Don’t throw the (associative-learning) baby out with the bathwater just yet: Backwards-blocking reasoning with *multiple* potential causes in human children

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

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

Abstract

Causal reasoning is a fundamental cognitive ability that enables humans to learn about the complex interactions in the world around them. However, the available evidence suggests that the mechanism or set of mechanisms that underpin causal reasoning are not well understood. It is unclear, for example, whether causal reasoning is underpinned by a Bayesian mechanism, an associative mechanism, or both. Some theorists have argued that a Bayesian mechanism underpins causal reasoning because it can better account for backward-blocking (BB) and indirect screening-off (IS) findings in children and adults (e.g., Sobel, Tenenbaum, & Gopnik, 2004). However, the evidence is mixed about the extent to which learners engage in both kinds of reasoning. Here, we report three experiments that examine to what extent adults engage in BB and IS reasoning using the blicket-detector design (e.g., Gopnik et al., 2001), what mechanism best explains their behavior in this task, and under what conditions are adults’ causal ratings consistent with the predictions of the three competing computational and analytical models. The results of Experiment 1 revealed that adults’ causal ratings in the backwards-blocking condition (as well as in the indirect screening-off condition) were consistent with the predictions of the traditional and modified Rescorla-Wagner models when asked to reason about two objects. The results of the present study suggest that adults use associative processes to reason about two objects but a Bayesian-inference-like process to reason about three or more objects.

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

There is perhaps no ability that is more central for learning about how the world works than causal reasoning or the capacity to reason about cause-and-effect relations. This is a key cognitive ability because it enables human learners to encode causal relations to inform prediction and inference (e.g., Oakes & Cohen, 1990; Rakison, Smith, & Ali, 2016; Schlottmann & Shanks, 1992), to intervene on those relations to generate new effects (e.g., Gopnik et al., 2001), and counterfactually to reason about causal events to determine what would have happened if the chosen intervention had not been undertaken (e.g., Harris, German, & Mills, 1996; Sobel, 2004).

Despite consensus among researchers about the importance of causal reasoning, there is much less consensus among theorists about the cognitive mechanism that underlies this capacity. For example, it is unresolved whether domain-general mechanisms such as associative learning underpins causal reasoning or whether—as has recently been suggested by some theorists (e.g., Gopnik et al., 2004; Walker, Lombrozo, Williams, Rafferty, & Gopnik, 2017)—causal reasoning is grounded in a Bayesian-inference mechanism. One empirical finding about which domain-specific and domain-general theorists have disagreed considerably concerns whether an associative-learning mechanism or a Bayesian-inference mechanism subserves human beings’ capacity to engage in a form of retrospective reevaluation called backwards-blocking reasoning. This form of reasoning involves learning blocking or discounting redundant causal cues when other cues are shown unambiguously and in isolation to produce effects (e.g., Blaser, Couvillon, & Bitterman, 2004; Shanks, 1985; Shanks & Dickinson, 1987; Sobel et al., 2004). The aim of the experiments reported here was twofold. First, it was designed to examine whether and to what extent human children engage in backwards-blocking reasoning in a new context. Specifically, in contrast to previous studies on backwards-blocking reasoning in human children that has tended to ask children to reason about two objects, here we examined whether children could engage in this form of reasoning when asked to reason about multiple objects. Second, this study was designed to illuminate whether an associative-learning mechanism or a Bayesian-inference mechanism underlies children’s backwards-blocking reasoning performance in the current situation.

**The** **emergence of BB reasoning**

The ability to reason about causal events is thought generally to emerge between 18 months and 5 years of age (e.g., Benton, Rakison, & Sobel, 2021; Gopnik & Sobel, 2000; Gopnik et al., 2001; Kimura & Gopnik, 2019; Meltzoff, Waismeyer, & Gopnik, 2012; Sobel & Kirkham, 2006, 2007; Sobel & Munro, 2006; Walker & Gopnik, 2014; cf. Sobel & Kirkham, 2005). Although researchers have used a variety of paradigms to examine causal reasoning in human children (for a review see Bullock, Gelman, & Baillargeon, 1982), here we focus on research that has used the blicket-detector design. We focus on this paradigm for three reasons. First, it has been used most extensively to test children’s causal-reasoning abilities as well as to assess their ability to engage in backwards-blocking reasoning. Second, variations on this design have been used to evaluate adults’ causal reasoning abilities (e.g., Griffiths et al., 2011), which may support cross-study and between-age comparisons. Third, we focus on this paradigm because the notion that human reasoners use Bayesian inference to reason about causal events was introduced within the context of the blicket-detector studies and in concert with key advances in computer science, philosophy, machine learning, and statistics (for a review, see Gopnik et al., 2004).

