Cognition 109 (2008) 211–223

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data

Laura E. Schulz [a,*], Noah D. Goodman [a], Joshua B. Tenenbaum [a], Adrianna C. Jenkins [b]

[a] Massachusetts Institute of Technology, MIT Department of Brain and Cognitive Sciences, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
[b] Harvard University

ARTICLE INFO

Article history:
Received 16 May 2007
Revised 14 July 2008
Accepted 28 July 2008

Keywords:
Preschoolers
Statistical learning
Abstract
Anomalous data
Unobserved causes
Bayesian inference

ABSTRACT

Given minimal evidence about novel objects, children might learn only relationships among the specific entities, or they might make a more abstract inference, positing classes of entities and the relations that hold among those classes. Here we show that preschoolers (mean: 57 months) can use sparse data about perceptually unique objects to infer abstract physical causal laws. These newly inferred abstract laws were robust to potentially anomalous evidence; in the face of apparent counter-evidence, children (correctly) posited the existence of an unobserved object rather than revise the abstract laws. This suggests that children's ability to learn robust, abstract principles does not depend on extensive prior experience but can occur rapidly, on-line, and in tandem with inferences about specific relations.

1. Introduction

Children's causal insights sometimes seem like acts of wizardry. From the evidence of only a handful of trials, they can make accurate predictions, generate effective interventions, and even posit the existence of unobserved variables (Baldwin, Markman, & Melartin, 1993; Gopnik & Sobel, 2000; Gopnik, Sobel, Schulz, & Glymour, 2001; Meltzoff, 1995; Schulz, Gopnik, & Glymour, 2007). Researchers have proposed that such rapid learning is possible because children's inferences are constrained by more abstract theories (Carey, 1985; Gopnik & Meltzoff, 1997; Keil, 1989). Such naïve theories may exist at different levels of abstraction, ranging from framework principles that carve phenomena into different domains (e.g., naïve physics, naïve psychology, and naïve biology; Wellman & Gelman, 1992) to more local causal laws (e.g., about the relationship between mass and balance, Karmiloff-Smith & Inhelder, 1975, or density and floating, Kohn, 1993).

The advantage of inferring abstract causal laws, even at a relatively low level of abstraction, is evident. (We will use the phrase "abstract causal law" to refer to knowledge about classes or kinds and the causal relationships that obtain among them.) Imagine for instance, that you think that there are three classes of chemicals: acids, oxydens, and indicators. You believe that these classes instantiate three distinct causal relations: if you combine an acid with an oxyden, the mixture will explode; if you combine an acid with an indicator, the mixture will turn blue, and if you combine an indicator with an oxyden, nothing will happen. Suppose then that you see a mystery vial and (being somewhat foolhardy) you combine the contents of the mystery vial with an acid. The mixture turns blue. Although you have only this single piece of data about the mystery vial, you might confidently conclude that the contents of the vial are an indicator and can be safely mixed with the oxyden. Your

* Corresponding author.
E-mail address: lschulz@mit.edu (L.E. Schulz).

0010-0277/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2008.07.017

belief in the causal law allows you to make inferences for which you have no direct evidence.

Suppose however that you are wrong. When you combine the contents of the mystery vial with the oxyden, the mixture explodes. This data is anomalous (and striking) with respect to your initial theory. Nonetheless, in the absence of any other challenge to the causal law, you might be reluctant to jettison your conviction that there are three causal types and three distinct causal relations. You might instead decide that the test tube in which you mixed the substances was contaminated with acid. That is, you might introduce a new hypothesis (about the existence of an unobserved entity) in order to reconcile the otherwise anomalous data with the abstract causal law. Thus constraints on hypotheses imposed by higher-order causal laws might account for the paradoxical combination of rapid, accurate inferences from minimal evidence (e.g., Gopnik et al., 2001) and apparent intransigence to counterevidence (Koslowski, 1996; Kuhn, 1989; Kuhn, Amsel, & O'Laughlin, 1988; Schauble, 1990) that seems to characterize much of causal reasoning.

Although some abstract principles might be derived from innate, domain-specific modules (Leslie, 1994) or innate concepts in core domains (Carey, in press; Keil, 1995; Spelke, Breinlinger, Macomber, & Jacobson, 1992), innate knowledge is unlikely to explain our knowledge of abstract laws about, for instance, chemical substances. One possibility is that such abstract causal laws are only learned very slowly, after repeated exposure to many specific causal relations. Indeed, Piagetian accounts of causal learning suggest that children first learn concrete causal relationships and only gradually construct more abstract theories (1930, 1953). Counter to such accounts however, evidence suggests that children frequently understand abstract principles in tandem with, or even before they understand the specific causal relations that underlie them. For instance, children seem to understand that causal mechanisms play an important role in mediating effects (e.g., if an object appears to move spontaneously, children will insist that "something made it go", positing even "invisible batteries", Gelman & Gottfried, 1996) despite having a limited and inaccurate understanding of many common causal mechanisms (e.g., Lehrer & Schauble, 1998; Rozenblit & Keil, 2002). Similarly, children seem to infer that the insides of animals are important to the animal's identity long before they have any clear understanding of what those insides are or do (Gelman, 2006; Keil, 1989). As some researchers have put it, young children's inferences often seem to be "sensible before being accurate" (Wellman & Gelman, 1992). However, these observations pose a 'chicken-and-egg' problem: if children rely on abstract, higher-order causal knowledge in order to constrain their inferences about specific causal relations, how do children acquire abstract causal knowledge so early, seemingly before they have extensive knowledge of specific causal facts?

Recent computational models (Kemp, Goodman, & Tenenbaum, 2007; Mansinghka, Kemp, Tenenbaum, & Griffiths, 2006; Tenenbaum, Griffiths, & Kemp, 2006; Tenenbaum & Niyogi, 2003) have suggested a formal Bayesian account of how one might infer abstract causal laws from minimal data about specific causal relationships. Yet to date there have been no empirical studies looking at whether children can rapidly learn abstract causal laws consistent with the predictions of these formal analyses. This is the goal of our work here. We address two specific questions. First, can children learn abstract causal laws from very minimal data? To address this question, we focus on the signature phenomenon of intransigence: people's tendency to invoke auxiliary hypotheses in order to save an accepted theory from apparent counterevidence (even at the cost of forcing a more complex interpretation of the apparent counterevidence). If we find that children are intransigent about causal laws recently inferred from minimal data, we can take this as evidence that children rapidly commit to abstract causal laws. However, this question leads quickly to a second one: how effective can a learning mechanism be if it quickly leaps to abstract conclusions and maintains those conclusions in the face of counterevidence? The second part of our argument is an idealized Bayesian analysis of learning abstract causal knowledge, which explains at least in outline form why and under what conditions it can be rational, both to infer abstract causal laws from minimal data and to be relatively conservative about revising such laws in the face of potentially anomalous evidence. This analysis motivates our experiment and frames our predictions; thus we begin with it and then turn to experiments showing that children learn abstract causal knowledge in accord with the prescriptions of this analysis.

From a Bayesian standpoint, the question of why one should infer abstract principles from sparse data and why these principles should be maintained in the face of potential counterevidence can be seen as two aspects of the same problem. The principles of induction that support inferences about abstract laws also prescribe their degree of entrenchment: the conditions under which they should be conserved or over-turned. When and why should one infer more abstract laws from a handful of specific observations? The answer, formalized in hierarchical Bayesian models, is that one should infer abstract laws when the observed evidence is more probable given the laws than without them. Because these analyses have been the subject of several recent papers (Kemp et al., 2007; Mansinghka et al., 2006; Tenenbaum & Niyogi, 2003; Tenenbaum et al., 2006), we only provide a sketch in intuitive terms here.

Imagine for instance, that you see a handful of trials in which red blocks activate blue blocks (causing a noise or lights) on contact. The observed pattern of evidence is highly probable given the abstract law: "red blocks activate blue blocks"; this law uniquely specifies the relations between the blocks (each red block will activate each blue block), and these relations in turn specify the few observed activations. By contrast, if each block were construed as a unique individual, many other specific relations would be possible and many other patterns of activation could have occurred and should have been observable in the data (e.g., each red block might or might not activate each blue block). Thus despite having evidence about only about a few specific blocks, we infer a more abstract law because the observed evidence is

more probable if the abstract law is true than if no abstract law holds.

