Don’t throw the associative baby out with the Bayesian bathwater: Children’s retrospective reasoning about multiple causes suggests multiple systems for causal inference

Deon T. Benton1, David Kamper2, Rebecca M. Beaton1, David M. Sobel2

1Vanderbilt University  
2Brown University

Address correspondence to Deon T. Benton, Department of Psychology and Human Development, Vanderbilt University, Peabody College, 230 Appleton Place #552, Nashville, TN 37235

Data availability statement: The code and model scripts are available upon reasonable request.

Conflicts of interests: none

Don’t throw the associative baby out with the Bayesian bathwater: Children’s retrospective reasoning about multiple causes suggests multiple systems for causal inference

Submitted to *Developmental Science* on DATE

Abstract

Causal reasoning is a fundamental cognitive ability that enables children to learn about the complex interactions in the world. The mechanisms that underpin children’s causal reasoning, however, are not well understood. An open question is the extent to which children retrospectively reevaluate causal efficacy given ambiguous information, based on observing novel patterns of data. Here, we report two experiments that test children’s capacity to engage in such inferences. We also fit those data to different computational frameworks – one more associative and one more Bayesian – to consider the strengths and weaknesses of each approach, and the possibility that these approaches together better explain children’s causal reasoning than either approach individually.

Keywords: causal reasoning; cognitive mechanisms; computational models; associative learning; Bayesian inference

Few capacities are more important than the ability to reason and make inferences about cause-and-effect relations. Causal reasoning enables human learners to make predictions and inferences (e.g., Bullock, et al., 1982; Gopnik & Sobel, 2000), to intervene on those relations to generate new effects (e.g., Butler et al., 2020; Gopnik et al., 2001; Schulz et al., 2007), and to reason about counterfactual claims – both about what might have been and how events could have turned out differently (e.g., Harris et al, 1996; Sobel, 2004; Walker & Nyhout, 2020). These and many other studies (e.g., Bonawitz & Lombrozo, 2012; Gopnik et al., 2001; Legare et al., 2010; Meltzoff et al., 2012; Walker & Gopnik, 2014) posit that young children have sophisticated causal reasoning capacities.

A fundamental question that underlies this research is *how*—that is, by what cognitive mechanism or mechanisms—children make such inferences. One answer to this question is that children’s causal inferences are best described by algorithms that underpinned by Bayesian inference. The crux of this idea is that learners use a simple form of Bayes’ rule to reason about causal events and to choose the causal hypothesis—among potentially infinitely many causal hypotheses—that is most consistent with the observed data (e.g., Bonawitz et al., 2014; Gopnik & Wellman, 2012; Griffiths & Tenenbaum, 2005, 2007; Xu, 2019). Crucially, causal reasoning starts with statistical learning capacities that are present in infancy (e.g., Gomez, 2002; Kirkham et al., 2002; Marcus et al., 1999; Saffran et al., 1996) but that develop into a system that infers abstract patterns of coherent causal structure from probabilistic data.

An alternative perspective is that associative learning alone is sufficient to describe children’s causal inferences. On this view, children build up a representation of causal structure from connecting and processing multiple associative relations and statistical regularities. Connectionist models—which learn largely via associative learning—have provided a proof of concept that causal learning can emerge from such associative processes (e.g., Benton et al., 2021; McClelland & Thompson, 2007). Additionally, comparative investigation between non-human animals and adults (e.g., Heyes, 2012) and studies of instrumental action and conditioning on human infants (e.g., Greco et al., 1990; Rovee-Collier, 1999) provide behavioral support for associative learning as a candidate mechanism for how children reason in the world.

One way to illustrate the tension between these hypotheses in development is through investigations of *backwards blocking* (Shanks, 1985). This is a form of reasoning that involves reevaluating the causal status of an ambiguous event based on learning more about the status of other unambiguous events (see also De Houwer et al, 2002; Larkin et al, 1998; Kruschke & Blair, 2000; Lovibond, 2003; Van Hamme & Wasserman, 1994, for other work on adults). For example, Sobel et al. (2004) introduced 3- and 4-year-olds to a machine called a “blicket detector” that lit up and played music when certain objects called “blickets” were placed on it (Gopnik & Sobel, 2000). Children were then shown that two novel objects, A and B, activated the machine when they were placed on it at the same time. Children were then shown that object A alone either did or did not activate the machine. On both types of trials, children were then asked whether each object was a blicket. Children judged that A was a blicket only when it activated the machine. Their judgments of object B also differed across these conditions. Children judged object B more likely to be a blicket when object A failed to activate the machine than when it did so. Using modified procedures, toddlers and even infants as young as 8 months showed a similar pattern of responses (Sobel & Kirkham, 2006).

These findings have since been interpreted as support for Bayesian inference rather than associative learning. This is because some associative models such as the Rescorla & Wagner (1972) predict that the strength between object B and the machine’s activation is equivalent between the backwards blocking (where A is effective) and indirect screening-off (where A is not effective) trials. However, there are two facets of these data that warrant further consideration. First, what is not clear in these data is whether and how children reevaluate the causal status of object B. For instance, do children increase their belief that B is a cause when A fails to activate the machine but decrease their belief that B is a cause when A activates the machine, or are both occurring? when object A fails to activate the machine, do they increase their belief that B is efficacious, when object A activates the machine, do they decrease their belief about B, or are both occurring (Beckers et al., 2005; McCormack et al., 2009)?

McCormack et al. (2009) showed children a similar backwards blocking sequence (AB+, A+) to Sobel et al. (2004): Two objects (A and B) activated the machine together, and then object A activated it alone. They compared children’s causal status judgments for object B with a sequence in which a third object, unrelated to the compound set, activated the machine (i.e., AB+, C+). The 4-year-olds did not differ in their judgments (although 5-year-olds did). This control measure—which we adopt here—is a superior measure of assessing whether children reevaluate their causal judgments. Although these studies used different control trials, the critical point of agreement between these investigations is that at some point in development, children have the capacity to backwards blocking reasoning (which is a form of ‘retrospective reevaluation'). A fundamental question remains, however: *How*—that is, by what cognitive mechanism—do children engage in this type of reasoning?

With this possibility in mind, the present study reconsiders children’s backwards blocking capacities in the context of an observation of the Griffiths et al. (2011) data. In their third experiment, 4-year-olds were shown two pairs of compound stimuli (A and B, and then A and C) were efficacious. The children they investigated categorized A as efficacious more often than B or C, and less so than ceiling, but not differently from individual objects presented as a single compound (X and Y that together activated the machine). In other words, children did not judge the likelihood that object A was efficacious as different from the efficacy of objects X and Y. At question is whether having to reason about more than two objects produced information processing demands that caused children to rely more on associations in their inferences (as all objects were associated with the machine’s activation).