In studies that use this design, children are introduced to a machine called the "blicket detector" and told that it lights up and plays music when certain objects—namely, "blickets"—are placed on it but not when other objects are placed on it. Following a series of events in which the detector activates (or not), children are then asked to determine which objects are blickets and to “make the machine go” by placing the blicket on the machine. Of the findings that have been reported by researchers who have used the blicket-detector methodology, perhaps none have been more controversial than that by Sobel, Tenenbaum, and Gopnik (2004). They showed that 4-year-old children—and to a lesser extent 3-year-old children—can engage in BB reasoning and IS reasoning. Children were shown initially that two novel objects A and B together caused the detector to activate and then that object A alone either failed to activate the detector (i.e., AB+ A-; IS condition) or activated the detector when placed on it (i.e., AB+, A+; BB condition). 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. It is worth noting that the BB condition is so called because after observing that A alone can activate the detector, children who engage in this form of reasoning are thought to disregard or block retrospectively object B as a potential cause because A was shown unequivocally to produce the effect. In contrast, the ISO condition is so called because B is assumed indirectly to "screen off" or to block object A as a potential cause given that A alone failed to activate the machine.

Sobel et al. (2004) found that when children were subsequently asked to make the machine go, the 4-year-olds during the ISO trial—and to a lesser extent the 3-year-olds during the same trial—responded by placing object B on the machine. In contrast, during the BB trial these same children responded by placing object A on the machine. Subsequent research by Sobel and Munro (2009) found that 3-year-olds, like the 4-year-olds in Sobel et al. (2004), could engage in BB and ISO reasoning if the activation of the detector represented desires rather than a physical effect: the 3-year-olds categorized object B as a blicket in the ISO condition but were less likely to do so in the BB condition but only when the machine was called “Mr. Blicket” and said to like blicket objects. These findings have since been interpreted not only as evidence that human children can engage in backwards-blocking reasoning but as evidence that this form of reasoning is underpinned by a Bayesian-inference mechanism rather than by an associative-learning mechanism. The crux of the Bayesian-inference account is that human learners use a simple form of Bayes’ rule to reason about causal events and to choose the causal hypothesis—within a space of hypotheses that is potentially super-exponentially large—that is most consistent with the observed data (e.g., Sobel et al., 2004; Gopnik & Wellman, 2012). More specifically, this process involves combining prior beliefs about each hypothesis with observed data to update the (posterior) probabilities of each of the hypotheses in the psychological hypothesis space.

One specific kind of associative-learning model that has received some criticism in the developmental causal literature is the traditional Rescorla-Wagner (henceforth, RW) model (e.g., Rescorla & Wagner, 1972; Griffiths et al., 2011; Sobel et al., 2004). The previous findings challenge the RW model for three key reasons. First, this model predicts that B should be treated equivalently across the BB and ISO conditions, which is a prediction that is at odds with participants’ actual treatment of object B across these conditions. The reason the RW model predicts that participants should treat B equivalently across the BB and ISO trials is because the association between object B and the outcome was identical across both conditions; that is, B was shown to produce the effect (in combination with object A) twice in both conditions. This model also only makes weighted adjustments to cues that are present, which B was not during the "A" phases in both the BB and ISO conditions. This means that because object B is absent during the A phases of the BB and ISO tasks, the RW model predicts that the associative strength between object B and the blicket effect should remain unchanged across the experimental trials in both conditions, and thus further predicts that participants should treat B equivalently across both conditions. Second, the RW model requires many learning trials for reliable associations to be established (assuming modestly set values for the salience parameters) and used to make causal inferences. In contrast, in the studies cited above participants engaged in BB (and ISO) reasoning based on only a handful of learning trials. Finally, the BB and ISO findings challenge the RW model because this model does not naturally encode base rates to which children have been shown to be sensitive.