The inference that, all else being equal, a hypothesis that can predict only the observed pattern of evidence is more probable than a hypothesis that predicts many different patterns of evidence, only some of which are observed, is an example of a principle known as *Bayesian Occam's razor*. This principle follows from Bayes' rule. According to Bayes' rule, one's belief in a hypothesis after observing evidence is partly determined by the likelihood of the evidence; the likelihood of the evidence will be smaller if many other patterns of evidence could have been generated by the hypothesis. This principle can be captured, with many additional subtleties, in the mathematics of probability (Jefferys & Berger, 1992). In what follows, we will also appeal to Bayesian Occam's razor to consider how abstract causal laws should be balanced against potentially anomalous data.

To evaluate the hypothesis that (1) children will spontaneously use evidence about a few specific causal relationships to infer more abstract causal laws and that (2) these inferences immediately constrain children's interpretation of subsequent evidence such that the laws will be relatively robust to potentially anomalous data, we show preschoolers data about novel objects in three phases (setting up a situation similar to the chemistry-lab example above). In the initial phase we introduce children to a few specific causal relations (see Fig. 1). For instance, in Conditions A and B, we show children that when a red block contacts a blue block the blue block emits a train noise and that when a blue block contacts a yellow block, the yellow block produces a siren noise. We also show children that no other causal relations obtain (i.e., red blocks do not activate yellow blocks). This provides evidence for specific causal relations among the blocks. However, this also provides minimal evidence for a more abstract law: that there are three *kinds* of blocks, instantiating distinct causal relations.

Note that intrinsic features (e.g., the color of the blocks) can be useful in identifying an entity's causal kind. However previous studies suggest that children are willing to override perceptual cues to kind membership – both to extend causal inferences on the basis of kind membership and to infer kind membership on the basis of causal features (Gelman & Markman, 1986; Gopnik & Sobel, 2000; Nazzi & Gopnik, 2003). Children shown for instance, that a red block called a 'blicket' activates a toy, will infer that a blue block called a blicket will activate the toy rather than a red non-blicket; similarly, children shown that a red block called a blicket activates a toy will extend the label 'blicket' to a blue block that activates the toy rather than to an inert red block. This ability to override perceptual features to causal kind membership is important given that perceptual features are not always reliable cues to causal kind. Thus for instance, the color of a cabbage is informative about whether or not its juice can act an indicator dye (purple cabbage juice can; green cabbage juice cannot) but the shape of the cabbage is not (both cabbages have the same shape).

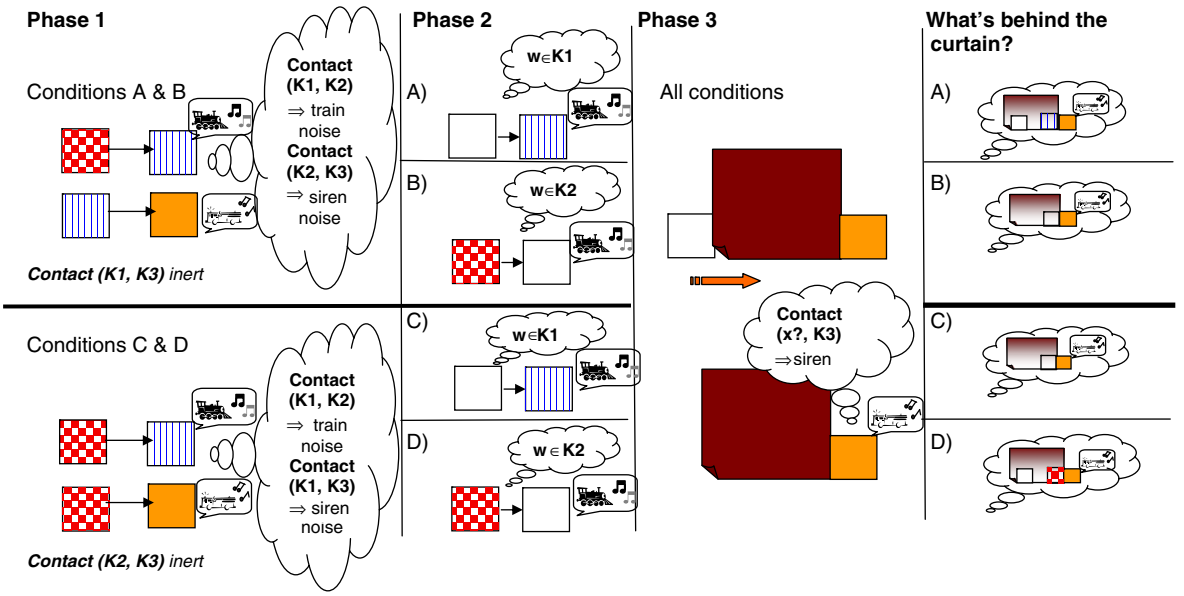

Fig. 1. *Design of Experiment 1.* Children were tested in one of four conditions, A, B, C or D, labeled below (n = 16 per condition). In Phase 1, children received evidence about red (here represented as checked), blue (represented as striped), and yellow (represented as solid) blocks consistent with there being three kinds of blocks, $K1$, $K2$, and $K3$, respectively. The evidence was consistent either with the causal law: Contact($K1$,$K2$) ⇒ train noise & Contact($K2$,$K3$) ⇒ siren noise or with the causal law: Contact($K1$,$K2$) ⇒ train noise and Contact($K1$,$K3$) ⇒ siren noise. Children also received evidence that all other contact relations were inert. In Phase 2, children saw evidence that a new white block plausibly played either the $K1$ or $K2$ role. In Phase 3, all children saw the white block go behind a curtain and the yellow block activate. If the children inferred the abstract laws in Phase 1, then the evidence in Phase 2 should lead to different inferences about whether there are one or two blocks behind the curtain.

We predict that if children treat intrinsic features as diagnostic, but not constitutive, of an entity's causal kind, they should be willing to use causal information to assign a novel, perceptually distinct, object to an existing kind, even if the different causal kinds are not established for the children in advance (i.e., even in the absence of established categories like blicket and non-blicket). There is some suggestive evidence that children may do this. Previous research (Nazzi & Gopnik, 2003) suggests that children as young as 30 months will pair perceptually dissimilar objects together based on their causal properties, even in tasks that do not involve naming (i.e., if told only, "give me the one that goes with this one"). However, in that study, the children might simply have identified objects with common causal properties; previous research falls short of showing that children can spontaneously posit a novel common kind for perceptually distinct objects.

In the current study, the initial learning phase provides evidence about specific causal relations that are also consistent with an abstract causal law: there are three causal kinds of blocks, $K1$, $K2$, and $K3$ such that (in Conditions A and B) a $K1$ contacting a $K2$ causes a train noise, and a $K2$ contacting a $K3$ causes a siren noise, but contact between two blocks of the same kind, or contact between a $K1$ and $K3$ has no effect. See Table 1 for details.

Phase 2 introduces children to a novel white block, and one of two pieces of evidence about this novel block. In Condition A, children see the novel block, w, contact a blue block, b, resulting in a train noise: that is, they see a specific instance of the causal relation, Contact$(w,b) \Rightarrow$ train noise. This evidence can be reconciled with the initial causal law if the novel block is in kind $w \in K1$ (and $b \in K2$). In Condition B, children see the novel block contact a red block, r, and produce a train noise: Contact$(r, w) \Rightarrow$ train noise. This evidence can be reconciled with the causal law if $w \in K2$ (and $r \in K1$). If children have learned the abstract causal law and reconciled the evidence of Phase 2 with this law by assigning the novel block to a known causal kind ($w \in K1$ in Condition A, $w \in K2$ in Condition B), then children should expect the novel block to have the other causal relations invested in that kind. In particular, in Condition B children should expect that if the novel block contacts a block of kind $K3$ (i.e., a yellow block) a siren noise will sound; in Condition A, contact between the novel block and a block of kind $K3$ should not cause a noise.