The presence of such information processing demands affecting children’s inferences might be surprising. Cohen et al. (2002) proposed numerous ways information processing demands, such as increased memory and attentional load, could interfere with children’s cognitive processing. They key idea is that information processing demands could limit more rational causal inferences in young children, which in turn can cause children to “drop back” to a more associative form of processing (see Cohen & Amsel, 1998; Cohen & Oakes, 1993). Similarly, although Sobel and Kirkham (2006) found that 8-month-olds engaged in backwards blocking inferences like preschoolers, 5-month-olds’ inferences on the same measure looked more associative in nature (Sobel & Kirkham, 2007). In addition, when infants make inferences about the reliability of others’ information, their judgments appear more associative in nature (Tummeltshammer et al., 2014). As children enter the preschool years, those judgments become more based in rational inferences, although occasionally they will default to certain kinds of associative inferences (Hermes et al., 2018; Luchkina et al., 2020). On this point, Sobel et al. (2017; see also Erb & Sobel, 2014) showed that between 4 and 7 years of age, children develop the capacity to form larger hypothesis spaces of the potential causes they might need to hold to engage in more rational inferences.

Here we asked whether children could engage in backwards blocking reasoning for three and four objects as opposed to two. Our design was like that used by McCormack et al. (2009). Children observed three objects (A, B, and C) together have causal efficacy, and then A by itself either have or fail to have that same efficacy. They were asked whether each object was efficacious. These trials were compared with control trials in which they again observed three objects (A’, B’, and C’) have efficacy together, and then a fourth object (D) either have or fail to have that efficacy. A retrospective inference involves the judgments of objects B and C being different across these two types of trials. When A is efficacious, judgments of the efficacy of B and C should be lower than the judgments of B’ and C’ when children see that a fourth, unrelated object is efficacious. When A is not efficacious, judgments of the efficacy of B and C should be higher than B’ and C’ when children see that a fourth, unrelated object is not efficacious. Moreover, judgments of B and C should differ between these two trials; B and C are more likely to be judged as efficacious when A is not efficacious than when A is. Because McCormack et al. (2009) found that 5 and 6-year-olds made such retrospective inferences, we considered the same age range here. After presenting these behavioral data, we present a pair of computational models that try to explain these results.

**Method**

**Participants.** Participants were 32 5-year-olds (16 boys and 16 girls; *M* = 64.81 months, range = 60-71 months, SD = 3.48) and 31 6-year-olds (17 boys and 15 girls; *M* = 77.81 months, range = 72-83 months, SD = 3.78). Sample size was determined based on previous studies on backwards blocking reasoning in human children (e.g., Gopnik & Sobel, 2000; Griffiths et al., 2011; Sobel et al., 2004). Two children were excluded from analysis for failing to participate (*N* = 1) or missing video (which made coding their responses impossible) (*N* = 1). We did not collect demographic information about the sample, but the demographic information about sample of children collected by the laboratory during this time was as follows (with % of the population of BLINDED as measured by 2020 Census in Parentheses): 82% White/Caucasian (compared with 83%), 3% Black/African American (9%), 4% Asian/Asian American (4%), 0.5% Native American (1%), and 11% of Mixed Descent (3%). Sixteen percent of the sample identified as Hispanic/Latinx (compared with 17% of the population). Similarly, the overall household income level of families tested in the lab during this time was as follows: Less than 30K: 7%, 30-50K: 7%, 50-70K: 14%, 70-90K: 9%, 90-120K: 25%, Over 120K: 38K. The median income for the population of BLINDED as measured by the 2020 Census was ~$74K.

**Materials.** The “device” used in the current study was a computer-animated version of the blicket detector (e.g., Gopnik & Sobel, 2000). The device was a white rectangle with a black border that measured 5.99 cm × 23.47 cm. If the device was “on”, the white region of the rectangle turned blue. No music was played when the machine activated. If the device was “off”, the white region remained white. A maximum of 4 differently colored circles were used, and each circle measured 2.67 cm × 2.67 cm (see Figure 1 below). The machine was designed such that it activated immediately when the bottommost edge of a circle—predetermined to be a blicket—contacted it. At the start of any given trial, three (for the backwards blocking or indirect screening-off trials) or four (for the backwards blocking or indirect screening-off control trials) equally spaced circles appeared above the machine. Finally, the videos contained a built-in script, which experimenters read. All video events were created in Microsoft PowerPoint.

**Procedure.** Participants were tested in a quiet room in local children’s science museum. At the beginning of the experiment, all participants were shown a pretraining video. The video consisted of a rectangular base (i.e., the previously mentioned “blicket detector”) and two shapes (i.e., a gray triangle and a gray pentagon). Crucially, these shapes were unrelated to the circles used during the experimental portion of the experiment. The pretraining phase began with the triangle (object A) and pentagon (object B) above the machine and next to one another. Object A then descended until it contacted and immediately activated the machine (i.e., the white region changed from white to blue). Object A then returned to its starting position above the machine. Object B then descended until it contacted and failed to activate the machine. Object B then returned to its starting position. Finally, both objects descended until they contacted and activated the machine. Participants were then asked whether each object was a blicket. This event ensured that participants understood the task and recognized that individual objects could activate the machine and that the it activated if at least one effective object was placed on it.

Following the pretraining phase, participants were given four trials. Half the participants received two backwards blocking trials and two backwards blocking control trials. The other half received two indirect screening off trials and two indirect screening off control trials. The order of these trials within each condition was counterbalanced using a Latin square design. Different colored objects were used across all trials to prevent carryover effects. A schematic of this procedure is shown in Table 1.

Diagram

Description automatically generated

Figure 1. Schematic of a Backwards Blocking experimental trial. The upper-right portion of the figure shows the backwards blocking event as it unfolded across time. The lower-left portion of the figure shows the three objects and the text, “Is this one a blicket?” above each object across time.

**Backwards Blocking Main and Control Trials.** The two backwards blocking experimental trials began with three differently colored objects, which were located above the machine. The text, “Look, I have these three toys. Let’s find the blickets. Watch what happens” appeared above the objects. All three objects (i.e., objects A, B, and C) then descended until they contacted and activated the machine. At this point, the text, “Look, these also make the machine go!” appeared above the objects. The objects then returned to their starting positions.

The left- or right-most (counterbalanced) object (which we will refer to here as object A) then descended until it contacted and immediately activated the machine. The text, “Look, this one makes the machine go!” then appeared above the objects. This object then returned to its starting position. Children were then asked whether each object was a blicket. Specifically, the text, “Is this one a blicket?” with a downward-facing arrow then appeared above each object, and participants were asked to indicate whether each object was a blicket. Children received two of these trials, which were identical except for the color of the objects.

The two backwards blocking control trials began with four differently colored objects (i.e., objects A, B, C, and D), which were located above the machine. Objects A, B, and C then descended until they contacted and activated the machine; object D remained in place while objects A-C descended onto the machine. Object D then descended by itself until it contacted and activated the machine. The left-right position of object D was counterbalanced. Children were then asked whether each object was a blicket. Children once again received two trials, which were identical except for the color of the objects.

**Indirect Screening-Off Main and Control Trials.** The procedures for the indirect screening-off experimental and control conditions were identical to the backwards blocking trials except that object A (experimental trials) and D (control trials) failed to activate the machine.