Despite these valid criticisms, caution should be exercised either before accepting these criticisms or arguments that stipulate that children use Bayesian inference to reason causally. One reason to exercise caution is because there are problems with Sobel et al.’s (2004) operationalization of BB reasoning (although for alternative operationalizations see De Houwer, Beckers, & Glautier, 2002; Larkin, Aitken, & Dickinson, 1998; Griffiths et al., 2011; Kruschke & Blair, 2000; Lovibond et al., 2003; Shanks, 1985; Van Hamme and Wasserman, 1994). These authors operationally defined BB reasoning as greater B choices in the ISO condition than in the BB condition. This way of operationally defining BB reasoning was presumably motivated by two key factors. First, if the causal status of object A—which can be determined unequivocally when object A is placed alone on the machine—causes participants retrospectively to reevaluate the causal status of object B, then participants should consider B to be less of a blicket in the BB condition (and thus retrospectively “block” it) than in the ISO condition. This is because A by itself fails to produce the effect in the ISO condition but produces the effect by itself in the BB condition. Second, proponents of this operationalization of BB reasoning have used the fact that participants do treat object B differently between the BB and ISO conditions as evidence against rudimentary associative-learning models such as the RW model given that this model predicts equivalent treatment of object B across the BB and ISO conditions.

However, this operationalization of BB reasoning has a notable shortcoming. Specifically, by operationalizing BB in terms of the difference in treatment of object B across the BB and ISO conditions, it is logically possible that participants treated object B differently between the BB and ISO conditions because they observed a positive effect during the elemental (i.e., A+) phase in the BB condition but a negative effect during the elemental (i.e., A-) phase in the ISO condition. This would mean that participants’ differential treatment of object B across the two conditions could have resulted from the fact that the two conditions differed in terms of their low-level perceptual features rather than from a true retrospective reevaluation of object B by participants based on A’s *relation to and effect* *on* object B across both conditions.

Given this limitation, we argue that a more (construct) valid operationalization of BB reasoning is to compare the treatment of object B following an AB+ A+ sequence of events (i.e., the BB experimental condition) to the treatment of B following an AB+ C+ sequences of events (i.e., the BB control condition). These two conditions differ in terms of the object that is shown during the elemental phase (i.e., A or C) and that object’s *relation* to B (and thereby the potential impact that this object has on how B is treated). For example, in the BB experimental condition, a dependency is presumably established between objects A and B because both objects appear together during the compound phase of the condition. This means that A’s causal status that is established during the subsequent elemental phase *should* affect participants’ (retrospective) treatment of B; that is, whether object A is shown to activate the machine should affect how participants treat object B. In contrast, in the BB control condition, object C never appeared with object B, which necessarily means that C’s causal status should not (retrospectively) impact how participants treat object B. Crucially, the blicket effect itself is held constant across the BB experimental and control conditions such that participants observe blicket-detector activation in both cases. If participants engage in BB reasoning in this context, then this would provide stronger evidence that participants have access to such a mechanism. In particular, if BB reasoning is treated as an indirect measure of the operation of a Bayesian-inference mechanism as has typically been the case (e.g., Griffiths et al., 2011; Sobel & Kirkham, 2006; Sobel et al., 2004), then demonstrating that participants treat object B differently across the BB experimental and control conditions would suggest that participants have access to and use Bayesian inference to reason causally.

**An open question**

A second reason to exercise caution before accepting the claim that human beings use Bayesian inference to reason about causal events is that it is not known whether human children engage in BB reasoning for three (or more) objects. The is because most, if not all, of the studies on BB reasoning in human children have tended to use two objects; that is, participants are shown an AB+ A+ sequence of events and then asked whether each object is a blicket. This research is important because it has revealed that BB reasoning may emerge by 3 years of age, but it leaves unaddressed whether children can engage in BB reasoning when asked to reason about three or more objects. It also remains unknown whether participants engage in BB reasoning when the elemental phase (i.e., the A+ phase in the BB condition or the A- phase in the ISO condition) consists of two rather than one object. These are important questions to answer because if a Bayesian-inference mechanism is assumed to underpin human causal reasoning—and it is further assumed that BB reasoning is an indirect measure of the operation of such a mechanism—then it is crucial to show that participants continue to engage in BB reasoning (and thus make use of Bayesian inference) even when they are asked to reason about three (or more) objects or even when the elemental phase consists of two rather than one object. In other words, if one of the goals of the larger research community is to elucidate the cognitive mechanisms by which human children reason about causality *in the real world*, then it is imperative to understand better how causal reasoning unfolds in situations that more closely approximate those that may be found the real world such as ones in which children must reason about more than two objects.