Table 1
The abstract causal law supported by the initial learning phase in Conditions 1 and 2 in the notation of first order logic (the law is similar in Conditions 2 and 3; see text for details)

$\exists$Types $K1$, $K2$, $K3$
$\forall\ r \in K1,\ b \in K2$ Contact$(r,b) \Rightarrow$ train noise
$\forall\ b \in K2,\ y \in K3$ Contact$(b,y) \Rightarrow$ siren noise
$\forall\ r \in K1,\ y \in K3$ Contact$(r,y) \Rightarrow \varnothing$
$\forall\ r, r' \in K1$ Contact$(r,r') \Rightarrow \varnothing$
$\forall\ b, b' \in K2$ Contact$(b,b') \Rightarrow \varnothing$
$\forall\ y, y' \in K3$ Contact$(y,y') \Rightarrow \varnothing$

To test whether children make this inference we present them with ambiguous evidence in Phase 3, the test phase: the white block goes behind an occluder and a yellow block adjacent to the occluder activates. Either the yellow block activated spontaneously (a hypothesis that is improbable given Bayesian Occam's razor: if blocks activated spontaneously, this evidence could have been observed in Phases 1 and 2; see also the discussion of determinism below) or some unknown entity, x, activated the yellow block: Contact$(x,y) \Rightarrow$ siren noise for some block x. The question is: what activated the yellow block? In Condition B the inference that the only block behind the occluder is the novel block, $x = w$, is consistent with the causal law; in Condition A this is not possible (since the law implies Contact$(w,y) \Rightarrow \varnothing$). This suggests that x is an unobserved block of kind $K2$. Thus, in Condition A, but not in Condition B, children who have learned and transferred at the abstract level should expect two blocks behind the curtain: the novel block and an unobserved blue block (i.e. of kind $K2$). In contrast, if children have learned only at the level of specific causal relations, they have no reason to expect the presence or absence of any specific causal relations for the novel block; the novel block might plausibly have its own unique set of causal properties. In this case the evidence in the test phase is innocuous in both conditions; children should not expect an unobserved block any more in Condition A than in Condition B.

The explanation described above for the evidence of Phases 2 and 3 is uniquely compatible with the abstract causal law, which itself is consistent with the evidence in the initial learning phase. However, children might also account for the anomalous evidence (in Condition A) by revising the causal law. There are several ways children might revise the causal law and these alternatives would not require children to posit an unobserved block in Condition A. We will argue that a rational learner should see each of these alternative explanations as less likely than the assumption that there is an unobserved block. Note however, that as long as no alternative explanation of the anomalous evidence is *more* likely than the inference of an unobserved block, children in Condition A, but not Condition B, will have reason to suspect that there is an unobserved block hidden behind the curtain. That is, under a rational Bayesian analysis, the learner should consider how well different alternative hypotheses account for the data. We suggest that the most plausible alternative hypotheses are not more plausible than the hypothesis that there is an unobserved block.

One way that children might explain the anomalous evidence in the test condition is by revising the causal laws relating the three kinds. To accommodate the evidence the laws must be altered to allow either a block of kind $K1$ contacting a block of kind $K3$ to sometimes cause $K3$ to produce a siren noise (hence, $w \in K1$ and $x = w$), or to allow a block of kind $K2$ when contacting another block of kind $K2$ to sometimes cause a siren noise (hence, $w \in K2$ and $x = w$) Notice, however, that each of these possible revisions replaces a deterministic causal law with a stochastic one. Previous research suggests that children assume that physical causal relations

are deterministic rather than stochastic (Schulz & Sommerville, 2006). This suggests that children will consider the revised laws less likely than the original laws. Further, in this situation, such a judgment is normatively correct: the stochastic law would have to account for why the causal interactions appeared to be deterministically effective in the initial learning phase (Phase 1). If the laws are genuinely probabilistic then the fact that children initially got deterministic evidence must be construed as a suspicious coincidence. (We may formalize this as a Bayesian Occam's razor argument: the evidence of the initial learning phase is fairly unlikely under the stochastic law but is uniquely determined, and thus very likely, under the deterministic law. By Bayes' rule, the likelihood of the evidence contributes to the posterior degree of belief in each law, and favors the deterministic one.)

The second kind of revision children could make to the causal law is to posit that the novel block belongs to a new causal kind. Indeed, if w belongs to a new kind, $K4$, and this kind participates in the correct causal laws ($\forall\ w \in K4$, $b \in K2$ $\text{Contact}(w,b) \Rightarrow$ train noise and $\forall\ w \in K4$, $y \in K3$ $\text{Contact}(w,y) \Rightarrow$ siren noise), then the evidence in Condition A is accounted for without the need for an unobserved block. However, positing an additional kind comes at the price of complexity, in two ways. First, the prior probability of there existing N causal kinds (within an abstract causal law) decreases with N, hence the original law, with three kinds, is a priori more likely than the revised law, with four kinds. Second, the likelihood that any particular object belongs to a given kind is lower when there are more kinds, thus the evidence from the initial learning phase (with all objects occupying only three kinds) is less likely under the revised law, and that law is in turn less likely given the evidence. (The latter argument is again a Bayesian Occam's razor.)

Therefore, while it is possible to accommodate the evidence without positing an unobserved object, any revision of the causal law that does so is more complex than the original causal law, hence less probable (in a sense that can be made precise in the language of Bayesian probability). Of course, the addition of an unobserved object is itself an added complication. For the reasons discussed, we believe this complication is more likely than the alternatives, however the critical point is that it is not less likely. As long as the hypothesis requiring the unobserved object is as likely as any other, children should entertain the possibility that such an object might be present.

So far we have presented this discussion only with reference to Conditions A and B. However, given only Conditions A and B, arguably some feature of the Phase 2 evidence itself might make children more likely to posit additional blocks (e.g., perhaps an agent role for the novel block is more interesting than the patient role, thus making the children more likely to posit additional blocks). We therefore also test children in two additional Conditions, C and D, which are formally identical to Conditions A and B except that the causal law is different. In Conditions C and D, we show children that contact between a red block and a blue block produces a train noise and that contact between a red block and a yellow block produces a siren noise (see Fig. 1). We also show children that no other causal relations obtain (i.e., the blue block does not activate the yellow block). Thus Conditions C and D provide evidence supporting different inferences about the agent/patient relationships than Conditions A and B. In Conditions A and B, the red block is a causal agent and the blue block is a causal patient in one interaction; in the other, the blue block is an agent and the yellow block is a patient. By contrast, in Conditions C and D there is only a single agent, the red block.[1] If children's inferences about the initial evidence in Phase 1 affects their interpretation of subsequent evidence, then, given the same evidence in Phases 2 and 3, children in Conditions C and D should draw different conclusions than children in Conditions A and B. In Condition C, when children see that the interaction between the novel white block and the blue block causes a train noise, they should infer that the white block is the same kind of block as the red block and thus could activate the yellow block (and thus they should not infer an unobserved block in the test condition). By contrast, if children see that the interaction between the novel white block and the red block causes the train noise (Condition D), children should infer that the white block is the same kind of block as the blue block and thus could not activate the yellow block; thus they should infer an unobserved block in the test condition.

Collectively, these four conditions allow us to investigate the interaction of children's (immediate) prior knowledge on their evaluation of subsequent evidence. Since the evidence in Phase 2 is identical in Conditions A and C, and Conditions B and D, if children do infer the presence of an unobserved block in Conditions A and D but not B and C, it suggests that children's inferences about the initial causal law affects their interpretation of the very next evidence they observe. In the experiments to follow, we look at whether children infer abstract causal laws from minimal data by looking at whether children infer the presence of unobserved entities to account for evidence that is innocuous if construed as pertaining to specific causal relations but is anomalous if construed with respect to an abstract causal law.

[1] The touching relationships were always asymmetric – the agent block was always moved into the (stationary) patient block in order to activate the noisemaker inside the patient block. We believe the children construed the causal relationships asymmetrically (i.e., as agent/patient relationships) because when the contact relations were inert, children frequently asked questions like, "Why didn't the red one make the yellow one go?" It is possible however, that the children might have believed that both blocks played equivalent roles in the interaction. Under this construal, nothing substantive changes in either the theoretical discussion or the empirical predictions for the comparison between Conditions A and B or the comparison between conditions C and D. However, the comparison between Conditions A and B on the one hand and conditions C and D on the other would change. Rather than supporting inferences about two different agent/patient relations (in Conditions A and B, r as an agent and b as the patient in one interaction; b as an agent and y as the patient in the other as opposed to, in Conditions C and D, r as the agent and b and y as patients) the causal relationships would be construed as symmetric interactions in all four conditions. Under this construal, rather than introducing a new causal law, Conditions C and D would simply replicate Conditions A and B (with the red block playing the role of the blue block). The other points continue to hold however: children in Conditions C and D should draw opposite conclusions to the children in Conditions A and B given the same evidence in Phases 2 and 3.