**Results**

Figure 2 shows the number of times children responded “yes” to the question “Is this a blicket” for each object. Using this dependent measure, the data were entered into a five-way linear mixed-effects model with Age as a continuous fixed effect, Condition (Backwards blocking vs. Indirect screening-off) as the between-participants fixed effect, Trial Type (Experimental vs. Control), Objects (A vs. B vs. C vs. D), and Trial Number (Trial 1 vs. Trial 2) as the within-participants fixed effects, and participant as the random effect. This analysis yielded several experimental-effects and two-way interactions, which were qualified by 3 three-way interactions. These included a three-way interaction among Age, Condition, and Object, χ*2*(3) = 7.90, *p* = .05, a three-way interaction among Condition, Trial Number and Object, χ*2*(3) = 13.31, *p* = .006, and a three-way interaction among Condition, Trial Type, and Object, χ*2*(2) = 64.85, *p <* .001.

To explore the interaction among Age, Condition, and Object, we constructed separate two-way linear mixed-effects models between Age and Object for each condition. Age was included as a continuous fixed effect, Condition as a between-participants fixed effect, Object as a within-participants fixed effect, and participants as a random effect. Both linear models only yielded experimental effects of Objects, both χ2’s > 31.88, both *p*-values < .001, which indicated that participated treated the objects differently. Specifically, in the backwards blocking condition, participants considered object A (*M* = .89, *SD* = .31) to be more of a blicket than object B (M = .67, SD = .47), *t*(30) = 4.95, *p* < .001, and C (*M* = .71, *SD* = .46), *t*(30) = 3.89, *p* < .001*.* However, participants treated objects A and D (*M* = .85, *SD* = .36) equivalently, *t*(30) = .76, *p* = .45. In contrast, in the indirect screening-off condition, participants were less confident that object A (*M* = .54, *SD* = .50) was a blicket than object B (*M* = .79, *SD* = .41), and C (*M* = .84, *SD* = .37), both *t*’s > -5.03, both *p*-values < .001. However, participants were more confident that object A was a blicket than object D (*M* = .36, *SD* = .48), *t*(30) = 2.36, *p* = .02. This reflected the fact that the scores for object A were collapsed over Trial Type (in which A was seen by itself during the experimental trials but in combination with other objects during the control trials). Finally, participants were less confident that object D was a blicket than objects B and C, both *t*’s > 6.10, both *p*-values < .001.

To explore the second interaction among Trial Number and Object for each condition, Trial Number and Object were included as within-participants fixed effects and participants were included as a random effect. Although both linear models yielded experimental effects of Object, both χ2’s > 31.86, both *p*-values < .001, only the two-way linear mixed-effects model for the Indirect Screening Off condition yielded an interaction between Trial Number and Object, χ*2*(3) = 9.57. This interaction reflected the fact that participants treated the objects differently between the two trials. During the trial 1 (when participants were asked to provide their first set of responses) participants treated object D (M = .55, SD = .50) as less of a blicket than object A (M = .72, SD = .45), *t*(114.97) = 2.31, *p* = .02, and object C (*M* = .78, *SD* = .42) as more of a blicket than either object B (*M* = .68, *SD* = .47), *t*(127) = -2.46, *p* = .02, or D, *t*(107.25) = 3.23, *p* < .01. In contrast, during trial 2, participants treated all the objects equivalently, all *t*’s < 1.78, all *p*-values > .08.

Chart, bar chart

Description automatically generated

Figure 2. Participants’ mean responses to whether each object was a blicket across the conditions and trial types. Bars show standard error.

To examine the final interaction among Condition, Trial Type, and Object, we constructed a set of one-way linear mixed-effects models for the experimental and control trials within the backwards blocking and indirect screening-off conditions. The Objects factor was treated as the sole within-participants fixed effect in these follow-up analyses. Participants were once again treated as a random effect to control for the within-participant variance from multiple responses. The one-way linear model for the control trials within the backwards blocking condition did not reveal a significant effect of Objects,χ2(3) = 1.34, *p* = .72. This means that participants treated the objects similarly in the control trials of the backwards blocking condition. In contrast, the second one-way linear model for the experimental trials within the backwards blocking condition revealed a significant experimental effect of Objects, χ2(2) = 55.20, *p* < .001. This experimental effect reflected the fact that participants considered object A to be more of a blicket (*M* = .98, *SD* = 0.13) than object B (*M* = .55, *SD* = 0.50), *t*(30) = 6.45, *p<* .001, or object C (*M* = .61, *SD* = 0.49), *t*(30) = 5.62, *p* < .001. Participants treated objects B and C equivalently, *t*(30) = -1.07, *p* = .29.

The third and fourth one-way linear models for the experimental and control trials within the indirect screening-off condition both revealed a significant experimental effect of Objects, both χ*2*-values > 76.81, both *p*-values < .001. Participants considered object A (*M* = 0.26, *SD* = 0.44) in the ISO experimental trials and object D (*M* = 0.36, *SD* = 0.48) in the ISO control trials to be less likely to be blickets than any of the other objects, all *t*-values > -7.45, all *p*-values < .001. Participants treated object B and C equivalently in the experimental trials, *t*(30) = -0.77, *p* = .29, and objects A-C equivalently in the control trials, all *t*-values < -1.07, all *p*-values > .29.

**Evidence of backwards blocking reasoning.** To examine whether participants engaged in backwards blocking reasoning—operationalized as higher combined ratings of objects A-C in the control trials than of objects B and C in the experimental trials—data were entered into a two-way linear mixed-effects model with Trial Type and Object as the within-participants fixed effects and participants as the random effect. This analysis revealed only a main effect of Trial Type, *χ2*(1) = 21.97, *p* < .001. This result indicated that participants did engage in backwards blocking reasoning: they provided higher combined ratings of objects A, B, and C in the backwards blocking control trials (*M* = 0.80, *SD* = 0.40) than the combined ratings of objects B and C in the backwards blocking experimental trials (*M* = 0.58, *SD* = 0.49).

For completeness, we ran the same analysis as above, but this time for the indirect screening-off condition. This analysis also only revealed a main effect of Trial Type, *χ2*(1) = 4.42, *p* = .04. The results mirrored the results for the backwards blocking condition. Participants provided higher combined ratings of objects A, B, and C in the indirect screening-off control trials (*M* = 0.84, *SD* = 0.36) than the combined ratings of objects B and C in the backwards blocking experimental trials (*M* = 0.77, *SD* = 0.42). Similar to the results above for the backwards blocking condition, this result indicated that when the object that is shown in isolation was also shown in combination with other objects participants show stronger retrospective reevaluations.

**Discussion**

This purpose of this study was to determine how children reason about a backwards blocking event that consisted of three rather than two objects. The results indicated that participants did engage in backwards blocking reasoning. Specifically, we found that participants were less confident that objects that were only shown in groups in the backwards blocking experimental trial (i.e., objects B-C) were blickets compared to the objects that were only shown in groups in the backwards blocking control trial (i.e., objects A-C). Although participants did engage in backwards blocking reasoning, an open question concerns whether participants’ causal inferences, overall or in parts, were best explained by an associative-learning mechanism, a Bayesian-inference mechanism, or some combination of both. We addressed this issue next.