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 really is meaningful. This is because these two settings differ by one (or at most, by two) candidate causes. However, if Bayesian inference is the cognitive mechanism that underpins human causal reasoning, then the difference between these two settings is far from trivial. This is because in the two-candidate-cause setting, participants need only to determine which of *four* candidate causal hypotheses generates the observed data. However, in the three- or even four-candidate-cause setting, participants need to determine which of *eight* (in the case of 3 candidate causes) or *sixteen* (in the case of 4 candidate causes) hypotheses is the one that generates the observed data. Thus, in the four-candidate-cause setting, participants must consider four times as many causal hypotheses as participants in the two-candidate-cause setting, which is far from a trivial difference.

Crucially, this difference may have important implications for whether an associative-learning mechanism or a Bayesian-inference mechanism underlies causal reasoning in children. For instance, it is possible that when children’s information-processing abilities are taxed—such as when they are asked to reason about three (or more) objects—they may resort to simpler modes of causal reasoning such as reasoning that is consistent with the predictions of the traditional RW model. This perspective is consistent with a view that was put forward by Cohen and colleagues (Cohen, 1998; Cohen & Cashon, 2001; Cohen, Chaput, & Cashon, 2002; Oakes & Cohen, 1990; see also Oakes, 1994). The crux of this perspective is that there is a bias for children to process information at the highest level (and perhaps in terms of the most sophisticated available cognitive mechanisms and processes). However, if the task that children face requires information-processing abilities that extend beyond what they possess, then there will be a tendency for them to lower levels and less sophisticated cognitive mechanisms.

Thus, if participants’ BB performance adheres to the predictions of the traditional RW model or even to the predictions of even simpler models of associative learning (see below) in a multiple-candidate-cause setting—which, here, would be in evidence if participants treated B equivalently across the BB experimental and control conditions regardless of the number of objects shown during the elemental phase—this would suggest that it may have been premature to conclude that the traditional RW model, on the one hand, or other associative processes, on the other hand, are inadequate models of causal reasoning in human children. This would also support the contention that there is a tendency for children to use simpler cognitive mechanisms and processes to reason about causal events when their information-processing abilities are stretched. Thus, by understanding whether participants engage in BB reasoning in a multiple-candidate-cause setting, we can gain greater insight into *how* children reason about causal events and under what conditions they use one kind of cognitive mechanism in lieu of another.

**Possible cognitive mechanisms underlying BB reasoning for multiple potential causes**

Given that the goal of the present paper is to determine what underlying cognitive mechanism best accounts for participants' performance in the present paper, it was important to determine what predictions a Bayesian-inference mechanism and an associative-learning mechanism—which was instantiated in a simple connectionist computational model that was trained with a variant of the traditional RW model—makes. Below we also outline the predictions of a second counting-based associative-learning mechanism. It should be noted that here we only describe these cognitive mechanisms at a high level and show, graphically, their predictions. However, we direct the reader to the Appendix for the formal details of the Bayesian model and the details of the connectionist simulation.

**Bayesian inference.** Proponents of the Bayesian-inference account maintain that human learners use a simple form of Bayes’ rule to reason about causal events. Specifically, the crux of this perspective is that learners’ responsibility is to determine which hypothesis—within a space that contains potentially an infinite number of psychological hypotheses—is that one that is producing some observed data. The proposed cognitive mechanism by which learners determine such a “winning” hypothesis is by combining their beliefs about the likelihood of each hypothesis (in the absence of data) with whether the observed data is likely to have been produced by the hypothesis that is currently under consideration. Crucially, learners will retain a hypothesis to the extent that the data that is being observed can be produced by that hypothesis. Learners are said to discard a hypothesis when it no longer can produce the data. A core assumption of Bayesian inference is that the probability of a given hypothesis (given some data; i.e., the “posterior probability” of a given hypothesis) at one time point becomes the prior probability of that hypothesis at the next point in time, and the process that was just described for how one isolates the winning hypothesis repeats when new data is encountered. Crucially, once the probabilities of each hypothesis (given some data) have been updated, learners may use them to update their beliefs about the likelihood that particular objects, rather than particular hypotheses, are blickets.

Thus, given that learners are asked to reason about a maximum of three candidate causes (i.e., objects A-C) during the experimental trials in the both the BB and ISO conditions and a maximum of four candidate causes during the control trials in both the BB and ISO conditions (i.e., objects A-D), the corresponding psychological hypothesis spaces consist, respective, of 8 and 16 total hypotheses (Figs. 1 & 2).

**Timeline

Description automatically generated with medium confidence**

Figure 1. 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.