2. Experiment 1

2.1. Method

2.1.1. Participants

Sixty-four preschoolers (mean: 57 months; range: 49–64 months) were tested individually in a quiet corner of a large metropolitan science museum. Sixteen children were randomly assigned to each of four conditions (see Fig. 1). Three children in Condition A, one child in Condition C, and two children in Condition D were dropped from the study and replaced due to failure to encode the specific causal relations, see below.

2.1.2. Materials

Seven 7 cm × 7 cm × 7 cm electronically activated blocks were used. Some of the blocks contained a noisemaker that activated when the noisemaker block was contacted by a block that activated a magnetic switch. The blocks were covered with colored (or white) paper to create two red blocks, two blue blocks, two yellow blocks and one white block. In Conditions A and B, contact between a red and blue block caused the blue block to emit a train noise and contact between a blue block and a yellow block caused the yellow block to emit a siren noise. In Conditions C and D, contact between a red and a blue block caused the blue block to emit a train noise and contact between a red block and a yellow block caused the yellow block to emit a siren noise. In Conditions A and C, contact between the white block and a blue block caused the blue block to emit a train noise. In Conditions B and D, contact between the white block and a red block caused the white block to emit a train noise. All other contact relations were inert. A black cardboard stage (45 cm × 32 cm × 15 cm) was also used. A red string curtain covered the front face of the stage so that children could not see into the stage but could reach through the curtain to the back of the stage. Two screens could be lifted on either side of the stage so that blocks could be slid into the stage from the side. The back of the stage, facing the experimenter, was open.

2.1.3. Procedure

2.1.3.1. Warm-up. We used a manual search task (loosely adapted from Feigenson & Carey, 2003) as an implicit measure of children's belief that an unobserved variable was present. We chose the manual search task both because we wanted a method that could later be adapted for use with younger children and because research suggests that children's implicit understanding of causal relationships may exceed their ability to offer explicit explanations (Chen & Klahr, 1999; Dunbar & Klahr, 1989; Inhelder & Piaget, 1958; Kuhn, 1989; Kuhn, Amsel, & O'Laughlin, 1988; Masnick & Klahr, 2003). The warm-up task familiarized children with reaching into the stage.

The experimenter showed the child a container containing the red, blue, and yellow blocks. She asked the child to count the blocks and name their colors. She then hid the container on a stool under the table, out of the child's sight, and inserted one or two of the blocks (color chosen at random; order of one-block, two-block reaches counterbalanced between children) into the stage from behind. She returned the container to the table so the child could see which block(s) were missing. She asked the child, "What do you think is behind the curtain? One block or two? Go ahead and see." The child was asked to reach into the stage to retrieve the block(s). The warm-up continued until the child had retrieved all six blocks.

2.1.3.2. Training and test. The experimenter told the child, "Let me show you what these blocks can do." She returned the container to the stool under the table (where it remained for the rest of the experiment) and retrieved two blocks at a time. In Conditions A and B, the experimenter pushed the red block into the blue block; the blue block emitted a train noise. The child was then allowed to place her hands on the blocks to try it along with the experimenter. (The experimenter 'helped' on the child's turn to ensure that all children saw identical evidence). The experimenter removed those blocks from view and then pushed a blue block into a yellow block so that the yellow block emitted a siren noise; again, the child was allowed to try. Finally, the experimenter pushed a red and a yellow block together, two red blocks together, two blue blocks together, and two yellow blocks together, showing the child that all other combinations of blocks were inert. All the demonstrations were repeated a second time. Conditions C and D were identical except that the child received evidence that the red block caused the blue block to make a train noise and caused the yellow block to make a siren noise. Again no other causal relations obtained. To verify that the children learned the specific causal relations, the experimenter selected two blocks at a time and asked the child to predict the outcome of each unique combination of red, yellow and blue blocks. Children who failed to learn the specific causal relations were dropped from the study and replaced. Thus at the end of Phase 1, children had seen a total of four trials of evidence for each combination of red and blue, blue and yellow, and red and yellow blocks.

The experimenter then reached under the table and retrieved the white block. She said, "Oh look. Here's a new, mystery block. Let's see what it does." In Conditions A and C, the experimenter pushed the white into the blue block causing the blue block to emit a train noise. In Conditions B and D, the experimenter pushed the red block into the white block causing the white block to emit a train noise. In all conditions, the experimenter then removed all the blocks from the table, leaving only the white block visible. She placed a yellow block next to one open side of the stage. She then lifted the doors on both sides of the stage and slid the white block into the stage so it was no longer visible (i.e., so that it could conceivably contact the yellow block on the far side of the stage). The yellow block emitted a siren noise. As in the training, she asked the child, "What do you think is behind the curtain? One block or two? Go ahead and see." The experimenter looked at her lap to avoid cuing the child and the child was allowed to reach into the box.

In all conditions, the experimenter surreptitiously inserted the necessary block (blue in Conditions A and B; red in Conditions C and D) into the stage to activate the yellow block. In all conditions, she then surreptitiously

moved that block to the back of the stage, ensuring that all children would retrieve the white block on their first reach. After children removed the white block, the experimenter always returned the target block to the front of the stage so that children would retrieve the target block if they reached a second time. Trials were terminated when the child did not reach into the stage for 3 consecutive seconds.

2.2. Results and discussion

Although all the children had seen only the white block go behind the curtain, children inferred the presence of an additional, unobserved block when the evidence was anomalous with respect to the causal laws (i.e., in Conditions A and D). As predicted, children were more likely to reach twice in Condition A than in Condition B (69% vs. 6%; $\chi^2(1, n = 32) = 13.33$, $p < .001$) and in Condition D than in Condition C (50% vs. 6%; $\chi^2(1, n = 32) = 7.57$, $p < .01$). There were no differences between Conditions A and D or B and C. Within Condition A, children were marginally more likely to reach twice than once ($n = 16$, $p = .07$ by binomial test) whereas within Condition B, children were more likely to reach once than twice ($n = 16$, $p < .001$ by binomial test). Similarly, within Condition C, children were more likely to reach once than twice ($n = 16$, $p < .001$ by binomial test) whereas in Condition D, children were as likely to reach twice as once ($n = 16$, $p = ns$ by binomial test).

Although we designed the experiment to look at children's implicit inferences, almost all the children (54 of the 64 children; 84%) also gave verbal responses to the (rhetorical) question "What do you think is behind the curtain? One block or two?" All verbal responses were consistent with the children's reaching behavior and the predicted results held when we considered only verbal responses. Of the children who gave verbal responses, children were more likely to say there were two blocks in Condition A than in Condition B (82% vs. 7%; $\chi^2(1, n = 25) = 14.31$, $p < .001$) and in Condition D than in Condition C (50% vs. 0%; $\chi^2(1, n = 29) = 9.89$, $p < .01$). There were no differences between Conditions A and D or B and C. Within Condition A, children were more likely to say there two blocks than one ($n = 11$, $p < .05$ by binomial test) whereas within Condition B, children were more likely to say there was one block than two ($n = 14$, $p < .001$ by binomial test). Similarly, within Condition C, children were more likely to say there was one block than two ($n = 15$, $p < .001$ by binomial test) whereas in Condition D, children were as likely to say there were two blocks as one ($n = 14$, $p = ns$ by binomial test).

Finally, some of the children interrupted the experimenter after she asked, "What do you think is behind the curtain?" to volunteer more specific predictions. Two children in Condition A and two children in Condition D spontaneously mentioned the correct (unobserved) target block (e.g., "the blue one!"); two other children in Condition A and four in Condition D specifically mentioned the (observed) white block; no children mentioned any other block. By contrast, in Condition B, 9 of the 16 children and in Condition C, 8 of the 16 children spontaneously mentioned the (observed) white block and none of the children named a block of any other color. That is, children were more likely to spontaneously (and correctly) reference the target non-white block in Conditions A and D than B and C ($\chi^2(1, n = 64) = 4.27$, $p < .05$) and were more likely to mention the observed white block in Conditions B and C than A and D ($\chi^2(1, n = 64) = 8.21$, $p < .01$).

These results are consistent with the hypothesis that children used the initial evidence about the red, blue, and yellow blocks to infer the existence of three kinds of blocks and the causal laws relating the blocks. This abstract inference constrained their interpretation of the next evidence they observed: the interaction between the novel white block and one of the known blocks. When children saw evidence that was inconsistent with the abstract causal law, they inferred the presence of an unobserved cause to maintain the law.

