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 network-modeling 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 human beings to learn about the complex interactions in the world. The mechanisms that underpin 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; causal mechanisms; computational models; analytical models; associative learning; Bayesian inference

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

Causal reasoning – the capacity to

Few capacities are more important than the ability to reason and make inferences about cause and effect relations – is central for learning about the world. Causal reasoning enables human learners to make predictions and inferences (e.g., Bullock, et al., 1982; Leslie & Keeble, 1987; Oakes & Cohen, 1990; Shultz, 1982), 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., 2014; 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 underpinned by a Bayesian-inference mechanism that is in place early in development. 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; Griffiths & Tenenbaum, 2005) describing how children make causal inferences (e.g., Gopnik & Wellman, 2012; Griffiths & Tenenbaum, 2005, 2007; Xu, 2019). On this view, children’s causal reasoning starts with associative and statistical learning capacities that are present in infancy (e.g., Gomez, 2002; Johnson et al., 2006; 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 explain children’s causal inferences. Connectionist accounts of causal reasoning (e.g., Rogers & McClelland, 2004), 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) suggest that associative processing is 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 retrospective inference in which learners reevaluate 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 et al., 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 were placed on it (Gopnik & Sobel, 2000). Children were then shown that two novel objects, A and B, together caused the detector to activate when they were placed on the machine. Children were then shown that object A alone activated the machine. Children in both conditions were then asked which of the two objects were blickets and to make the machine go by placing the blicket on the detector. Sobel et al. (2004) found that the 4-year-olds (and to a lesser extent, 3-year-olds) were less likely to place B on the machine than on trials in which object A did not make the machine go by itself (a condition referred to as *indirect screening off*). Using an anticipatory eye-tracking procedure, Sobel and Kirkham (2006) found that 8-month-olds showed this same response pattern.

These findings have been interpreted to mean that children’s causal inferences are governed by Bayesian inference rather than associative learning. This is because some associative models (e.g., Rescorla-Wagner model) 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, some caution should be exercised before accepting the conclusion that Bayesian inference rather than associative learning underpins how children process backwards blocking events. One reason is that it is unclear whether the difference in how children treated object B was due to backwards blocking, indirect screening-off, or some combination of both. Both Beckers et al. (2005) and McCormack et al. (2009) were the first to raise this objection. McCormack et al. (2009) showed when a more appropriate operationalization is adopted—in which participants treatment of object B following the standard backwards blocking event (i.e., AB+; A+) is compared to their treatment of object B following a closely matched control event (i.e., AB+; D+)—4-year-olds treated object B equivalently across both trials; the 5-year-olds, in contrast, were more likely to place object B on the machine in the control condition than in the backwards blocking condition. The critical point of agreement between these investigations is that at some point in development, children have the capacity to make retrospective inferences. But the fundamental question remains as to how children engage in such reasoning.

**Bayesian vs. Associative Reasoning**

Another reason to exercise caution before accepting the claim that human beings use Bayesian inference to engage in backwards blocking reasoning is that it remains unknown whether human children engage in backwards blocking reasoning for three (or more) objects. Consider a modified version of the standard backwards blocking event in which children first see an ABC+ sequence followed by an A+ sequence. If backwards blocking reasoning is unaffected by the number of presented objects, then children should be less likely to label objects B *and* C as blickets compared to the same objects in a control event in which ABC+ is followed by D+. This question is worth addressing because if the goal is to elucidate and better understand the nature of the cognitive mechanisms that subserve causal reasoning *in the real world*, then it is crucial that we understand how causal reasoning unfolds in situations that mirror children’s natural environments.

Sobel et al. (2004) also demonstrated that changing the base rate of causally efficacious objects affected children’s backwards blocking inference. Four-year-olds observed that only two out of the first twelve objects they saw activated the machine, a backwards blocking inference about the 13th and 14th objects they observed on the machine was different than if they saw 10 out of the first twelve objects activate the machine. Children discounted the activation of object B with object A on the machine when few objects activated the machine, but did not when many objects did so. Three-year-olds did not show this inference, but they did if they understood that the objects held a deterministic relation to the machine’s activation because of a nonobvious property, regardless of the causal domain in which children were tested (Sobel & Munro, 2009).

These data were fit to a Bayesian model that these and other researchers have used to describe children’s causal inference (e.g., Gopnik & Wellman, 2012; Griffiths et al., 2011; Tenenbaum & Griffiths, 2001). For example, Griffiths et al. (2011) showed that both children and adults reasoned about backwards blocking with base rates according to a Bayesian algorithm, but that algorithm made predictions about other kinds of causal inferences. Particularly, both adults and 4-year-olds inferred that when three objects (A, B, and C) were placed on the machine in combinations (A and B activated it, and then A and C activated it), when efficacious objects were rare, object A was most likely to be efficacious, but not at ceiling levels, and B and C would have equal likelihood at the end of the sequence, greater than baseline, but lower than A.

