blickets are necessary to activate the machine, and the bias reflects how noisy the relationship is. Lucas and Griffiths found that exponential priors predicting a high mean gain (3.34) and a low mean bias (0.23) – or reliable OR relationship – lead to model predictions that closely match adults' judgments. If children are happier believing that a relationship could be conjunctive or noisy, the priors that best capture their inferences should lead to *a priori* gains and biases closer to 1. This space of likelihood functions is intended to cover a range of relationships that are appropriate to the cover story and participants' prior knowledge, and we do not claim it includes all relationships that people could conceivably learn, such as those in which blickets prevent the machine from activating.

Participants

Children Thirty-two children were recruited from university-affiliated preschools, divided evenly between the *AND* and *OR* conditions. Children in the *AND* and *OR* conditions had mean ages of 4.46 (SD=0.27) and 4.61 (SD=0.31) years, respectively.

Adults UC Berkeley undergraduates received course credit for participating during lectures of an introductory psychology course. There were 88 participants in the *AND* condition and 55 in the *OR* condition. Five participants in the *AND* condition were excluded for declining to answer one or more questions.

Methods

Children Each child sat at a table facing the experimenter, who brought out three objects, each painted a different color, as well as a green box with a translucent panel on top, describing the box as "my blicketness machine".

At the beginning of the experiment, children were prompted to help the experimenter name the objects using their colors, e.g., "red". They were then told that the goal of the game was to figure out which of the objects were blickets, that blickets have blicketness inside them, and blickets cannot be distinguished from non-blickets by their appearance. No other information was provided about the relationship between blickets and the activation of the machine.

The children then observed a set of training events in which the experimenter placed objects alone or in pairs on the machine, which activated in some cases by lighting up and playing music. These events corresponded to either the *OR* condition or *AND* condition training given in Figure 2. After the children saw these events, they were asked whether each object was a blicket or not. Next, the experimenter brought out three objects that the children had not seen before. After the children named the new objects, the experimenter demonstrated the test events listed in Figure 2 and asked whether

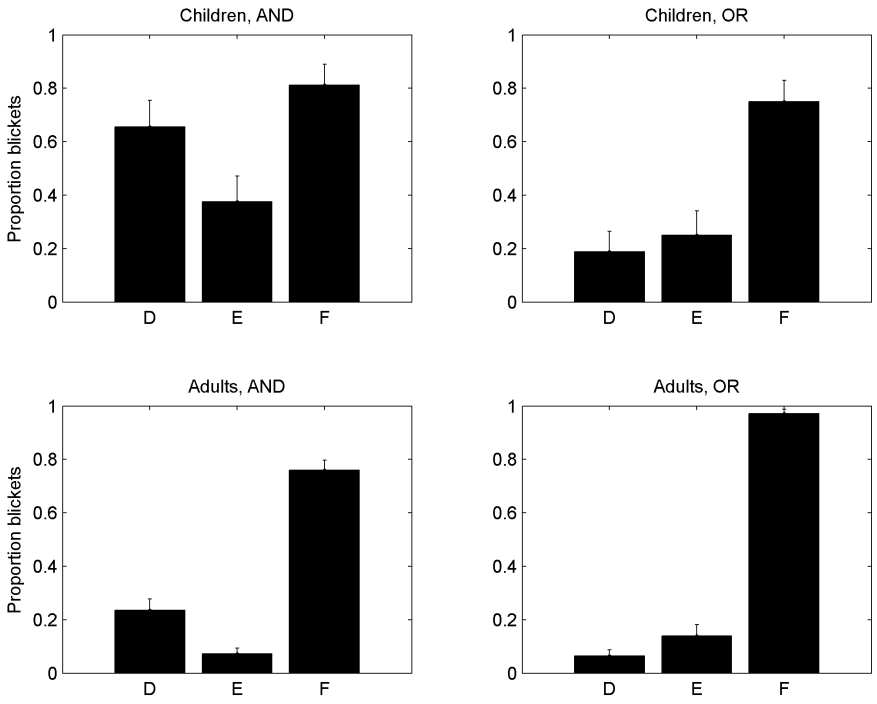

Figure 3: Proportions of objects that were judged to be blickets for children (top row) and adults (bottom row) for the *AND* (left column) and *OR* (right column) conditions. Error bars represent standard error of the mean.

each of these new objects was a blicket or not. In a departure from Lucas and Griffiths's design, the experiment was repeated a second time for each child, using the same patterns of evidence, but with a distinct set of objects that varied by shape and had a uniform gray color. The identities of the individual objects were counterbalanced, as was whether the children saw the different-shaped or different-colored objects first.

Adults The adults were tested in groups, and saw demonstrations that were almost identical to what the children saw in the corresponding conditions. Unlike the children, the adults were not asked to name the objects, and they recorded their judgments on sheets of paper rather than responding verbally.

Results

Children The critical prediction was that children would be more likely to judge object *D* to be a blicket in the *AND* condition than in the *OR* condition, indicating that they were (1) learning about the form of the relationship between blickets and the machine's activation, and (2) transferring that abstract knowledge to make better inferences about novel objects and otherwise ambiguous events.

Children were more likely to judge object D to be a blicket in the *AND* condition than in the *OR* condition ($p < 0.005$, two-tailed permutation test). There was also a change in the predicted direction for object E, albeit non-significant.