Note that in the anomalous evidence conditions, only some of the children inferred the existence of an unobserved block (69% in Condition A; 50% in Condition D). We will discuss possible explanations for this mixed responding in Section 4. One concern however, is that children were always asked whether there were one or two blocks behind the curtain. It is possible that although the children did distinguish the anomalous evidence from the predicted evidence (hence the differences in responding between Conditions A and B, and between Conditions C and D), children might not have actually posited the presence of an unobserved block in response to the data; rather, the anomalous evidence might have merely confused the children and led them to choose at chance.

In Experiment 2, we address this concern by replicating Conditions A and B of Experiment 1, but eliminate the forced choice question (" ... one block or two blocks?") and include two explicit dependent measures, in addition to the implicit reaching measure. Specifically, we ask the children "What do you think is behind the curtain?" and allow them to respond both verbally and by pointing to pictures of the blocks. If children are confused by the anomalous evidence and decide to name and/or point to blocks other than the white block at random, they should be as likely to name a red block as a blue block (given that the white block has disappeared behind the curtain and the yellow block is visible to one side of the curtain). However, if children are genuinely positing an unobserved variable in order to explain the otherwise anomalous data, then children should be more likely to identify two blocks in Condition A than Condition B, and should selectively indicate the blue block (in addition to the white block) rather than the red block.

Note also that in these studies, the children are not given any information about the mechanism behind the blocks' activation. They are simply shown an intervention on the blocks, immediately followed by the target noise. Many researchers have suggested that neither adults nor children draw causal inferences merely from covariation information (Ahn, Kalish, Medin, & Gelman, 1995; Koslowski, 1996; Shultz, 1982). Arguably therefore, in the absence of mechanism information, children might treat these events as arbitrary associations, rather than genuinely causal events.

However, many other researchers have suggested that it is possible to have genuine causal knowledge without much understanding of causal mechanisms (e.g., Gopnik et al., 2004; Keil, 2003; Schulz, Gopnik et al., 2007). Recently, philosophers, psychologists, and computer scientists (Gopnik & Schulz, 2004; Pearl, 2000; Spirtes, Glymour, & Scheines, 2001; Woodward, 2003) have suggested that the crucial piece missing from both the covariation and the mechanism accounts is the notion of intervention. Specifically, researchers have suggested that knowing that X directly causes Y means knowing that, all else being equal, intervening to change X can change Y (Gopnik & Schulz, 2007; Schulz, Gopnik et al., 2007; Woodward, 2003). Such an interventionist account of causation underlies the principle of experimental design in science: even if we do not know the mechanisms underlying a relationship between events, we can often use interventions to infer that a causal relation exists (as for instance, when we infer a link between serotonin levels and depression because manipulating one changes the other).

Goal-directed human actions have many of the formal criteria for causal inference (e.g., they are exogenous interventions, typically unconfounded by other potential causes), and considerable evidence suggests that preschoolers are willing to treat effects that follow goal-directed human actions as caused by those actions, even when artificial, arbitrary stimuli are used. Thus for instance, if an adult intentionally places a block on a toy and the toy lights up, children infer that the adult's action caused the toy to activate (Gopnik & Sobel, 2000; Kushnir & Gopnik, 2005); similarly, if an adult slides a knob and a toy activates, children assume the sliding knob activated the toy (Schulz, Hooppell, & Jenkins, 2008). By assuming that novel effects are the consequences of the immediately preceding intentional action, even very young children may be able to infer the cause of events that would otherwise be mechanically far too complex for them to understand (e.g., that flipping a switch turns on a light or that pushing a remote turns on a TV). If children do infer that intentional actions are the cause of immediately subsequent, otherwise unexplained events, the children in Experiment 1 should infer that the intentional action to contact one block with the other as the cause of the target noise. In Experiment 2, we test this intuition by asking the children to perform an intervention of their own to "make the block go". If children infer that the contact relation caused the block to make noise, they should readily perform the target action on the blocks.

3. Experiment 2

3.1. Method

3.1.1. Participants

Twenty-four preschoolers (mean: 57 months; range: 48–66 months) were tested individually in a quiet corner of a large metropolitan science museum. Twelve children were randomly assigned to each of two conditions: A and B. One child assigned to Condition B became distracted before completing the training and was replaced.

3.1.2. Materials

The red, blue, yellow, and white blocks and the stage from Experiment 1 were used as stimuli. The same causal relations obtained as in Experiment 1: a red block hitting a blue block caused the blue block to emit a train noise, and a blue block hitting a yellow block caused the yellow block to emit a siren noise. In Condition A, the white block hitting a blue block caused the blue block to emit a train noise; in Condition B, the red block hitting the white block caused the white block to emit a train noise. All other combinations were inert. Four 8 cm $\times$ 6 cm pictures, each illustrating one of the colored blocks against a black background, were also used.

3.1.3. Procedure

The procedure was identical to that in Conditions A and B of Experiment 1, with three exceptions. First, a pointing measure was added to the warm-up and test trials. Whenever a child was asked what he or she thought was inside the stage, pictures of the blocks were presented to the child. Before reaching to retrieve the actual blocks, the child was asked to point to the picture(s) of what was inside. Second, to determine whether children encoded the relations as causal, at the end of Phase 1, each of the 'patient' blocks (blue and yellow) were placed on the table in turn and children were asked "Can you make the blue block make the train noise?" and "Can you make the yellow block make the siren noise?" (order counterbalanced between children). Children were then shown the evidence as in Phase 2 of Experiment 1, Conditions A and B. In Phase 3, the experimenter placed the pictures in front of the child and asked, "What do you think is behind the curtain?" Children had the opportunity to verbalize their responses and/or point to the pictures before the experimenter asked them to reach behind the stage. We coded three dependent measures: children's verbal response (if any), their picture choice response, and their reach response. As in Experiment 1, the trial terminated when the child did not reach into the stage for three consecutive seconds.

3.2. Results and discussion

All but two children (one in Condition A and one in Condition B) responded to the prompt to make the blocks go by immediately reaching for the correct 'agent' blocks (red and blue) and using them to contact the 'patient' blocks (blue and yellow). The two children who did not respond correctly attempted the intervention on the blocks but chose the wrong blocks. These children spontaneously corrected themselves when they saw that their action was ineffective. This suggests that children genuinely treated the contact events as causally effective interventions.

As in Experiment 1, although all the children had seen only the white block go behind the curtain, children inferred the presence of an additional, unobserved block when the observed evidence was anomalous with respect to the causal law (Condition A) but not when the observed evidence was consistent with the inferred causal law (Condition B). Eleven of the 12 children (92%) in Condition A, and 10 of the 12 children (83%) in Condition B gave verbal responses when asked, "What's behind the curtain?" Of the

children who gave verbal responses, children were more likely to mention two blocks (either by color or by saying "two blocks") in Condition A than in Condition B (55% vs. 10%; $\chi^2(1, n = 21) = 4.68$, $p < .05$). Within Condition A, children were as likely to refer to two blocks as one ($n = 11$, $p = ns$ by binomial test) whereas within Condition B, children were more likely to refer only to one block than two ($n = 10$, $p = .01$ by binomial test).

All of the children gave pointing responses (i.e., including the three who did not give verbal responses). Children's pointing responses were consistent with their verbal responses (that is, if the children referred to two blocks, they pointed to two blocks and if they named the colors of the blocks, they pointed to those colors; if they said there was one block, they pointed to one). In Condition A, seven of the twelve children pointed to both the white and the blue blocks; the remaining children pointed only to the white block. In Condition B, two children pointed only to the red block, one child pointed to the white and the yellow (visible) block, and the remaining children pointed only to the white blocks. Children were significantly more likely to point to the unobserved blue block in Condition A than the unobserved red block in Condition B (58% vs. 17%; $\chi^2(1, N = 24) = 4.44$, $p < .05$).

Finally, replicating Experiment 1, children were more likely to reach twice in Condition A than Condition B (58% vs. 0%; $\chi^2(1, n = 24) = 9.88$ $p < .005$). Children were as likely to reach twice as once in Condition A ($n = 12$, $p = ns$, by binomial test) but significantly more likely to reach once than twice in Condition B ($n = 12$, $p < .001$ by binomial test). These results suggest that children are not merely confused by evidence that is anomalous with respect to previously inferred causal laws but genuinely posit a causally appropriate unobserved variable to explain anomalous evidence.

4. General discussion

The results of these studies suggest that from a few minutes of exposure to novel, specific causal relationships, very young children make more abstract inferences that are robust to potential counter-evidence. Given sparse data about individual entities, preschoolers seemed to spontaneously infer the existence of three causal kinds instantiating distinct causal relations. In the face of evidence that might challenge this causal law, children maintained the causal law by hypothesizing the presence of an unobserved entity.