However, one could also consider the evidence in favor of Bayesian inference as indicating that children make higher-older correlations among events. Benton et al. (2021) suggested that children can make causal inferences based on second-order correlations – associations of associations (see also Cuevas et al., 2006; Rakison & Benton, 2019; Yermolayeva & Rakison, 2016), providing a proof of concept that the same data in support of Bayesian models as a description of children’s causal inferences could be explained by higher-order statistical learning models, such as connectionist networks. McClelland and Thompson (2007) presented a connectionist account of the initial findings on backwards blocking on children; we suspect that deep learning architectures could be adopted to explain findings about such inferences that also took the base rates information of the candidate cause into account (although to our knowledge, no such model has been published).

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 – when shown that two pairs of compound stimuli (A and B, and then A and C) were efficacious, the 4-year-olds 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). That is, 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 be less likely to use a rational Bayesian mechanism and more likely to look more associative in their inferences.

The presence of such information processing demands affecting children’s inferences might be surprising. SENTENCE ABOUT INFORMATION PROCESSING IN INFANT PERCEPTION OF CAUSALITY. 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). 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 they still show default to certain kinds of associative inferences (Hermes et al., 2018; Luchkina et al., 2020). Relevant to the present investigation, Sobel et al. (2017; see also Erb & Sobel, 2014) showed that between the ages of 4-7, children develop the capacity to form larger hypothesis spaces of the potential causes they might need to hold to engage in more rational inferences. Because of this finding, we considered 5- and 6-year-olds in the present study.

Here, we asked children to make retrospective causal inference about three potential causes as opposed to only two, using a logic like that of McCormack et al. (2009). In , 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, followed by a second experiment that considers a prediction that emerges from the comparison of these models.

**Experiment 1**

Participants were 32 5-year-olds (16 boys and 16 girls, *Mage* = 64.81 months, SD = 3.48) and 32 6-year-olds (17 boys and 15 girls, *Mage* = 77.81 months, SD = 3.78). Sample size was determined based on previous studies on BB reasoning in human children (e.g., Gopnik & Sobel, 2000; Griffiths et al., 2011; Sobel et al., 2004). Two additional children were tested, but not included in the analysis for failing to participate (*N* = 1) or missing video (which made coding their responses impossible, *N* = 1)..

***Materials***

We presented children with a computer-animated version of the blicket detector (e.g., Gopnik & Sobel, 2000). On a computer screen, the machine1 below).At the start of any given trial, three (for the BB or ISO experimental trials) or four (for the BB or ISO control trials) equally spaced circles appeared above the machine.

***Procedure***

This event ensured that participants understood the task and recognize that individual objects could activate the machine and that the machine activated if at least one efficacious 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.

1.The upper-right portion of the figure shows the BB event as it unfolded across time.

**Backwards Blocking Trials and Control.** The two BB main trials began with three differently colored objects, which were located above the machine. The experimenter read the text on the screen, “Look, I have these three toys. Watch what happens.”objects A, B, and C) then descended until they rested on top of the machine, which activated by lighting up blue and making a sound. At this point, text appeared on the screen, which the experimenter read: “Look, these also make the machine go!” The objects then returned to their starting positions.

Then the left- or right-most (counterbalanced) object (which we will refer to here as object A) descended until it contacted and immediately activated the machine. The experimenter read the text on the screen, “Look, this one makes the machine go!” Children received two of these trials, which were identical except for the color of the objects.

The two BBAgain, children received two trials, which were identical except for the color of the objects.

**Indirect Screening Off Trials and Control.** The procedure for children who received the indirect screening off trials and controls was identical to the backwards blocking trials, except that when the A object (Experimental trial) or the D object (Control trial) was placed on the machine by itself, the machine did not activate.

Table 1. Schematic of Experiment. Children receive two trials of BB Experimental and Control or ISO Experimental and Control

|  |  |  |  |
| --- | --- | --- | --- |
| Schematic of Experiment | | | |
|  |  |  |  |
| BB Experimental trial | ABC+ | A+ | Is A/B/C a blicket? |
| BB Control trial | ABC+ | D+ | Is A/B/C/D a blicket? |
| ISO Experimental trial | ABC+ | A- | Is A/B/C a blicket? |
| ISO Control trial | ABC+ | D- | Is A/B/C/D a blicket? |

Table 1. The +/- signs corresponds to whether the machine activates (+) or not (-)

Figure 2 shows the mean percentages of responses to the “Is it a blicket?” question for each object at the end of each type of trial. To account for the within-subject nature of this question, we analyzed the frequency with which children stated each object was a blicket across the trial types and conditions with a Generalized Estimating Equation assuming a robust correlation matrix. Condition (Backwards Blocking vs. Indirect Screening Off), Trial Type (Experimental vs. Control), Object (A, B, C, and D), and age (in months) were fixed factors.