**Computational Models**

We fit two different computational models to the behavioral data. The first was a model based on Bayesian inference. This model was described initially by Sobel et al. (2004) and in more detail in Griffiths et al. (2011). The second was a simple connectionist model.

**Bayesian Model.** The Bayesian model starts with a set of hypotheses *H*. Each hypothesis *h* ∈ *H* is assigned a *prior probability*, *p*(*h*), which indicates the initial belief in that a learner has in a particular hypothesis prior to seeing data. After the learner observes data, *d*, the learner computes a posterior probability, *p*(*h* | *d*), given an updated belief about each hypothesis given the data. This is done using Bayes’ rule, shown in Equation 1:

(1)

In this formula, *p*(*d | h*) is the probability of the data *d* given each a particular hypothesis *h* (also known as the *likelihood*).

Forming the initial hypothesis space relies on assuming that there is a set of objects *O* and a set of detectors *D*, such that any object *o* ∈ *O* can potentially cause any detector *d* ∈ *D* to activate. Given that participants are shown that the machine activates when blicket objects are placed on its surface, a hypothesis *h* corresponds to a structure that posits whether individual objects have the causal efficacy to activate the detector (see Griffiths & Tenenbaum, 2005, for more computational details). Griffiths et al. (2011) describe the formal parameterization of this hypothesis space and model that results in the hypothesis space shown in Figure 3.

To instantiate the model, each hypothesis is given a prior probability *p*(*h*), which is a function of the child’s belief about the base rate of blickets **. This prior corresponds to the number of blickets posited by the hypothesis. For example, in the figure, Hypothesis 0 posits 3 blickets, so its *p*(*h*) = ** Hypotheses 1, 2, and 4 posit exactly 2 blickets, so their *p*(*h*) = ****Hypotheses 3, 5, and 6 each posit 1, making their *p*(*h*) = ****Finally, Hypothesis 7 posits no blickets, making its *p*(*h*) = (1-**

**Timeline

Description automatically generated with medium confidence**

Figure 3. The eight different causal hypotheses indicating the possible causal relations for a causal event that involves three objects and one blicket detector. *A*, *B*, and *C* correspond to the three objects that were used on the machine and *E* indicates the activation of the machine.

For the purposes of this demonstration, we will assume that the model itself assumes that objects with causal efficacy will act deterministically on detectors.[[1]](#footnote-1) As a result, the likelihood of each hypothesis is equal to 1 if that hypothesis could produce the data and 0 if not. This allows each model to be updated based on Bayes’ rule given the data. The way the model determines the probability that an object is a blicket is based on the posterior probability of the models in the hypothesis space; that is, the probability that any object *o* is a blicket given the data *d* can be calculated by the equation in (2)

where *p*(*o*→*E* | *h*) is 1 if there is an edge between that object and the detector in h, and 0 otherwise.

Crucially, because the predictions of this (or any) Bayesian model will depend on the prior probability that any given object is a blicket, below we show the predictions of the model when the probability of a blicket is .5, .65, .8, .95, and 1. Figure 4A-E shows the model’s predictions for Experiments 1 and 2 for the various probabilities. Our rationale for plotting the model’s predictions for various prior probabilities was that it was unclear what participants’ baseline assumptions would be about the prior probability of blickets in the absence of explicit manipulations to those probabilities. Thus, by deriving the model’s predictions for various prior probabilities, it was possible to compare (qualitatively and quantitatively) the model’s predictions for the different probabilities to children’s actual treatment of the objects.

With the exception of Figure 4E in which the baseline probability that an object is a blicket is 100% (and thus children should treat all objects as blickets except those that are shown explicitly not to activate the machine), the model makes two notable qualitative predictions (see below for the precise quantitative predictions). First, when participants observe backwards blocking sequences, they should be most confident that objects A (during the experimental trials) and D (during the control trials) are blickets for all base rates. In contrast, participants in the indirect screening-off condition should be most confident that objects A (during the experimental trials) and D (during the control trials) are not blickets. Second, within the backwards blocking condition, the model predicts that participants should engage in backwards blocking reasoning: Participants should be more confident that objects in the control trials are blickets than objects in the experimental trials, but with diminishing differences as the base rate increases.

**Connectionist model**. We also built a two-layer connectionist model. The model architecture is shown in Figure 4. The model used to simulate the behavioral experiment consisted of an input layer and an output layer—there were no hidden layers in these models. The rationale for building a two-layer model was to explore whether a simple learning model (the Delta Rule, Kruschke, 1992; Widrow & Hoff, 1960) could be used to explain these data. The Delta rule is formally equivalent to the traditional Rescorla-Wagner model (Danks, 2003; Gluck & Bower, 1988). This is an instantiation of an associative learning model that might better explain retrospective inferences as the number of potential causes increases. If such a model can capture the present behavioral data, then the conclusion that the Rescorla-Wagner model is insufficient to explain children’s causal reasoning (e.g., Sobel et al., 2004) may be premature (we return to this issue in the General Discussion).

The input layer for the model consisted of four units, and the output layer consisted of a single unit. Each input unit corresponded to each of the four possible objects used in the experiment. Whenever an object was present, the activation value of its corresponding input unit was set to a value of “1”; the activation of these units was set to a value of “0” if the corresponding objects were not present. The input units could not take on any other values beside 0 or 1. If a predetermined blicket was presented at the input layer, then the model was trained to turn on the single output unit (i.e., to produce an activation of 1). This process corresponded to an object activating the blicket machine. All simulations used a learning rate of .05 but no momentum. Model weights were initialized to small random values (distribution range = ± 0.1), and the output units used sum-squared activation functions (which enabled the weights to be modified with training). The activation of the single output unit was interpreted as the model’s confidence (or prediction) that a given object was a blicket and could range between 0 and 1 due to the sigmoid activation function (unlike the input units). Thus, if object A was presented to the model (i.e., its input unit was set to 1) and the model produced an output activation of 0.50, this indicated that the model was uncertain about A’s causal status.

Diagram

Description automatically generated  
Figure 4. The connectionist model used to simulate Experiment 1.

The procedure for training the models was the same as that for children. For example, models were assigned randomly to the indirect screening-off condition or to the backwards blocking condition. To match the behavioral experiment, models experienced two of each kind of event within a given condition. For example, during the two experimental trials for models in the backwards blocking condition, the first three input units were turned on (i.e., the activation of each input node was set to a value of 1, whereas the activation of the fourth node was set to 0), and the model’s task was to learn to activate the single output unit (i.e., to set the activation of the single output unit to 1). The way this worked in practice was the following: If a blicket was presented to the model, the input node that stood in for that object was set to 1, and the target for the single output unit was set to 0. Real-valued activity then propagated from the input layer directly to the output layer. Once at the output layer, the activation of the output node—which was a nonlinear (sigmoidal) function of the input to it—was compared to the target. If there was a difference between the observed and target output, the input-to-output weights were adjusted via the Delta rule to reduce this difference. Similar to children, we trained 16 models (i.e., ‘participants’) per condition, and like the children each model received two trials. Each new participant began with a fresh set of small random weights (sampled uniformly between ±0.1). Data were aggregated over the responses of each model, as was the case for the children. Turning on the first three input units simulated placing objects A, B, and C on the machine, and training the model to turn on the single output unit corresponded to teaching the model that the machine activated when objects A-C were placed on it. This segment of training corresponded to the ABC+ events.