Our findings add to the considerable body of research suggesting that very young children can make rapid and accurate causal inferences from small amounts of data (Baldwin et al., 1993; Gopnik et al., 2001; Gopnik et al., 2004; Kushnir & Gopnik, 2005; Meltzoff, 1995; Schulz, Gopnik et al., 2007). In particular, as noted, children will override perceptual dissimilarities to infer that objects with common causal rather than common perceptual properties are members of an established category (Gelman & Markman, 1987; Gopnik & Sobel, 2000; Nazzi & Gopnik, 2003). Our research extends these findings by suggesting that children can themselves posit the existence of kinds (e.g., *K*1, *K*2, and *K*3 in our studies) and can generalize the relevant causal inferences appropriately.

Additionally, our findings are consistent with recent research showing that even very young children infer the existence of unobserved causes in response to evidence that is otherwise inconsistent with their prior knowledge. For instance, 12-month-olds will infer the presence of an unobserved agent if an inanimate object appears to move spontaneously (Saxe, Tenenbaum, & Carey, 2005) and preschoolers will infer the presence of an unobserved inhibitory cause if a switch activates a toy probabilistically (Schulz & Sommerville, 2006). Importantly however, in those studies the evidence was anomalous with respect to fundamental, well-established assumptions about the physical world: the assumption that objects do not move spontaneously and the assumption that causes do not act unreliably. By contrast in the current study, all the evidence was consistent with the laws of physical causality. The evidence in Conditions A and D was only surprising if construed with respect to more abstract inferences that were themselves novel and based on minimal data. Of course, children's inferences in this experiment were also supported by considerable prior knowledge about the physical world (e.g., that objects are solid and move on continuous paths, that physical causes tend to act deterministically) and there is compelling evidence for the claim that some of this abstract knowledge about objects might be innate (Spelke et al., 1992). However, the results of this study suggest that abstract causal laws can arise, not just from innate constraints or extensive prior experience, but also rapidly, on-line, and in tandem with inferences about specific causal relations.

These findings have implications for computational accounts of children's causal reasoning. The primary contrast with our hierarchical Bayesian approach is the causal Bayes net approach, which has recently been influential in developmental theories of causal learning (see e.g., Gopnik et al., 2004). (Note that despite the potentially confusing overlap in terminology, causal Bayes net representations are distinct from and do not require Bayesian learning.) Some researchers (including the first author of this paper) have suggested that the learning algorithms associated with causal Bayes nets might support the representation and learning of abstract, coherent causal knowledge (Gopnik et al., 2004; Schulz, Gopnik et al., 2007; Schulz, Kushnir, & Gopnik, 2007; Sobel & Kirkham, 2007). Could children's inferences on our tasks be explained by the same learning algorithms?

Bayesian network learning algorithms can model complex causal structures, and can support effective predictions, interventions, and inferences about unobserved causes (see Glymour, 2001; Glymour & Cooper, 1999; Jordan, 1998; Pearl, 1988; Pearl, 2000; Spirtes, Glymour, & Scheines, 1993; Spirtes et al., 2001). On our tasks, they can explain some of the specific knowledge that children likely acquired in Phases 1 and 2, but they cannot generate the abstract knowledge necessary to explain the critical inferences children made in Phase 3. The latter would require either an ability to construct new nodes in the network that are based on inferred abstract variables rather than simply observed events, or an ability to posit appropriate new links

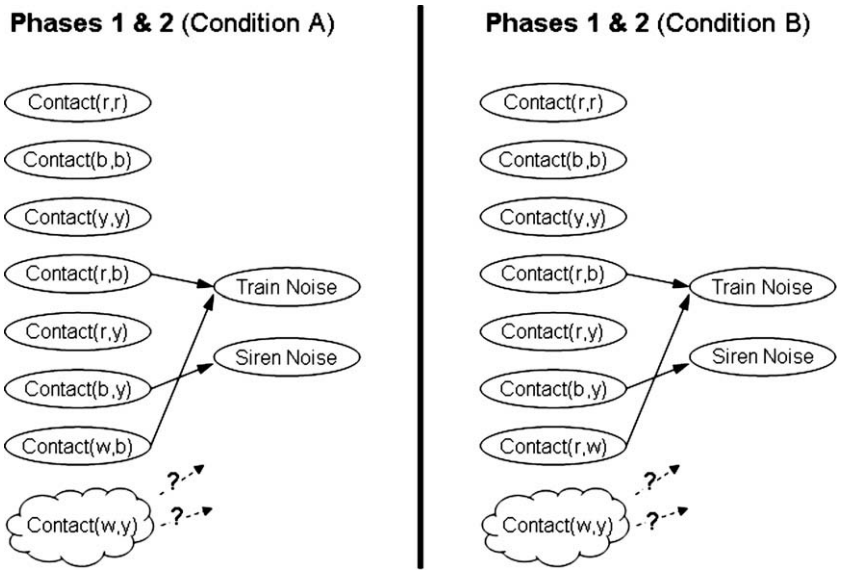

Fig. 2. *Causal Bayesian networks that could be constructed from the evidence observed in Experiment 1, Phases 1 and 2.* These networks represent the best possible causal models that children could acquire in Experiment 1, if they were learning using standard algorithms for learning causal Bayesian networks from observed data. The arrows indicate causal dependencies between variables and the nodes represent observed variables of interest: *r* is a red block, *b* a blue block, *y* a yellow block, and *w* a white block; Contact(*r*,*b*) denotes the event of a red block contacting a blue block, and similarly for all other contact events between blocks.

in the network for which there is no evidence in the observed data; standard Bayesian network learning algorithms have neither of these capacities.

Fig. 2 shows the best causal Bayesian networks that could be learned after Phases 1 and 2 in Conditions A and B, by standard Bayesian network learning algorithms. Nodes represent observed events: for instance, Contact(*r*, *b*) represents a contact event between a red block and a blue block. Causal links exist between each observed contact event that led to a sound and the appropriate sound, while there are no links coming out of observed contact events that appeared to have no causal effects. For instance, there are links from Contact(*r*,*b*) to Train Noise and from Contact(*b*,*y*) to Siren Noise, but no link from Contact(r,y) to any other event. Phase 2 provides evidence for one causal link, from Contact(*w*,*b*) (in Condition A) or from Contact(*r*,*w*) (in Condition B) to Train Noise. But after Phase 2, the network contains no representation of the effects of other contact events involving white blocks, because these events have never been observed. In particular (as illustrated in Fig. 2), the network has no expectations about the causal relations that might hold between Contact(*w*,*y*), the event in which a white block contacts a yellow block, and either of the two possible sound effects. These expectations are the key beliefs needed to explain preschoolers' interpretations of the evidence presented in Phase 3, and their differential predictions in Conditions A and B of the experiments. Because standard Bayesian network learning algorithms do not create new variables representing events defined on abstract categories of objects, or attempt to learn more abstract causal structure relating classes of variables, the learned network has no way to represent the inferred commonality between the red and white blocks in Condition A, or the blue and white blocks in Condition B; thus it has no expectation that a link between Contact(*w*,*y*) and Siren Noise is more likely in Condition B than in Condition A.

By contrast, recent work on hierarchical Bayesian inference models (Kemp et al., 2007; Mansinghka et al., 2006; Tenenbaum & Niyogi, 2003; Tenenbaum et al., 2006) provides an account of how more abstract relationships, including inferences about classes of variables and relationships among classes, might be inferred, even from sparse data. This experiment provides evidence that this additional level of representation, more abstract than causal Bayes nets, and consistent with inferences prescribed by the formal analysis, is learned and used by children even on first encountering novel evidence. Although the details are beyond the scope of this paper, more details of the computational approach, and explicit comparison with a few possible alternatives, can be found in Kemp, Goodman, and Tenenbaum (2008). However, while in principle there may be a number of specific mechanisms that could give rise to these inferences, we know of no other formal proposals for how such abstract knowledge could be learned, or for determining when, normatively, such knowledge should be preserved or revised.