Figure 2. Participants’ mean responses to whether each object was categorized as a blicket across the conditions and trial types. Bars show standard error.**Chart, bar chart

Description automatically generated**

This model revealed a no significant main effects, but two significant interactions: a significant two-way interaction between object and condition, Wald 2(2) = 8.05, *p =* .05, and a significant three-way interaction between object, condition, and age, Wald 2(3) = 9.43, *p =* .02. In particular, parameter estimates of this model suggest that as children got older, they treated objects B and C differently than object A between the two conditions, B = 0.04, SE = 0.005, 95% CI [0.003, 0.02], Wald2(1) = 6.83, *p =* .009, for object B, and B = 0.02, SE = 0.01, 95% CI [-0.003, 0.04], Wald2(1) = 3.03, *p =* .08 for object C.

To analyze these interactions, we divided children by condition and age group. In the Backwards Blocking condition, both the 5-year-olds (60% of the time vs. 85%) and 6-year-olds (56% vs. 75%) categorized the B and C objects as less likely to be blickets in the experimental trials than the control trials, *t*(14) = 2.42 and *t*(15) = 2.42 respectfully, both *p*-values = .03. In contrast, in the Indirect Screening Off condition, the 5-year-olds did not differ in their average judgment of the efficacy of objects B and C in the experimental trials (80%) and the control trials (85%), *t*(15) = 1.57, *p* = .14, while the 6-year-olds showed a larger difference (77% vs. 86%), but only at a non-significant trend, *t*(16) = 1.86, *p* = .08.

Finally, we considered performance between the trials. Children treated object A differently in the experimental trials of the Backwards Blocking and Indirect Screening Off conditions, judging A as a blicket on 98% of the trials in the former and 26% of the trials on the latter, *t*(62) = 9.11, *p* < .001. Their judgments of objects B and C also differed between these conditions, with them stating these objects were blickets on 58% of the backwards blocking trials and 77% of the indirect screening off trials, *t*(62) = -2.23, *p* = .03.

Experiment 1 showed evidence that children retrospectively reevaluated the causal efficacy of ambiguous data, although the evidence suggesting that they do so was not as clear as in previous investigations. When shown ambiguous evidence that three objects together have efficacy, and then that one of those objects alone has the same efficacy (backwards blocking), children were less likely to judge the other two objects as having efficacy compared to a control condition in which a fourth, independent object had such efficacy and compared to the case in which that object individually did not have efficacy (indirect screening off). However, in that latter case, the judgment that the other two objects did have efficacy did not differ compared to the same type of control condition, particularly for the younger children we tested.

**Model Fits**

We fit two competing computational models to these data. The first was a model based on Bayesian inference, described initially by Sobel et al. (2004) and in more detail in Griffiths et al. (2011). The second was a connectionist model following work by Rogers and McClelland (2014). The computational details of both models are described in the Supplemental Materials. Below, we briefly describe each 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 hypothesis a learner has 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*).

The initial hypothesis space relies on assuming 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 told that the machine in front of them detects blickets, which is defined by the object having the efficacy to activate the detector, 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 about this form of structure learning). Griffiths et al. (2011) describe the formal parameterization of this hypothesis space and model that results in the hypothesis space shown in Figure 2. This figure represents the eight possible hypotheses reasoners might have about the experimental trials.

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-**

Figure 2. Possible hypothesis for the experimental trials of the backwards blocking and indirect screening off conditions in which Objects A-C could each activate the detector (E). Note that this hypothesis space is similar for the control trials, but must include a fourth D object.

**Timeline

Description automatically generated with medium confidence**

Because the model assumes that objects with causal efficacy will act deterministically on detectors, 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 posit the causal relation between the object and the detector. 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.

Given the pretest, in which children observe two objects, one of which activates the machine and one which does not, it is possible to assume **= 0.5 as a starting point for the model. Table 2 presents the probability that each object is a blicket given the data presented to them in each type of trial. When 5- and 6-year-olds are considered as separate ages, this model fits the data the level of *r* = .86, with a Bayes Factor of 0.00000028, suggesting that the model is strongly related to these data.

*Table 2. Results of Bayesian Model, assuming* **= 0.5.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Object A | Object B | Object C | Object D |
| Backwards Blocking Experimental | 1 | .5 | .5 | N.A. |
| Backwards Blocking Control | .57 | .57 | .57 | 1 |
| Indirect Screening Off Experimental | 0 | .67 | .67 | N.A. |
| Indirect Screening Off Control | .57 | .57 | .57 | 0 |

**Connectionist Network. DEON TO ADD**

One may question whether asking children to reason about three to four objects can really tell us more about the cognitive mechanisms that underpin causal reasoning than asking children to reason about two objects. This is because the two situations differ trivially by at most two potential causes. However, if Bayesian inference is the cognitive mechanism that underpins backwards blocking reasoning in human beings, then the difference between these two settings is far from trivial. This is because in the two-cause setting, 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 and children are asked to reason about four blickets, then there are 24 possible combinations of blickets and non-blickets. In contrast, in a three- or four-cause setting like that just discussed, participants need to determine which of *eight* (in the case of 3 objects) or *sixteen* (in the case of 4 objects) hypotheses is the right none. This means that participants must consider up to four times as many causal hypotheses across these two situations.