By application of Bayes’ rule, the prediction that this perspective makes for how participants should treat the objects after the BB main trial is shown below (see the Appendix for more details on how these predictions were derived).

|  |  |  |
| --- | --- | --- |
| **Bayesian Model Predictions (3 objects) – BB main** | | |
| **Graphs** | **Prior (*p*(*h*))** | **After AB+ A+ (*p*(*h|d*))** |
| Model 0 | *p*3 | *p*2 |
| Model 1 | *p*2(1-*p)* | *p*(1-*p*) |
| Model 2 | *p*2(1-*p)* | *p*(1-*p*) |
| Model 3 | *p*(1-*p*)2 | (1-*p*)2 |
| Model 4 | *p*2(1-*p*) | 0 |
| Model 5 | *p*(1-*p*)2 | 0 |
| Model 6 | *p*(1-*p*)2 | 0 |
| Model 7 | (1-*p*)3 | 0 |
| Probabilities of objects A, B, and C | | |
| Object A | *p* | 1 |
| Object B | *p* | p |
| Object C | *p* | *p* |

As can be seen, this model predicts that for any probability *p,* following the AB+ A+ BB main event participants should be maximally confident that object A is a blicket but should treat objects B and C equivalently. The predictions that this model makes after the ISO main trial are shown below.

|  |  |  |
| --- | --- | --- |
| **Bayesian Model Predictions (3 objects) – ISO main** | | |
| **Graphs** | **Prior (*p*(*h*))** | **After AB+ A- (*p*(*h|d*))** |
| Model 0 | *p*3 | 0 |
| Model 1 | *p*2(1-*p)* | 0 |
| Model 2 | *p*2(1-*p)* | 0 |
| Model 3 | *p*(1-*p*)2 | 0 |
| Model 4 | *p*2(1-*p*) | *p*2 |
| Model 5 | *p*(1-*p*)2 | *p*(1-*p*) |
| Model 6 | *p*(1-*p*)2 | *p*(1-*p*) |
| Model 7 | (1-*p*)3 | (1-*p*)2 |
| Probabilities of objects A, B, and C | | |
| Object A | *p* | 0 |
| Object B | *p* | *p* |
| Object C | *p* | *p* |

As is shown here, this model predicts that for any probability *p*, following the ABC+ A- ISO main trials participants should be maximally confident that object A is *not* a blicket but treat objects B and C equivalently. In terms of the control trials for the BB and ISO condition, the hypothesis space—which consists of 16 hypotheses—is shown below.

Timeline

Description automatically generated

Figure 2. The sixteen different causal hypotheses indicating the possible causal relations for a causal event that involves three objects and one blicket detector. *A*, *B*, *C*, and D correspond to the four objects that were used on the machine and *E* indicates the activation of the machine.

The predictions that this model makes for the BB control trials—that is, for the ABC+ D+ trials—are shown below.

|  |  |  |
| --- | --- | --- |
| **Bayesian Model Predictions (4 objects) – BB control** | | |
| **Graphs** | **Prior (*p*(*h*))** | **After ABC+ D+ (*p*(*h|d*))** |
| Model 0 | *p*4 | *p*3 |
| Model 1 | *p*3(1-*p*) | 0 |
| Model 2 | *p*3(1-*p*) | *p*2(1-*p*) |
| Model 3 | *p*2(1-*p*)2 | 0 |
| Model 4 | *p*3(1-*p*) | *p*2(1-*p*) |
| Model 5 | *p*2(1-*p*)2 | 0 |
| Model 6 | *p*2(1-*p*)2 | *p*(1-*p*)2 |
| Model 7 | *p*(1-*p*)3 | 0 |
| Model 8 | *p*3(1-*p*) | *p*2(1-*p*) |
| Model 9 | *p*2(1-*p*)2 | 0 |
| Model 10 | *p*2(1-*p*)2 | *p*(1-*p*)2 |
| Model 11 | *p*(1-*p*)3 | 0 |
| Model 12 | *p*2(1-*p*)2 | *p*(1-*p*)2 |
| Model 13 | *p*(1-*p*)3 | 0 |
| Model 14 | *p*(1-*p*)3 | (1-*p*)3 |
| Model 15 | (1-*p*)4 | 0 |
| Probabilities of objects A, B, C, and D | | |
| Object A | *p* | *p* |
| Object B | *p* | *p* |
| Object C | *p* | *p* |
| Object D | *p* | 1 |