During the subsequent A+ trials, only the first input unit was turned on, but again the model’s task was to activate the single output unit. The backwards blocking control trials were identical to the backwards blocking experimental trials except that the fourth input unit (corresponding to object D) rather than first input unit was turned on following the ABC+ phase. The indirect screening off experimental and control trials were identical to the backwards blocking experimental and control trials except that the model was trained to turn off the single output unit (i.e., to produce an output activation of 0) during the A- and D- phases of the indirect screening-off experimental and control trials. The ABC+ and A- phases—which were shown twice to be consistent with the behavioral study—lasted anywhere between 200 and 1,000 epochs. This meant that one complete simulation lasted anywhere between 800 (i.e., 200 × 4) and 4,000 (i.e., 1,000 × 4) epochs. Networks were trained for different numbers of epochs to ensure that the model-fit results were not idiosyncratic to the precise number of training epochs. The model’s predictions for the different numbers of training epochs is shown below in Figure 5A-D.

|  |  |  |
| --- | --- | --- |
| Chart, bar chart  Description automatically generated  A | Chart, bar chart  Description automatically generated  B | Chart, bar chart  Description automatically generated  C |
| Chart, bar chart  Description automatically generated  D | Chart, bar chart  Description automatically generated  E |  |

Figure 5A-E. The connectionist model’s predictions for how participants should treat the objects between the experimental and control trials of the backwards blocking and indirect screening-off conditions in Experiment 1 after 800 (A), 1,600 (B), 2,000 (C), (3) 3,000, and (4) 4,000 epochs of training.

As can be seen, the connectionist model predicts that participants should treat the objects that were only shown in groups equivalently across the different trials, conditions, and training epochs. However, the model predicts that participants should be more confident that object A in the backwards blocking experimental trials is a blicket than the objects that were only shown in groups. Finally, for the indirect screening-off experimental trials, the model predicts that participants should be less (though not maximally) confident that objects A and D in the experimental and control trials are blickets compared to the objects that were only shown in groups across the same trials.

Clearly, the Bayesian and connectionist models make distinct predictions for how participants should respond to the objects across the various conditions, trials. Where these models especially diverge is in their predictions for the backwards blocking control condition: The connectionist model predicts that participants should treat objects A-D equivalently during regardless of the number of training epochs; the Bayesian model predicts that participants should treat object D differently than the other objects (which are treated equivalently), with diminishing differences as the base rate increases (until, at a base rate of 1, the model predicts that everything is a blicket). The models also differ in terms of whether they predict backwards blocking reasoning; the Bayesian model, but not the associative model, predicts backwards blocking reasoning. It therefore should be possible to determine which model participants relied on by fitting these models to their data.

**Qualitative and Quantitative Model fits**

To assess the quantitative fit of the predictions of the connectionist and Bayesian models to the data, we computed the root mean square (RMSE) and mean absolute error (MAE) between each model’s predictions (for the connectionist model these were the average activation of the single output unit in response to each object; for the Bayesian model these were point estimates) and participants’ mean responses to objects A-C during the backwards blocking and indirect screening-off experimental trials and objects A-D during the backwards blocking and indirect screening-off control trials. One or both metrics have been used in previous simulation studies to assess a model’s quantitative fit to behavioral data (e.g., Bhat et al., 2022; Buss & Spencer, 2014; Spencer et al., 2022; Stojnic et al., 2023). Lower values on each metric indicate better model fit. Table 2 below shows the fits for the different connectionist and Bayesian model instantiations.

|  |  |  |
| --- | --- | --- |
| Computational Models | root mean square (RMSE) | mean absolute error (MAE) |
| **Connectionist Model (800 epochs)** | **.10** | **.07** |
| Connectionist Model (1600 epochs) | .13 | .09 |
| Connectionist Model (2000 epochs) | .15 | .11 |
| Connectionist Model (3000 epochs) | .17 | .14 |
| Connectionist Model (4000 epochs) | .18 | .15 |
| *Average Connectionist Model Fit* | .15 | .11 |
| Bayesian model (.5) | .21 | .19 |
| Bayesian model (.65) | .16 | .13 |
| Bayesian model (.80) | .15 | .10 |
| Bayesian model (.95) | .22 | .19 |
| Bayesian model (1) | .25 | .22 |
| *Average Bayesian Model Fit* | .17 | .17 |

Table 1. Model fit indices for the various models and instantiations for the data overall. The rows shaded light gray correspond to the best fitting individual connectionist and Bayesian models. Bold denotes the best-fitting individual model.

It should be clear from the table above that the connectionist model provided a better quantitative fit to the behavioral data than any of the Bayesian models. This is most evident when one considers the model’s predictions for participants’ judgements during the backwards blocking control trials and the model’s predictions for these trials. The connectionist model predicted that participants should treat the four objects equivalently, which the behavioral data supported. In contrast, the Bayesian model predicted that participants should treat object A differently than the other objects which the behavioral data did not support.

Although the connectionist model provided a better overall fit to the data than did the Bayesian model, it is possible that these models provided better accounts for different aspects of the data. For example, it is possible that one of the two models would provide a better fit to the backwards blocking data, whereas the other of the two models would provide a better quantitative fit to the indirect screening-off data. Likewise, it is possible that one model would provide a better fit to the experimental data, whereas another model would provide a better fit of the control data. To explore whether this was the case, we first fit both models to the backwards blocking data. This is shown below in Table 3.

|  |  |  |
| --- | --- | --- |
| Computational Models | root mean square (RMSE) | mean absolute error (MAE) |
| Connectionist Model (800 epochs) | .14 | .09 |
| Connectionist Model (1600 epochs) | .18 | .15 |
| Connectionist Model (2000 epochs) | .19 | .16 |
| Connectionist Model (3000 epochs) | .22 | .19 |
| Connectionist Model (4000 epochs) | .23 | .20 |
| *Average Connectionist Model Fit* | .19 | .16 |
| Bayesian model (.5) | .19 | .17 |
| **Bayesian model (.65)** | **.10** | **.10** |
| Bayesian model (.80)++ | .14 | .09 |
| Bayesian model (.95) | .26 | .24 |
| Bayesian model (1) | .31 | .29 |
| *Average Bayesian Model Fit* | .20 | .18 |

Table 3. Model fit indices for the various models and instantiations for the BB experimental and control conditions. The rows shaded light gray correspond to the best fitting individual connectionist and Bayesian models. Bold denotes the best-fitting individual model.

As is shown, the best-fitting individual Bayesian model outperformed the best-fitting individual connectionist model on the backwards blocking behavioral data. However, it should also be clear that on average the connectionist models provided a better fit to these data than did the Bayesian models.