In light of these unresolved issues, the present investigation had two goals. The first goal was to examine backwards blocking in the presence of multiple potential causes and using a logic like that of McCormack et al. (2009). Five- and 6-year-old children were introduced to a computer-animated “blicket detector” machine and were told that their task was to determine which objects activated the machine. We chose to test 5- and 6-year-olds as opposed to 4-year-olds because the latter group likely has more robust information-processing capacities than the former group and thus a greater chance of engaging in Bayesian inference. Research by Sobel et al. (2017; see also Erb & Sobel, 2014) is consistent with this contention. Participants then received either two backwards blocking and two backwards blocking control trials or two indirect screening-off and two indirect screening-off control trials. Participants in both conditions were then asked to indicate whether the objects in each trial were blickets. The second goal was to determine whether children’s causal inferences were best explained by an associative-learning mechanism or a simple Bayesian mechanism. We did this by fitting a connectionist (associative learning) model and a Bayesian model to participants’ data. The Experiment below addresses the first goal. The “Computational Models” section addresses the second goal.

Figure 2A-E shows these predictions.

As shown in Figure 2A-E, the model predicts that during the BB experimental and control trials participants should be maximally confident that objects A and D are blickets. In contrast, during the ISO experimental and control trials, participants should be maximally confident that objects A and D are not blickets. Importantly, the model makes these predictions regardless of the prior probability of blickets. In contrast, the model predicts that participants should categorize objects B and C at the same rate across the main trials in the BB and ISO conditions and objects A-C at the same rate across the corresponding control trials.

**Associative learning**

In contrast to the Bayesian model, we built a simple, two-layer connectionist computational model (Figure 3). This model was designed to simulate the current experiments

Diagram

Description automatically generated  
Figure 3.

Chart, bar chart

Description automatically generated

A

B)

Figure 4. Connectionist model predictions for how participants should treat the objects between the main and control trials of the BB and ISO conditions.

As can be seen in Figure 4 above, the model predicts that participants should treat objects A-C equivalently during the BB experimental trials. In contrast, the model predicts that participants should treat object A as more of a blicket than objects B and C during the same trials. For the ISO experimental trials, the model predicts that participants should treat object A as less of a blicket than objects B and C during the ISO experimental trials.

It should be noted that the Bayesian and connectionist model make identical qualitative predictions for all the conditions and trials except for the BB control condition: the simple connectionist model predicts that participants should treat objects A-D equivalently during this trial; the simple Bayesian model predicts that participants should only treat objects A-C equivalently but should be maximally confident that object D is a blicket. Thus, it should be possible to determine which model participants relied on based on their performance during the BB control condition. Interestingly, both models predict that participants’ treatment of the redundant causes between the BB experimental and BB control trials should not differ. Likewise, both models predict that participants’ treatment of the redundant causes between the BB main and ISO main conditions should not differ. Thus, the simple connectionist model and Bayesian model do not predict BB reasoning according either to the new or old operationalization of BB reasoning. The present study was designed to test these predictions.

**The present investigation**

The present investigation had two goals. First, it was designed to use a more valid measure of BB reasoning to determine whether 5- and 6-year-olds could engage in this form of reasoning for three and four objects. Second, it was designed to illuminate how—that is, by what cognitive mechanism—children reasoned about the present causal events. Our specific aim was to determine which of two cognitive mechanisms—Bayesian inference or associative learning—best explained children’s causal inferences in the present context.

**Current study**

**.** Participants were 32 5-year-olds (16 boys and 16 girls; *M* = 64.81 months, range = 60-71 months) and 31 6-year-olds (17 boys and 15 girls; *M* = 77.81 months, range = 72-83 months). 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).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.

The upper-right portion of the figure shows the backwards blocking event as it unfolded across time.

Following the pretraining phase, participants were given four test trials—either the two backwards blocking experimental trials and 2 backwards blocking control trials or two indirect screening-off experimental trials and 2 indirect screening-off control trials—in counterbalanced order using a Latin square. Differently colored objects were used across all trials to prevent carryover effects.

The two backwards blocking main trials began with three differently colored objects, which were located above the machine. The text, “Look, I have these three toys. The first and second backwards blocking experimental trials were identical except for the object colors (see Figure 5 for a schematic of the backwards blocking experimental event).

The two backwards blockingThe experimental and control trials used the same text.