Adults Adults were also more likely to judge object D to be a blicket in the *AND* condition than in the *OR* condition ($p < 0.005$, two-tailed permutation test), consistent with the results in Lucas and Griffiths (2009). See Figure 3, bottom row, for a summary of their judgments for the test objects.

Differences In the *AND* condition, the adults judged object D to be a blicket less frequently than children ($p < 0.005$, Fisher's exact test). See Figure 3 for a summary of ratings in the three conditions. Children's ratings were also higher for object E ($p < 0.001$, two-sided permutation test), which is consistent with their being quicker to learn that an AND relationship applies: under an AND relationship, the event where E fails to activate the machine is uninformative, so judgments of E being a blicket should reflect the base rate of blickets occurring. The high frequency of other objects being blickets under an AND relationship (4 of 5), plus a belief that blickets are not rare, should lead a learner to expect that a novel object is somewhat likely to be a blicket.

Model fits We converted children's judgments about blickets to probabilities in order to examine them using the previously-mentioned hierarchical Bayesian model and sigmoid space of hypotheses. We treated is-a-blicket judgments as assertions that objects were definitely blickets, and not-a-blicket judgments as assertions that objects were definitely not blickets. Lucas and Griffiths (2009) found that priors favoring disjunctive, deterministic relationships – predicting a mean gain of 3.34 and a mean bias of 0.23 – fit adults' judgments closely, with a mean squared error of 0.29 per judgment on a zero to ten scale. We found that similar priors best

captured adults' judgments in our experiment, giving a mean squared error of 0.80 with a mean gain of 5.30 and bias of 0.11.

These same priors were wildly inconsistent with children's inferences, giving a mean squared error of 6.12. In contrast, priors giving a mean *a priori* gain and bias of 1 – favoring neither AND nor OR relationships – were much more accurate, with a mean squared error of 0.58. The priors that best fit the children's judgments gave a mean gain and bias of 1.45 and 0.85, respectively, with mean squared error of 0.15.

Discussion

Our experiment was designed to explore two questions: whether children could make high-level generalizations about the form of causal relationships, and whether they were more willing to do so than adults. Our results show that children are capable of making such inferences, and that their judgments were more strongly influenced by the available evidence than adults, whose inferences reflected a bias toward OR relationships. Our results thus support the view that when learning about cause and effect, children are flexible learners whose inexperience may sometimes let them learn better from sparse evidence, especially in novel situations. These results are also consistent with treating the acquisition and application of causal knowledge as a matter of hierarchical Bayesian inference, where a learner has beliefs expressed at multiple levels of abstraction, with abstract theories driving specific hypotheses which, in turn, enable prediction and categorization.

Before closing, we will address two alternative explanations for our results. The first is that children are more likely than adults to judge any object to be a blicket. This is less consistent with the data than our interpretation, given that adults were more likely than children to call object *F* a blicket in the *OR* condition, and nearly as likely in the *AND* condition (75 percent of the objects versus 81 percent). A second alternative is that the children were confused by the training data in the *AND* condition, and responded to the novel objects by guessing randomly. This explanation can be ruled out by noting that children judged objects *D* and *F* to be blickets more often than chance would predict ($t(15) = 3.529$, $p < 0.005$).

The results of our experiment have implications for understanding causal learning, and for understanding cognitive development more generally. In terms of causal learning, these results suggest that the fundamental biases that lie beneath causal inference are more subtle and abstract than *a priori* preferences for specific kinds of causal relationships. We believe that trying to understand these biases is fertile ground for future research. For cognitive development, the idea that children are more flexible in their commitments about the way that causal systems tend to work seems like not just a necessary consequence of a hierarchical Bayesian approach, but an important insight for understanding how it is that children see the world differently from adults. The plasticity of beliefs that this implies helps to explain the bold exploration and breathtaking innovation that characterizes children's interactions with the world.

Acknowledgments. This research was supported by the James S. McDonnell Foundation's Causality Collaborative Initiative and the Air Force Office of Scientific Research, grant FA9550-07-1-0351.

References

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367-405.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective*. Cambridge, MA: MIT Press.

Glymour, C. (1998). Learning causes: Psychological explanations of causal explanation. *Minds and Machines*, *8*, 39-60.

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

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 1-31.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354-384.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-Based Causal Induction. *Psychological Review*, *116*(4), 661–716.

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

Leslie, A. M. (1994). ToMM, ToBY, and agency: Core architecture and domain specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture.* Cambridge: Cambridge University Press.

Lu, H., Yuille, A., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2006). Modeling causal leanring using bayesian generic priors on generative and preventive powers. In R. Sun & N. Miyake (Eds.), *Twenty-eighth conference of the cognitive science society* (p. 519-524). Erlbaum.

Lucas, C., & Griffiths, T. (2009). Learning the Form of Causal Relationships Using Hierarchical Bayesian Models. *Cognitive Science*, *34*(1), 113–147.

Markman, E., & Wachtel, G. (1988). Childrens use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157.

Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared make your tummy ache? naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology*.

Schulz, L. E., Goodman, N., Tenenbaum, J., & Jenkins, A. (2008). Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data. *Cognition*, *109*(2), 211–223.

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

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