Next, we fit the models to the indirect screening-off data. This is shown below in Table 4.

|  |  |  |
| --- | --- | --- |
| Computational Models | root mean square (RMSE) | mean absolute error (MAE) |
| **Connectionist Model (800 epochs)** | **.04** | **.04** |
| Connectionist Model (1600 epochs) | .06 | .05 |
| Connectionist Model (2000 epochs) | .08 | .06 |
| Connectionist Model (3000 epochs) | .09 | .09 |
| Connectionist Model (4000 epochs) | .11 | .10 |
| Average Connectionist Model Fit | .08 | .07 |
| Bayesian model (.5) | .22 | .21 |
| Bayesian model (.65) | .13 | .11 |
| Bayesian model (.80) | .14 | .13 |
| Bayesian model (.95) | .22 | .19 |
| Bayesian model (1) | .19 | .18 |
| *Average Bayesian Model Fit* | .18 | .16 |

Table 4. Model fit indices for the various models and instantiations for the ISO experimental and control conditions. The rows shaded light gray correspond to the best fitting individual connectionist and Bayesian models. Bold denotes the best-fitting individual model.

Here, the best-fitting individual connectionist model not only provided the best fit to participants’ indirect screening-off data than the best-fitting individual Bayesian model, but on average the connectionist models provided a better fit to these data than the Bayesian models.

We next fit the models to participants’ responses in the experimental trials (Table 5).

|  |  |  |
| --- | --- | --- |
| Computational Models | root mean square (RMSE) | mean absolute error (MAE) |
| Connectionist Model (800 epochs) | .13 | .11 |
| Connectionist Model (1600 epochs) | .18 | .15 |
| Connectionist Model (2000 epochs) | .19 | .16 |
| Connectionist Model (3000 epochs) | .21 | .18 |
| Connectionist Model (4000 epochs) | .22 | .19 |
| *Average Connectionist Model Fit* | .19 | .16 |
| Bayesian model (.5) | .13 | .11 |
| **Bayesian model (.65)** | **.12** | **.08** |
| Bayesian model (.80) | .17 | .14 |
| Bayesian model (.95) | .26 | .23 |
| Bayesian model (1) | .29 | .26 |
| *Average Bayesian Model Fit* | .19 | .16 |

Table 5. Model fit indices for the various models and instantiations for the experimental trials. The rows shaded light gray correspond to the best fitting individual connectionist and Bayesian models. Bold denotes the best-fitting individual model.

Although the best-fitting individual Bayesian model provided a better fit to participants’ experimental data than the best-fitting individual connectionist model, both models provided equivalent fit on average.

Lastly, we fit both models to participants’ control data, which is shown below in Table 6.

|  |  |  |
| --- | --- | --- |
| Computational Models | root mean square (RMSE) | mean absolute error (MAE) |
| **Connectionist Model (800 epochs)** | **.07** | **.05** |
| Connectionist Model (1600 epochs) | .08 | .06 |
| Connectionist Model (2000 epochs) | .10 | .07 |
| Connectionist Model (3000 epochs) | .12 | .09 |
| Connectionist Model (4000 epochs) | .13 | .11 |
| *Average Connectionist Model Fit* | .1 | .08 |
| Bayesian model (.5) | .26 | .25 |
| Bayesian model (.65) | .19 | .17 |
| Bayesian model (.80) | .14 | .08 |
| Bayesian model (.95) | .18 | .16 |
| Bayesian model (1) | .21 | .19 |
| *Average Bayesian Model Fit* | .2 | .17 |

Table 6. Model fit indices for the various models and instantiations for the control trials. The rows shaded light gray correspond to the best fitting individual connectionist and Bayesian models. Bold denotes the best-fitting individual model.

It is clear not only that the best-fitting individual connectionist model provided a better fit to participants’ control data than the best-fitting individual Bayesian model, but on average the connectionist models provided a better fit to the data than the Bayesian models.

There are three key takeaways from these model-fit indices. The first is that the best-fitting *individual* Bayesian model best explained participants’ causal responses in the experimental trials as well as within the backwards blocking condition. The second is that at no point did the Bayesian models, on average, best explain the data—at best the connectionist and Bayesian models provided an equivalent fit to the experimental data. The third is that these data were largelybest explained, individually and broadly, by the connectionist models. Individually, the connectionist models best explained participants’ responses overall, in the indirect screening-off condition, and in the control condition. Broadly, the connectionist models provided the best account of participants’ responses overall and in the backwards blocking, indirect screening-off, and control conditions.

Nonetheless, an important caveat is worth noting. Although children did display a clear tendency to process the present events associatively, this was not always the true as could be seen in their backwards blocking responses. This suggests that participants may simultaneously be relying on Bayesian inference *and* associative reasoning, even when there is a greater tendency to rely on associative processing to reason about multiple potential causes. These data clearly do not support the conclusion that it is only Bayesian inference or associative learning that underpins children’s causal inferences. Instead, these data support the conclusion that children weight these two cognitive mechanisms differently depending on the number of potential causes about which they are asked to reason. Bayesian inference may be given more weight than associative learning when there are a small number of potential causes (e.g., Sobel et al., 2004). However, as the number of causes and the information processing demands of the task increase participants give more weight to associative learning, even if both are operating (albeit to different degrees).

General Discussion

The purpose of this study was to examine how—that is, by what cognitive mechanism or mechanisms—children reasoned about backwards blocking events. Children in the backwards blocking condition were shown that three objects (A-C) activated a machine, and then shown that object A or an unrelated object, object D activated the machine. Children in the indirect screening-off condition saw the same events except that objects A and D did not activate the machine. We found that children did engage in backwards blocking reasoning: They were less confident that objects that were only shown in groups (and never in isolation) in the backwards blocking experimental trial (i.e., objects B-C) were blickets compared to the objects that were only shown in groups in the backwards blocking control trial (i.e., objects A-C). We subsequently fit a Bayesian model and a connectionist model to the data and found that the Bayesian model, but not the connectionist model, best explained children’s backwards blocking performance. Specifically, the best-fitting individual Bayesian model provided a better explanation of participants’ backwards blocking responses than did the best-fitting individual connectionist model.

Yet, we also found that this was the only place that the Bayesian model outperformed the connectionist model. In terms of the individual models, the connectionist models provided the best explanation for participants’ responses overall, in the indirect screening-off condition, and in the control condition. In terms of the average performance of the models, the connectionist model also provided the best account of participants’ responses overall and in the backwards blocking, indirect screening-off, and control conditions. The models were tied in their ability to explain participants’ responses during the experimental trials. Thus, these data seem to implicate the operation of a Bayesian inference and an associative learning mechanism, with a tendency for participants to process the events associatively. This finding is particularly noteworthy because the argument has been made that the associative learning captured by the Rescorla-Wagner model—a learning procedure that the connectionist model effectively implemented—is insufficient to explain how children (e.g., Sobel et al., 2004) and adults (Griffiths et al., 2011) reason causally. This finding extends previous research on this topic by showing that when children are asked to reason about three and four causes (as opposed to the standard two objects used in previous research), children reason normatively and associatively, with a clear tendency for associative processing.