Finally, the indirect screening-off experimental and control conditions were identical to the backwards blocking experimental and control conditions except that objects A (during the indirect screening-off main trials) and D (during the indirect screening-off control trials) failed to activate the machine (see Table 1 for a schematic).

|  |  |  |  |
| --- | --- | --- | --- |
| Schematic of Experiment 1 | | | |
|  |  |  |  |
| backwards blocking experimental trial | ABC+ | A+ | Is A/B/C a blicket? |
| backwards blocking control trial | ABC+ | D+ | Is A/B/C/D a blicket? |
| indirect screening-off experimental trial | ABC+ | A- | Is A/B/C a blicket? |
| indirect screening-off control trial | ABC+ | D- | Is A/B/C/D a blicket? |

Table 1. The +/- signs corresponds to whether the machine activates (+) or not (-)

Chart, bar chart

Description automatically generated

Figure 6. Participants’ mean responses to whether each object was a blicket across the conditions and trial types. Bars show standard error.

Figure 6 shows the results for this experiment. The dependent measure was the number of times that participants responded “Yes” to the “Is this a blicket” question. Thus, across two trials, the maximum number of times that a participant could respond “Yes” was 2; the minimum number of times that a participant could respond “Yes” was 0. 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-subjects fixed effect, Trial Type (experimental vs. control),Objects (A vs. B vs. C vs. D), and Phase Order (Phase 1 vs. Phase 2) as the within-subjects fixed effects, and subjects as the random effect. This analysis yielded several main-effects and two-way interactions, which were qualified by 3 three-way interactions. These included a three-way interaction between Age, Condition, and Object, *χ2*(3) = 12.75, *p* = .005, a three-way interaction between Condition, Phase Order, and Object, *χ2*(3) = 13.91, *p* = .003, and a three-way interaction between Condition, Trial Type, and Object, *χ2*(2) = 78.59, *p <* .0001. Given that the lattermost interaction was the only theoretically relevant and meaningful one, all subsequent analyses focused on this interaction. This three-way interaction is shown in Figure 3.

We followed up this three-way interaction with separate one-way linear mixed-effects models for the main and control trials within the backwards blocking and indirect screening-off conditions. The Objects factor was treated as the sole within-subjects fixed effect in these follow-up analyses. Subjects were once again treated as a random effect. The first one-way linear model for the control trials within the backwards blocking condition did not reveal a significant effect of Objects, *χ2*(3) = 1.08, *p* = .78. This means that participants treated the objects similarly during the control trials of the backwards blocking condition. In contrast, the second one-way linear model for the main trials within the backwards blocking condition revealed a significant main effect of Objects, *χ2*(2) = 29.78, *p* < .0001. This main effect reflected the fact that participants considered object A to be more of a blicket (*M* = 1.97, *SD* = 0.18) than object B (*M* = 1.21, *SD* = 0.83), *t*(27) = 4.70, *p<* .001, or object C (*M* = 1.32, *SD* = 0.82), *t*(27) = 4.12, *p* < .001. Participants treated objects B and C equivalently, *t*(27) = -0.72, *p* = .48.

The third and fourth one-way linear models for the main and control trials within the indirect screening-off condition both revealed a significant main effect of Objects, both *χ2*’s > 60.30, both *p*’s < .0001. This reflected the fact that participants considered object A (*M* = 0.52, *SD* = 0.87) to be significantly less of a blicket than objects B (*M* = 1.70, *SD* = 0.67) and C (*M* = 1.74, *SD* = 0.59) during the main condition, both *t*’s > -5.97 and both *p*’s < .0001. Likewise, participants considered object D (*M* = 0.73, *SD* = 0.72) to be less of a blicket than objects A (*M* = 1.76, *SD* = 0.58), B (*M* = 1.69, *SD* = 0.59), and C (*M* = 1.76, *SD* = 0.50) during the control trials, all *t*’s > -6.01, all *p*’s < .001.

To examine whether there was evidence specifically of backwards blocking, data were entered into a two-way linear mixed-effects model with Trial Type and Objects as the within-subjects fixed effects and subjects as the random effect. This analysis revealed only a main effect of Trial Type, *χ2*(1) = 10.14, *p* < .005. This result reflected the fact that participants were more likely to respond that a redundant object was a blicket during the control trials (*M* = 1.60, *SD* = 0.71) than during the main trials (*M* = 1.27, *SD* = 0.82), *F*(1, 139) = 5.28, *p* = .02. This analysis indicates that participants did engage in backwards blocking reasoning.

The results of Experiment 1 indicated that participants engaged in backwards blocking reasoning. Specifically, participants were more likely to treat the redundant objects as blickets during the backwards blocking control trials than during the backwards blocking experimental trials. These results indicate that participants show some evidence of backwards blocking reasoning when asked to make inferences about multiple candidate causes. However, an open question concerns whether participants will continue to engage in backwards blocking reasoning when two rather than one object participants on the machine during the second (i.e., the A+) phase of the backwards blocking condition.