Children's rapid learning and apparent resistance to counter-evidence highlight an apparent paradox of causal reasoning. We have suggested that preschoolers' ability to draw robust conclusions from minimal data is a remarkable feat of inductive inference. However many researchers cite children's (and adults') tendency to jump to conclusions and to ignore or discount data that challenges initial theories as evidence of the fallibility of everyday causal reasoning (Brewer & Chinn, 1991; Champagne, Klopfer, &

Anderson, 1980; Chi, 1992; Chinn & Brewer, 1992; Chinn & Brewer, 1993; diSessa, 1982; Klahr & Dunbar, 1988; Klahr, Dunbar, & Fay, 1990; Koslowski, 1996; Kuhn, 1989; Kuhn, Amsel, & O'Laughlin, 1988; Piaget, 1980; Samarapungavan, 1992; Schauble, 1990; Wisniewski & Medin, 1991). We want to suggest that this apparently paradoxical combination of rapid, accurate learning from minimal evidence and obduracy in the face of counterevidence arise from a common underlying process. Inductive biases constrain the hypothesis space for better and for worse. Our oft-noted difficulty in learning from data that violates our prior beliefs may be the inevitable byproduct of the inferential processes that support rapid learning initially.

In our experiment, for instance, the abstract causal law that there were three causal kinds instantiating distinct causal relations supported the prima facie surprising (albeit correct) hypothesis that an unseen block had been introduced behind the occluder. However, the inductive bias that led children to treat the white block as a member of one of the already inferred kinds necessarily implied that children would be less likely to infer that the novel block was a member of a new kind. Had the white block in fact been a new kind of block – a kind that could interact effectively with all the other kinds – the children might have seemed strangely resistant to learning from the data, stubbornly unwilling to accept even the seemingly straightforward evidence that the yellow block activated adjacent to the curtain where the white block had just disappeared. As philosophers have long noted (Goodman, 1983; Hume, 1748), inductive inference works best to the extent that nature is uniform; since that assumption often obtains, the benefit of rapid learning may be worth the occasional cost (or delay) in accuracy.

This is not of course to imply that young children's initial beliefs are inflexible. Consistent with Bayesian inference models, children's causal judgments seem to reflect a graded interaction of prior knowledge and evidence: preschoolers' prior beliefs affect their interpretation of evidence but evidence also affects children's beliefs. Thus for instance, children are less likely to draw accurate causal inferences from statistical data that violates their naïve theories (e.g., about contact causality or domain boundaries) than from data consistent with their prior beliefs – but they are significantly more likely to accept such theory-violating causal relationships than are children not exposed to such statistical evidence (Kushnir & Gopnik, 2007; Schulz, Baraff Bonawitz, & Griffiths, in press; Schulz & Gopnik, 2004). Moreover, children seem to be able to integrate even recently acquired information into their evaluation of subsequent evidence. If for instance preschoolers first observe that relatively few blocks activate a toy, and then see (A) that a red and a blue block together activate a toy and (B) that the red block by itself activates the toy, they tend to deny that the blue block is causally effective. However, if they first learn that activating blocks are common, they tend to infer that both the red and blue blocks are causally effective (Sobel, Tenenbaum, & Gopnik, 2004). Thus the current study adds to recent evidence suggesting that preschoolers are sensitive to the interaction between statistical data and prior knowledge and that this interaction obtains, not just in cases where children have substantial prior knowledge (e.g., about contact causality or about domain boundaries) but even when the prior knowledge is quite recent: immediately 'prior' to the very next evidence the children observe.

Note also, that although our experiments focused specifically on how abstract knowledge might support inferences about causal relations between objects, inferences about kind membership might support many other types of inductive inference (e.g., about the insides of objects, about object labels, about other perceptual properties, etc.). We do not mean to suggest that causal reasoning is the only type of reasoning supported by such abstract knowledge. Indeed, we expect that children's ability to infer abstract laws from minimal data supports inductive inferences in a wide variety of domains.

As noted however, many of the children in these experiments did not seem to suspect the presence of an unobserved block. There are several possible explanations for children's fragile responding. One possibility is that the children might have been confused by the task demands. In particular, some of the children might have interpreted the question "What is behind the curtain?" as a request to respond to the white block's conspicuous disappearance, rather than as a request to reason about the observed causal relations. Another possibility is that the competence posited in this paper might be fragile in preschoolers; some children might have learned only the specific causal relations among the blocks rather than the more abstract causal law. Alternatively, children might have inferred the more abstract causal law but failed to generalize it to the novel, white block – perhaps because they were unwilling to conclude that the two perceptually dissimilar blocks could be members of a common kind. Additionally, as noted, the hypothesis positing an unobserved block was as, or more plausible, than the alternatives – but it was not the only plausible hypothesis. Rather than positing the auxiliary hypothesis about the unobserved block, children might have revised the initial causal law by accepting that the white block was a member of a novel kind with novel causal properties, or by concluding that the blocks could activate spontaneously or stochastically. Future research might determine the extent to which children fail to infer abstract causal laws and/or fail to maintain them in the face of potentially anomalous data, and the extent to which children instead genuinely entertain alternative hypotheses.

More generally, future research might investigate how prior belief in a causal law, the strength of the anomalous evidence, and the availability of alternative hypotheses, interact to affect whether children respond to anomalous evidence by introducing auxiliary hypotheses or by revising their causal beliefs. Given that in this experiment children had only minimal data for the initial causal law, any strengthening of the evidence against this causal law might have led children to revise rather than maintain it. If for instance, the novel object had been more perceptually distinguishable from the initial objects (e.g., a ball rather than a block – or possibly even a less neutral novel color than white) the children might have been more willing to treat it as a member of a new causal kind rather than as a member of a known kind. Similarly, the children might

have been more willing to revise their initial causal law had the apparently anomalous evidence been more difficult to contravene. Positing unobserved entities behind a curtain is a relatively simple auxiliary hypothesis; the initial causal law would presumably not have been so robust had the auxiliary hypothesis been more improbable (e.g., had the children actually seen the white block contact the yellow block and thus needed to posit the presence of a genuinely unobservable entity).

Conversely however, had children's prior belief in the initial causal law been stronger (e.g., had children been deeply convinced that there were only three kinds of activating blocks in the world), the causal law might have been even more robust to potential counter-evidence. Indeed, as philosophers of science have long noted (Duhem, 1954; Lakatos, 1970; Popper, 1959; Quine, 1953; Strevens, 2001) even the most prima facie improbable auxiliary hypotheses can gain plausibility when they rescue well-accepted abstract causal laws from otherwise anomalous data. Critically, if the causal law is correct, such hypotheses can support important discoveries in their own right. Thus in the 19th century, scientists posited the, previously unsuspected, existence of an eighth planet (Neptune) to explain perturbations in the orbit of Uranus inconsistent with Newtonian mechanics; today, scientists posit the existence of dark matter in order to rescue the laws of gravity from otherwise incompatible data about the orbit of galaxies.

The inferential gap between suspecting the presence of an unseen block and hypothesizing the existence of dark matter is vast. Nonetheless, both insights result from a willingness to posit unobserved entities in order to reconcile otherwise surprising evidence with an abstract causal principle. Children's ability to engage in this reasoning suggests that imaginative leaps of scientific insight may have humble origins in our early ability to infer abstract principles that – by limiting the hypotheses we consider – lead us to consider more than meets the eye.

Acknowledgements

This research was supported by a James H. Ferry, Jr. Fund for Innovation in Research Education to L.S. and a McDonnell Foundation Collaborative Initiative Causal Learning grant to L.S. and J.T. Thanks to Elizabeth Baraff Bonawitz and Rebecca Saxe for helpful comments and suggestions. Thanks also to the Discovery Center at the Museum of Science, Boston, and to participating parents and children.

References

Ahn, W. K., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54*(3), 299–352.

Baldwin, D. A., Markman, E. M., & Melartin, R. L. (1993). Infants' ability to draw inferences about nonobvious object properties: Evidence from exploratory play. *Child Development, 64*(3), 711–728.

Brewer, W. F., & Chinn, C. A. (1991). Entrenched beliefs, inconsistent information, and knowledge change. In L. Birnbaum (Ed.), *Proceedings of the 1991 International Conference on the Learning Sciences* (pp. 67–73). Charlottesville, VA: Association for the Advancement of Computing in Education.

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

Carey, S. (in press). *The Origin of Concepts*. New York: Oxford Unversity Press.

Champagne, A. B., Klopfer, L. E., & Anderson, J. H. (1980). Factors influencing the learning of classical mechanics. *American Journal of Physics, 48*(12), 1074–1079.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098–1120.

Chi, M. T. H. (1992). Conceptual change within and across ontological categories: Examples from learning and discovery in science. In R. Giere (Ed.), *Cognitive models of science: Minnesota studies in the philosophy of science*. Minneapolis: University of Minnesota Press.