One may question whether the difference between a setting in which participants are asked to reason about two candidate causes and one in which they are asked to reason about three or even four candidate causes is theoretically meaningful. Such skepticism would stem from the fact that the two situations differ by at most two potential causes (i.e., 2 vs. 4 causes). However, if Bayesian inference is the description of how the cognitive mechanism that underpins children’s causal inferences works, then the difference between these two settings is substantial. This is because with two causes participants need only to determine which of *four* candidate causal hypotheses generated the observed data—if each object can either be a blicket or not, then there are 22 possible combinations of blickets and non-blickets. In contrast, with three and four objects participants need to determine which of *eight* (23) or *sixteen* (24) hypotheses generated the data, respectively. This means that participants must consider up to four times as many hypotheses across these two situations. However, if the children tested here lacked the requisite processing resources to reason over what to them is such an expansive hypothesis space, this may explain why they processed the present events largely associatively (though see Bonawitz et al., 2014 and Gopnik et al., 2015 for potential solutions to this “search” problem).

Although children largely processed the present events associatively, a question that we have not yet answered concerns the exact nature of this associative learning. Specifically, how did the connectionist models (and by extension, the children)—through associative learning—arrive at their causal judgements? To understand how the connectionist model’s judgements arose mechanistically (and perhaps by extensive, how children may have arrived at their causal judgements), consider the control trial in the backwards blocking (i.e., ABC+ D+) condition. During the simulation of this trial, when all four objects were first presented to the model, the resulting difference at the output layer between the activation of the single output unit and the predicted activation of that unit was equivalent for all four objects. Thus, because the difference between the observed and predicted activation of the output unit was equivalent for all four objects, the model made equivalent weight adjustments in sign and magnitude to the connections between each object and the output unit. Crucially, these connections instantiated each object’s association with the machine’s activation. Stated plainly, the model’s responses were based on a simple associative “counting” strategy. This strategy, in turn, was based on the number of times that a given object appeared with the blicket effect. As such, because objects A-D were shown with the “machine’s activation” (i.e., the output of the output unit) an equal number of times in the control trials of the backwards blocking condition, the strength of the association between each object and the machine’s activation was equivalent. Given that participants’ responses mostly matched the model’s predictions, this suggests that in those cases in which children’s responses were better explained by the connectionist model than by the Bayesian model children relied on a similar associative process.

It turns out that there is a wealth of data that is consistent with the contention that children rely on simpler modes of thinking when their information-processing capacities are stretched (Doebel & Zelazo, 2015; Frye, Zelazo, & Palfai, 1995; Zelazo, Frye, & Rapus, 1996; Zelazo et al., 2003). One such recent study by Kenderla and Kibbe (2023) showed that when 8- and 10-year-old children’s information-processing abilities were stretched in a virtual memory game—such as when children were asked to find three cards that shared one feature and differed on another feature—they relied less on working memory and more on manual exploration. Given that children were not required actively to maintain information in memory when manually exploring, manual exploration was an ostensibly simpler and less cognitively effortful strategy than one that required an already resource-limited system such as working memory. In a similar vein, Richland et al. (2006) found that 3- and 4-year-old children made more featural and relational errors when asked to reason about multiple relations or when the task included a salient distractor than when asked to reason about a single relation without a distractor. Finally, there is evidence that preschool-age children's performance on theory-of-mind and social-problem-solving tasks was adversely affected when they first completed tasks that taxed their information-processing abilities compared to when such capacities were not taxed (Caporaso & Marcovitch, 2021; Powell & Carey, 2017; Steinbeis, 2018).

Together, this research demonstrates that although children can process information at higher levels, if the task that they are given requires information-processing abilities that extend beyond what they possess, then there will be a tendency for them to process information at lower levels and to rely on less sophisticated strategies and cognitive mechanisms (e.g., Cohen et al., 1998). A testable prediction of this account is that there should be a point at which children go from using a simple associative-based counting mechanisms in contexts like the present one to more rationale processes like Bayesian inference. This issue should be explored more fully in future research.

Before closing, some potential criticisms are worth noting. First, we cannot be sure that there are no contexts in which of the balance of associative processing to Bayesian inference can be flipped. In the present study, children mostly relied on associative processing and minimally on Bayesian inference, but it is possible that in the right situation children would mostly rely on Bayesian inference and minimally on associative processing. For example, if participants assumed that blickets were common in the present context—which is plausible given both how frequently the detector activated in the present study and the fact that the model with a base rate of 1 occasionally outperformed models with lower base rates—then participants should be *less* likely to block redundant causes; in other words, participants should be *more* likely to treat all potential blickets (expect for the ones that are explicitly shown not to be causal) equally. This explains participants’ performance across the board, although it does run into difficulty explaining participants’ backwards-blocking performance (which likely accounts for the point at which participants relied on more rational inferences captured by algorithms that implement Bayesian inference).

A second potential criticism concerns the connectionist model’s performance relative to that of the Bayesian model. Specifically, one may raise the objection that the connectionist model's superior overall performance compared to the Bayesian model was due to overfitting to the data. Although it is true that the connectionist model tended to produce lower values on these two model fit indices, this criticism is inconsistent with the data. This is because if the connectionist model's superior performance over that of the Bayesian model resulted from overfitting, it would be expected to outperform the Bayesian model in a few specific instances rather than across the board as we have observed. This criticism is also weakened by the fact that the model’s learning parameters and its architecture remained constant throughout the simulations. Overfitting, which can occur when the model is modified to fit various aspects of the data in different tasks, is therefore unlikely to have influenced the present results.

A third criticism concerns the artificial nature of the paradigm used here, which was necessitated by the pandemic. This may have introduced a level of noise in the data that is fundamentally different than testing in person with real objects. Future studies should replicate our study. If such a study revealed that participants performed more normatively than associatively in person, this would suggest that children’s normative inferences may not be as robust as originally thought—it is present when tested in person but nearly absent when tested on a computer. Such a finding would be interesting regardless because it would add nuance to the literature on children’s causal inferences.

**Conclusion**

This study constitutes one of the first systematic attempts to examine backwards blocking and indirect screening-off reasoning in human children in the context of multiple three and four candidate causes. A longstanding view has been that the cognitive mechanism by which people reason about causal events is Bayesian inference rather than associative processes. The experiments reported here support a different conclusion: children rely on associative learning *and* Bayesian inference to reason about causal events.

References

Beckers, T., Vandorpe, S., Debeys, I., & De Houwer, J. (2009). Three-year-olds’ retrospective revaluation in the blicket detector task: Backward blocking or recovery from overshadowing?. *Experimental Psychology*, *56*(1), 27-32.

Benton, D. T., Rakison, D. H., & Sobel, D. M. (2021). When correlation equals causation: A behavioral and computational account of second-order correlation learning in children. Journal of Experimental Child Psychology, 202, 105008.

Bhat, A. A., Spencer, J. P., & Samuelson, L. K. (2022). Word-Object Learning via Visual Exploration in Space (WOLVES): A neural process model of cross-situational word learning. Psychological Review, 129(4), 640.

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-Stay, Lose-Sample: A simple sequential algorithm for approximating Bayesian inference. Cognitive psychology, 74, 35-65.

Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: children's use of simplicity and probability to constrain inference. Developmental psychology, 48(4), 1156.

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. The developmental psychology of time, 209-254.