**Experiment 2**

**Method**

**Participants.** ADD RELEVANT INFO WHEN WE GET IT.

**Stimuli, Design, and Procedure.**  All aspects of Experiment 2 were identical to Experiment 1 with one exception: Two objects participated on the machine during the elemental phases of the backwards blocking and indirect screening-off events. Thus, participants in the backwards blocking condition saw an ABC+ AB+ series of events during the experimental trials and an ABC+ DE+ series of events during the control trials. In contrast, participants in the indirect screening-off condition saw an ABC+ AB- series of events during the experimental trials and an ABC+ DE- series of events during the control trials. Similar to Experiment 1, to prevent potential carryover effects differently colored objects were used across all trials to prevent carryover effects and the left-right positions of objects A and B during the experimental trials and D and E during the control trials in both conditions were counterbalanced.

Computational Models

**A simple Bayesian computational model**

A key assumption of Bayesian inference is that causal induction is a process that involves representing the entire space of candidate causal hypotheses—which can be expressed as parameterized causal graphical models with nodes that are connected by edges that encode the Markov condition—and then choosing the hypothesis that is most consistent with the data. We can use Bayes rule to choose among these hypotheses. Formally, it is assumed that at the beginning of a learning episode, an ideal learner represents all possible candidate hypotheses, *H*, whereby each hypothesis, *h* ∈ *H*, is assigned some prior probability, *p*(*h*). This prior probability represents the learners’ confidence that a given hypothesis generated that causal data. Figure 7 below shows the hypothetical hypothesis space for three objects.

.

**Timeline

Description automatically generated with medium confidence**

Figure 7. 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.

Following observations of data, *d*, the learner uses Bayes' rule to compute and assign posterior probabilities to each hypothesis, *p*(*h*|*d*),

,

where *p*(*d*|*h*) represents the likelihood or the probability of the data *d* under a given hypothesis *h*. The denominator serves as the normalizing term—it allows the posterior probabilities the hypotheses to sum to 1. Given that the machine behaved deterministically in the present context (i.e., objects either produce or do not produce detector activation), the likelihoods are set to 1 whenever a link (i.e., causal relation) exists in the hypothesis and is consistent with the observed data; otherwise, they are set to 0. Once we have determined whether such a link exists for a particular object, we can compute the likelihood that any of the objects is a blicket by taking the product of the likelihood that that object activated the detector under each hypothesis and the prior probability of each hypothesis and then summing this product. For example, to determine the probability that object B is a blicket, we can compute the following equation

,

where equals 1 if a causal link between *B* and *E* existsfor a specific hypothesis *h*; otherwise, equals 0.

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 8A-E1 shows the model’s predictions for Experiments 1 and 2. Figure 8A-E shows these predictions. We also plotted the model’s predictions for various prior probabilities because 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 the model’s predictions for the different probabilities to children’s actual treatment of the objects. Figure 8A-E shows these predictions.

|  |  |
| --- | --- |
| 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  Description automatically generated  E | Chart, bar chart  Description automatically generated  A1 |
| Chart, bar chart  Description automatically generated  B1 | Chart, bar chart  Description automatically generated  C1 |
| Chart, bar chart  Description automatically generated  D1 | Chart, bar chart  Description automatically generated  E1 |

Figure 8A-E1. A-E displays the Bayesian model’s predictions for the backwards blocking and indirect screening-off conditions in Experiment 1 when *P*(Blickets) = .5 (3A), *P*(Blickets) = .65 (3B), *P*(Blickets) = .8 (3C), *P*(Blickets) = .95 (3D), *P*(Blickets) = 1 (3E). A1-E1 displays the Bayesian model’s predictions for the backwards blocking and indirect screening-off conditions in Experiment 2 for the same probabilities.

With the exception of Figure 8E and 8E1 in which the baseline probability that an object is a blicket is 100% (and thus all objects should be considered to be blickets except those that are shown explicitly not to activate the machine), the model makes several notable predictions for how participants should treat the objects between the backwards blocking and indirect screening-off experimental and control trials in Experiments 1 (A-D) and 2 (A1-D). First, participants in Experiment 1 should be maximally confident that objects A and D are blickets in the backwards blocking condition main and control trials, respectively, but maximally confident that objects A and D are not blickets in the indirect screening-off condition main and control trials. Second, participants in Experiment 2 should be more confident that objects A and B (during the main trials) and D and E (during the control trials) are blickets than the causally redundant objects (i.e., objects C in the main condition, and objects A-C in the control condition) in the backwards blocking condition main and control trials, but maximally confident that objects A and B (during the main trials) and D and E (during the control trials) are not blickets in the indirect screening-off condition. Third, 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 main trials.

**A simple connectionist computational model**

To examine whether participants’ inferences in the present experiment are best explained by an associative-learning mechanism we built a simple two-layer connectionist model. The network architecture is shown below in Figure 9

Diagram

Description automatically generated  
Figure 9.