Chinn, C. A., & Brewer, W. F. (1992). Psychological responses to anomalous data. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 165–170). Hillsdale, NJ: Erlbaum.

Chinn, C. A., & Brewer, W. F. (1993). Factors that influence how people respond to anomalous data. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 318–323). Hillsdale, NJ: Erlbaum.

diSessa, A. A. (1982). Unlearning Aristotelian physics: A study of knowledge-based learning. *Cognitive Science, 6*(1), 37–75.

Duhem, P. (1954). The aim and structure of physical theory. Translated by P.P. Wiener. Princeton, NJ: Princeton University Press.

Dunbar, K., & Klahr, D. (1989). Developmental differences in scientific discovery processes. In D. Klahr & K. Kotovsky (Eds.), *The 21st Carnegie-Mellon symposium on cognition: Complex information processing: The Impact of Herbert A. Simon*. Hillsdale, NJ: Lawrence Erlbaum.

Feigenson, L., & Carey, S. (2003). Tracking individuals via object files: Evidence from infants' manual search. *Developmental Science, 6*(5), 568–584.

Gelman, S. A. (2006). Early conceptual development. In K. McCartney & D. Phillips (Eds.), *Blackwell handbook of early childhood development* (pp. 149–166). Malden, MA: Blackwell.

Gelman, S. A., & Gottfried, G. M. (1996). Children's causal explanations of animate and inanimate motion. *Child Development, 67*(5), 1970–1987.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23*(3), 183–209.

Gelman, S. A., & Markman, E. M. (1987). Young children's inductions from natural kinds: The role of categories and appearances. *Child Development, 58*(6), 1532–1541.

Glymour, C. (2001). *The mind's arrows: Bayes nets and causal graphical models in psychology*. Cambridge, MA: MIT Press.

Glymour, C., & Cooper, G. F. (1999). *Computation, causation, and discovery*. Cambridge, MA: MIT/AAAI Press.

Goodman, N. (1983). The new riddle of induction. In *Fact, fiction, and forecast* (4th ed., pp. 59–83). Cambridge, MA: 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.

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

Gopnik, A., & Schulz, L. E. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Sciences, 8*(8), 371–377.

Gopnik, A., & Schulz, L. E. (Eds.). (2007). *Causal learning: Psychology, philosophy and computation*. New York, NY: Oxford University Press.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development, 71*(5), 1205–1222.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patters of variation and covariation. *Developmental Psychology, 37*(5), 620–629.

Hume, D. (1748). *An enquiry concerning human understanding*. Oxford: Oxford University Press [A.P.S. Milgram, Trans.].

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books.

Jefferys, W., & Berger, J. (1992). Ockham's razor and Bayesian analysis. *American Scientist, 80*, 64–72.

Jordan, M. (Ed.). (1998). *Learning in graphical models*. Cambridge, MA: MIT Press.

Karmiloff-Smith, A., & Inhelder, B. (1975). If you want to get ahead, get a theory. *Cognition, 3*(3), 195–212.

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

Keil, F. C. (1995). The growth of causal understandings of natural kinds. In D. P. D. Sperber (Ed.), *Causal cognition: A multidisciplinary debate.*

Symposia of the Fyssen Foundation; Fyssen Symposium, 6th Jan 1993, Pavillon Henri IV, St-Germain-en-Laye, France. New York, NY: Clarendon Press/Oxford University Press.

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences, 7*(8), 368–373.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2007). Learning causal schemata. *Proceedings of the 29th Annual Conference of the Cognitive Science Society*.

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

Klahr, D., Dunbar, K., & Fay, A. L. (1990). Designing good experiments to test 'bad' hypotheses. In J. Shrager & P. Langley (Eds.), *Computational models of discovery and theory formation* (pp. 355–402). San Mateo, CA: Morgan-Kaufman.

Kohn, A. S. (1993). Preschoolers' reasoning about density: Will it float? *Child Development, 64*(6), 1637–1650.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Cambridge, MA: MIT Press.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review, 96*(4), 674–689.

Kuhn, D., Amsel, E., & O'Laughlin, M. (1988). *The development of scientific thinking skills*. Orlando, FL: Academic Press.

Kushnir, T., & Gopnik, A. (2005). Young children infer causal strength from probabilities and interventions. *Psychological Science, 16*(9), 678–683.

Kushnir, T., & Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental Psychology, 44*, 186–196.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). New York: Cambridge University Press.

Lehrer, R., & Schauble, L. (1998). Reasoning about structure and function: Children's conception of gears. *Journal of Research in Science Teaching, 35*(1), 3–25.

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; Based on a conference entitled "Cultural Knowledge and Domain Specificity," held in Ann Arbor, MI, Oct 13–16, 1990* (pp. 119–148). New York, NY: Cambridge University Press.

Mansinghka, V. K., Kemp, C., Tenenbaum, J. B., & Griffiths, T. L. (2006). Structured priors for structure learning. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* (UAI 2006).

Masnick, A. M., & Klahr, D. (2003). Error matters: An initial exploration of elementary school children's understanding of experimental error. *Journal of Cognition and Development, 4*(1), 67–98.

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month–old children. *Developmental Psychogy, 31*, 838–850.

Nazzi, T., & Gopnik, A. (2003). Sorting and acting with objects in early childhood: An exploration of the use of causal cues. *Cognitive Development, 18*, 219–237.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Mateo, CA: Morgan Kaufman.

Pearl, J. (2000). *Causality*. New York: Oxford University Press.

Piaget, J. (1930). *The child's conception of physical causality*. London: Kegan Paul.

Piaget, J. (1980). The psychogenesis of knowledge and its epistemological significance. In M. Piatelli-Palmarini (Ed.), *Language and Learning* (pp. 23–34). Cambridge, MA: Harvard University Press.

Popper, K. R. (1959). *The logic of scientific discovery*. New York: Basic Books.

Quine, W. V. O. (1953). Two dogmas of empiricism. In *From a logical point of view*. Cambridge, MA: Harvard University Press.

Rozenblit, L. R., & Keil, F. C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*, 521–562.

Samarapungavan, A. (1992). Children's judgments in theory choice tasks: Scientific rationality in childhood. *Cognition, 45*(1), 1–32.

Saxe, R., Tenenbaum, J., & Carey, S. (2005). Secret agents: 10 and 12-month-olds infer an unseen cause of the motion of an inanimate object. *Psychological Science, 16*(12), 995–1001.

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

Schulz, L. E., Bonawitz, E. B., & Griffiths, T. (2007). Can being scared make your tummyache? Naive theories, ambiguous evidence and preschoolers' causal inferences. *Developmental Psychology, 43*(5), 1124–1139.

Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology, 40*(2), 162–176.

Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science, 10*(3), 322–332.

Schulz, L. E., Hooppell, K., & Jenkins, A. C. (2008). Judicious imitation: Young children imitate deterministic actions exactly, stochastic actions more variably. *Child Development, 79*(2), 395–410.

Schulz, L. E., Kushnir, T., & Gopnik, A. (2007). Learning from doing: Interventions and causal inference. In A. Gopnik & L. E. Schulz (Eds.), *Causal Learning; Psychology, Philosophy and Computation*. New York: Oxford University Press.

Schulz, L. E., & Sommerville, J. (2006). God does not play dice: Causal determinism and preschoolers' causal inferences. *Child Development, 77*(2), 427–442.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development, 47*(1), 1–51.

Sobel, D. M., & Kirkham, N. Z. (2007). Interactions between causal and statistical learning. In A. Gopnik & L. E. Schulz (Eds.), *Causal learning: Psychology, philosophy and computation* (pp. 139–153). 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*(3), 303–333.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of Knowledge. *Psychological Review, 99*(4), 605–632.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search (Springer Lecture Notes in Statistics)*. New York: Springer-Verlag.

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

Strevens, M. (2001). The Bayesian treatment of auxiliary hypotheses. *The British Journal for the Philosophy of Science, 52*(3), 515–537.

Tenenbaum, J. B., & Niyogi, S. (2003). Learning Causal Laws. *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10*(7), 309–318.

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

Wisniewski, E. J., & Medin, D. L. (1991). Harpoons and long sticks: The interaction of theory and similarity in rule induction. In D. Fisher & M. Pazzani (Eds.), *Computational Approaches to Concept Formation* (pp. 237–278). San Mateo, CA: Morgan Kaufman.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.