Buss, A. T., & Spencer, J. P. (2014). The emergent executive: A dynamic field theory of the development of executive function. Monographs of the Society for Research in Child Development, 79(2), vii.

Butler, L. P., Gibbs, H. M., & Tavassolie, N. S. (2020). Children’s developing understanding that even reliable sources need to verify their claims. Cognitive Development, 54, 100871.

Caporaso, J. S., & Marcovitch, S. (2021). The effect of taxing situations on preschool children’s responses to peer conflict. *Cognitive Development*, *57*, 100989.

Danks, D. (2003). Equilibria of the Rescorla–Wagner model. Journal of Mathematical Psychology, 47(2), 109-121.

Doebel, S., & Zelazo, P. D. (2015). A meta-analysis of the Dimensional Change Card Sort: Implications for developmental theories and the measurement of executive function in children. *Developmental Review*, *38*, 241-268.

Erb, C. D., & Sobel, D. M. (2014). The development of diagnostic reasoning about uncertain events between ages 4–7. PloS one, 9(3), e92285.

Frye, D., Zelazo, P. D., & Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive development*, *10*(4), 483-527.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: an adaptive network model. Journal of Experimental Psychology: General, 117(3), 227.

Gomez, R. L. (2002). Variability and detection of invariant structure. Psychological Science, 13(5), 431-436.

Gopnik, A., Griffiths, T. L., & Lucas, C. G. (2015). When younger learners can be better (or at least more open-minded) than older ones. Current Directions in Psychological Science, 24(2), 87-92.

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 patterns of variation and covariation. *Developmental psychology*, *37*(5), 620.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, *138*(6), 1085.

Greco, C., Hayne, H., & Rovee-Collier, C. (1990). Roles of function, reminding, and variability in categorization by 3-month-old infants. Journal of Experimental Psychology: Learning, memory, and cognition, 16(4), 617.

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive science*, *35*(8), 1407-1455.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. Cognitive psychology, 51(4), 334-384.

Griffiths, T. L., & Tenenbaum, J. B. (2007). From mere coincidences to meaningful discoveries. Cognition, 103(2), 180-226.

Harris, P. L., German, T., & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition*, *61*(3), 233-259.

Heyes, C. (2012). Simple minds: a qualified defence of associative learning. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1603), 2695-2703.

Houwer, J. D., Beckers, T., & Glautier, S. (2002). Outcome and cue properties modulate blocking. *The Quarterly Journal of Experimental Psychology: Section A*, *55*(3), 965-985.

Kenderla, P., & Kibbe, M. M. (2023). Explore versus store: Children strategically trade off reliance on exploration versus working memory during a complex task. *Journal of Experimental Child Psychology*, *225*, 105535.

Kimura, K., & Gopnik, A. (2019). Rational higher‐order belief revision in young children. *Child Development*, *90*(1), 91-97.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. Cognition, 83(2), B35-B42.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. Psychological review, 99(1), 22.

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin and Review*, *7*(4), 636-645.

Larkin, M. J., Aitken, M. R., & Dickinson, A. (1998). Retrospective revaluation of causal judgments under positive and negative contingencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1331.

Legare, C. H., Gelman, S. A., & Wellman, H. M. (2010). Inconsistency with prior knowledge triggers children’s causal explanatory reasoning. Child development, 81(3), 929-944.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality?. *Cognition*, *25*(3), 265-288.

Lovibond, P. F. (2003). Causal beliefs and conditioned responses: retrospective revaluation induced by experience and by instruction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(1), 97.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. Science, 283(5398), 77-80.

McClelland, J. L., & Thompson, R. M. (2007). Using domain‐general principles to explain children's causal reasoning abilities. Developmental Science, 10(3), 333-356.

McCormack, T., Butterfill, S., Hoerl, C., & Burns, P. (2009). Cue competition effects and young children’s causal and counterfactual inferences. *Developmental psychology*, *45*(6), 1563.

Meltzoff, A. N., Waismeyer, A., & Gopnik, A. (2012). Learning about causes from people: observational causal learning in 24-month-old infants. *Developmental psychology*, *48*(5), 1215.

Oakes, L. M., & Cohen, L. B. (1990). Infant perception of a causal event. *Cognitive Development*, *5*(2), 193-207.

Powell, L. J., & Carey, S. (2017). Executive function depletion in children and its impact on theory of mind. *Cognition*, *164*, 150-162.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Classical conditioning II: Current research and theory, 2, 64-99.

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children’s development of analogical reasoning: Insights from scene analogy problems. *Journal of experimental child psychology*, *94*(3), 249-273.

Rogers, T. T., & McClelland, J. L. (2014). Parallel distributed processing at 25: Further explorations in the microstructure of cognition. Cognitive science, 38(6), 1024-1077.

Rovee-Collier, C. (1999). The development of infant memory. Current directions in psychological science, 8(3), 80-85.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. Science, 274(5294), 1926-1928.

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

Shultz, T. R. (1982). Rules of causal attribution. Monographs of the society for research in child development, 1-51.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *The Quarterly Journal of Experimental Psychology Section B*, *37*(1b), 1-21.

Sobel, D. M. (2004). Exploring the coherence of young children's explanatory abilities: Evidence from generating counterfactuals. *British Journal of Developmental Psychology*, *22*(1), 37-58.

Sobel, D. M., & Kirkham, N. Z. (2006). Blickets and babies: the development of causal reasoning in toddlers and infants. *Developmental psychology*, *42*(6), 1103.

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.

Spencer, J. P., Ross‐Sheehy, S., & Eschman, B. (2022). Testing predictions of a neural process model of visual attention in infancy across competitive and non‐competitive contexts. Infancy, 27(2), 389-411.

Steinbeis, N. (2018). Taxing behavioral control diminishes sharing and costly punishment in childhood. *Developmental science*, *21*(1), e12492.

Stojnić, G., Gandhi, K., Yasuda, S., Lake, B. M., & Dillon, M. R. (2023). Commonsense psychology in human infants and machines. Cognition, 235, 105406.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and motivation*, *25*(2), 127-151.

Walker, C. M., & Gopnik, A. (2014). Toddlers infer higher-order relational principles in causal learning. *Psychological science*, *25*(1), 161-169.

Walker, C. M., & Nyhout, A. (2020). Asking “why?” and “what if?”: The influence of questions on children’s inferences. The questioning child: Insights from psychology and education, 252-280.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. Stanford Univ Ca Stanford Electronics Labs.

Xu, F. (2019). Towards a rational constructivist theory of cognitive development. Psychological review, 126(6), 841.

Zelazo, P. D., Frye, D., & Rapus, T. (1996). An age-related dissociation between knowing rules and using them. *Cognitive development*, *11*(1), 37-63.

Zelazo, P. D., Müller, U., Frye, D., Marcovitch, S., Argitis, G., Boseovski, J., ... & Carlson, S. M. (2003). The development of executive function in early childhood. *Monographs of the society for research in child development*, i-151.

1. The Griffiths et al. (2011) model assumes that this can be learned through a hierarchical process; we are presenting a simpler model for the purposes of this investigation. [↑](#footnote-ref-1)