For example, networks, like children, were assigned randomly to the indirect screening-off condition or to the backwards blocking condition.For example, during the two “experimental trials” for networks in the backwards blocking condition, the first three input units were turned on (i.e., activated the machine.During the subsequent A+ trials, only the first input unit was turned on, but again the network’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. The indirect screening off experimental and control trials were identical to the backwards blocking experimental and control trials except that the network was trained to turn off the single output unit (i.e., set its value to 0) during the A- and D- phases of the indirect screening-off experimental and control trials, respectivelyABC+) and elemental (e.g., A+/D+; A-/D-)) epochs. The predictions that this model makes for how participants should treat the different objects across the trials and conditions are shown below in Figure 8

|  |  |
| --- | --- |
| 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  A1 | Chart, bar chart  Description automatically generated  B1 |
| Chart, bar chart  Description automatically generated  C1 | Chart, bar chart  Description automatically generated  D1 |
|  |  |

Figure 10A-D1. A-D shows the connectionist model’s predictions for how participants should treat the objects between the main and control trials of the backwards blocking and indirect screening-off conditions in Experiment 1 after 800 (A), 1,600 (B), 2,000 (C), and (3) 3,000 epochs of training. A1-D1 shows the connectionist model’s predictions for how participants should treat the objects between the main and control trials of the backwards blocking and indirect screening-off conditions in Experiment 2 after the same amount of training.

The connectionist model predicts that participants should treat the redundant objects equivalently across the different trials and conditions but should be more confident that object A and objects A and B (in the backwards blocking main and control trials, respectively) than the redundant objects. 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 main and control trials in Experiment 1, respectively) and objects A and B and D and E (in the main and control trials in Experiment 2, respectively) are blickets compared to the causally redundant objects across the same trials.

The derived predictions from the simple Bayesian model and the connectionist model indicate clearly that they make distinct predictions for how participants should respond to the objects across the various conditions, trials, and two experiments. However, these models especially diverge 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 simple Bayesian model predicts that participants should only treat the objects differently. The models also differ in terms of whether they predict backwards blocking reasoning; the Bayesian model, but not the associative model, predict backwards blocking reasoning. Thus, it should be possible to determine which model participants relied on based on their performance during the backwards blocking control condition.

|  |  |
| --- | --- |
| **Qualitative and Quantitative Model fits: Experiment 1**  **Chart, bar chart  Description automatically generated**  A | |
| Chart, bar chart  Description automatically generated  B | Chart, bar chart  Description automatically generated  C |

Figure 4A-C. Participants data (6A) compared to the predictions of the best-fitting connectionist model trained for 800 epochs (6B) and the best fitting Bayesian model, where the *P*(Blicket) = .80 (6C).

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 and participants’ mean responses to objects A-C during the backwards blocking and indirect screening-off main trials and objects A-D during the backwards blocking and indirect screening-off control trials. These two metrics have been used in previous simulation studies to assess model’s quantitative fit to behavioral data (e.g., Bhat et al., 2022). Lower values on each metric indicate better model fit. Table 2 shows the fits for the different model instantiations.

|  |  |  |
| --- | --- | --- |
| Computational Models | root mean square (RMSE) | mean absolute error (MAE) |
| Connectionist Model (800 epochs)++ | .19 | .15 |
| Connectionist Model (1600 epochs) | .22 | .16 |
| Connectionist Model (2000 epochs) | .25 | .17 |
| Connectionist Model (3000 epochs) | .28 | .22 |
| Bayesian model (.5) | .46 | .43 |
| Bayesian model (.65) | .34 | .29 |
| Bayesian model (.80)++ | .29 | .20 |
| Bayesian model (.95) | .39 | .33 |
| Bayesian model (1) | .45 | .39 |

Table 2. Model fit indices for the various models and instantiations. ++ indicates the best fitting connectionist and Bayesian models.

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. It should also be clear that the connectionist model provides a better *qualitative* fit to the data than the Bayesian model. 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.

Given that the connectionist model instantiated associative learning, these results suggest that participants may have used associative learning rather than Bayesian inference to process the present events. We discuss below the exact nature of this associative learning.

General Discussion

This study had two aims. The first was to determine whether 5- and 6-year-olds would engage in backwards blocking reasoning for 3 and 4 objects. This departs from the typical convention of using two objects to study causal reasoning in human children. The second aim was to clarify how exactly (i.e., the cognitive mechanism by which) children reasoned about the present causal events. We were specifically interested in whether children’s causal inferences best conformed to the predictions of a simple Bayesian model or a connectionist (associative learning) model.

With respect to the first aim, we found evidence of backwards blocking reasoning under the old operationalization of backwards blocking reasoning but only minimal evidence of such reasoning under a new and more valid measure of backwards blocking reasoning. This finding extends previous research to show that when children are asked to reason about three objects and a more valid of operationalization of backwards blocking reasoning is used, children minimally engage in backwards blocking reasoning.

With respect to the second aim, the data were most consistent with the connectionist (associative learning) models. However, a question that we have not yet answered concerns the exact nature of the connectionist models’ associative learning. Specifically, how did these models—via associative learning—arrive at their causal judgements? To understand how the connectionist model’s judgements arose mechanistically, consider the backwards blocking (i.e., ABC+ D+) control trial. The rationale for focusing on this condition is that the connectionist uniquely predicts participants performance during the backwards blocking control 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 single output unit was equivalent for all four objects, the network made the same weight adjustments both in sign and magnitude to the connections (which encode associations) between each object and the single output unit. This explains why the network responded equivalently when “asked” whether each object (i.e., objects A-D) was a blicket. Given that participants’ causal responses mirrored the predictions of the connectionist model, this suggest that children arrived at their causal judgements via a similar associative-learning process. This finding is significant because it has been suggested that causal reasoning in human children is best explained by Bayesian inference and rational processes rather than by associative processes. The present data add nuance and context to this debate by suggesting that whether children engage in Bayesian inference or associative learning may depend on how causal reasoning is measured and the number of objects about which they are asked to reason.

These aims aside, some potential criticisms are worth noting. One such potential criticism is that the results are inconsistent with the findings from previous studies on backwards blocking reasoning in human children. Such previous research showed that children unequivocally engage in backwards blocking reasoning when asked to reason about two objects; in contrast, the current study only provided equivocal evidence for backwards blocking reasoning when three and four objects were used. However, we believe that the present results extend rather than are at odds with such previous research. Specifically, the present study likely demonstrates that when children’s information-processing capacities are stretched such as when they are asked to reason about multiple potential causes, they may deploy and rely on simpler associative processes. Although the numerical difference between three and four objects is miniscule, by contrast the corresponding increase in the size of the underlying psychological hypothesis space is substantial. Such an increase in the size of the underlying psychological hypothesis space may have important ramifications on the cognitive mechanism that gets deployed by children, especially if children are sensitive to and affected by this increase. For example, children who are asked to reason about two candidate causes need only to represent and choose among *four* candidate causal hypotheses. Four candidate causal hypotheses may well be within the information-processing capacities of 5- and 6-year-olds. In contrast, children who are asked to reason about three candidate causes must now consider *eight* candidate causal hypotheses—this may exceed their restricted information-processing capacities.

It turns out that there is a wealth of data that is consistent with this general proposal (Doebel & Zelazo, 2015; Frye, Zelazo, & Palfai, 1995; Zelazo, Frye, & Rapus, 1996; Zelazo et al., 2003). One 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 working memory. In a similar vein, Richland, Morrison, and Holyoak (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. This may provide a developmental explanation for why children in the present study did not engage in backwards blocking reasoning or show evidence that they relied on Bayesian inference. 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.

A second potential criticism is that we cannot be sure that a simple Bayesian-inference mechanism underpinned participants’ performance in the present study. For example, if participants assumed that blickets were common in the present context—which is plausible given how frequently the detector activated in the present study—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 could explain participants’ performance in the backwards blocking control condition—in that condition, participants treated all objects equally. However, this explanation cannot explain *all* the present data. This is because this explanation predicts that participants should have also treated objects A-C equivalently in the backwards blocking experimental condition as well, but this was not the case: Participants treated object A differently than either objects B or C in the backwards blocking experimental condition. This explanation is also unlikely given that, over all the Bayesian models provided a poorer fit to the behavioral data than the connectionist models. This would not be expected if participants relied on Bayesian inference. Nonetheless, because we did not systematically manipulate base-rate information, this alternative explanation cannot be ruled out entirely. However, if we are correct that participants do not rely on Bayesian inference when asked to reason about multiple causes, we predict that their performance in this proposed future study would not differ from participants’ performance in the current study. However, if children’s causal judgements are affected by base-rate information, such that, overall, how they process backwards blocking event changes with changes to base-rate information, then this would suggest that participants may use Bayesian inference after all to reason about multiple candidate cause, at least when a Bayesian-inference mechanism is primed by explicitly and systematically manipulating base-rate information.

**Conclusion**

These potential criticisms notwithstanding, this study constitute one of the first systematic attempts to examine backwards blocking and indirect screening-off reasoning in human children in the context of three and four objects. A longstanding view has been that the cognitive mechanism by which human beings reason about causal events is Bayesian inference (e.g., Gopnik et al., 2004) rather than associative processes. The experiments reported here support a different conclusion: associative learning is sufficient to explain 5- to 6-year-old children’s reasoning about multiple potential causes.

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.

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

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.

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

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.

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.

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

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.

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.

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.

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.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart J. L. McClelland, & the PDP Research Group (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

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., & Munro, S. (2006, August). When Mr. Blicket wants it, children are Bayesian. In *Proceedings of the Cognitive Science Society* (pp. 810-816).

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.

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

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.

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.