As is evident, the model predicts that for any probability *p*, following the ABC+ D+ BB control trials participants should be maximally confident that object D is a blicket after the BB control trials but should treat objects A-C equivalently. Finally, the predictions that the model makes for the ISO control trials—that is, for the ABC+ D- trials—are shown below.

|  |  |  |
| --- | --- | --- |
| **Bayesian Model Predictions (4 objects) – ISO control** | | |
| **Graphs** | **Prior (*p*(*h*))** | **After ABC+ D- (*p*(*h|d*))** |
| Model 0 | *p*4 | 0 |
| Model 1 | *p*3(1-*p*) | *p*3 |
| Model 2 | *p*3(1-*p*) | 0 |
| Model 3 | *p*2(1-*p*)2 | *p*2(1-*p*) |
| Model 4 | *p*3(1-*p*) | 0 |
| Model 5 | *p*2(1-*p*)2 | *p*2(1-*p*) |
| Model 6 | *p*2(1-*p*)2 | 0 |
| Model 7 | *p*(1-*p*)3 | *p*(1-*p*)2 |
| Model 8 | *p*3(1-*p*) | 0 |
| Model 9 | *p*2(1-*p*)2 | *p*2(1-*p*) |
| Model 10 | *p*2(1-*p*)2 | 0 |
| Model 11 | *p*(1-*p*)3 | *p*(1-*p*)2 |
| Model 12 | *p*2(1-*p*)2 | 0 |
| Model 13 | *p*(1-*p*)3 | *p*(1-*p*)2 |
| Model 14 | *p*(1-*p*)3 | 0 |
| Model 15 | (1-*p*)4 | (1-*p*)3 |
| Probabilities of objects A, B, C, and D | | |
| Object A | *p* | *p* |
| Object B | *p* | *p* |
| Object C | *p* | *p* |
| Object D | *p* | 0 |

Finally, as can be seen above, the model predicts that for any probability *p*, following the ABC+ D- ISO control trials participants should be maximally confident that object D is *not* a blicket but should treat objects A-C equivalently. Taken together, a simple Bayesian model predicts that learners should be maximally confident about the status of a candidate cause when it is shown in isolation, but should treat objects that are shown in combination equivalently.

**Associative learning: the traditional RW model.** In contrast to a simple Bayesian model, learners who rely on a cognitive mechanism that can be captured by the traditional RW model should adjust their beliefs about the status of a candidate cause by an amount that is equal to the difference between the outcome that participants observe (i.e., whether or not the blicket machine activates) and participants’ beliefs about whether a given candidate cause will activate the machine. On this account, this difference is “scaled” by the salience of the effect (i.e., how causal the effect appears) and the salience of the candidate causes. Crucially, unlike the Bayesian inference perspective, learners who use this mechanism to reason about causal events need not represent a psychological hypothesis space. Rather, learners’ inferences are based on prediction error between what they observe and what they expect. The predictions that this model makes for how participants should treat the objects after the BB main and control trials and the ISO main and control trials are shown below.

|  |  |  |
| --- | --- | --- |
| **The traditional RW model (3 objects) – BB main** | | |
| A | *p* | 1 |
| B | *p* | *p* |
| C | *p* | *p* |
| **The traditional RW model (3 objects) – ISO main** | | |
| A | *p* | 0 |
| B | *p* | *p* |
| C | *p* | *p* |
| **The traditional RW model (4 objects) – BB control** | | |
| A | *p* | *p* |
| B | *p* | *p* |
| C | *p* | *p* |
| D | *p* | 1 |
| **The traditional RW model (4 objects) – ISO control** | | |
| A | *p* | *p* |
| B | *p* | *p* |
| C | *p* | *p* |
| D | *p* | 0 |

As is shown above, this model, like the simple Bayesian model, predicts that learners should be maximally confident about the status of a candidate cause when it is shown in isolation, but should treat objects that are shown in combination equivalently. Thus, if participants’ performance align with the predictions of these two models, it should be impossible to determine whether a Bayesian mechanism or an associative-learning mechanism—based on the mechanics of the traditional RW model—underlies children’s performance in this task.

**Associative learning: a simple “counting” cognitive mechanism.** A third potential cognitive mechanism that children may rely on to reason about the present causal events is based on a simple associative “counting” strategy based on the frequency with which (i.e., the number of times that) a given object—either individually or in combination with other objects—was seen with the machine’s activation or inactivation. To understand how this cognitive mechanism works to produce causal judgements, consider the ABC+ D+ BB control trial. If the cognitive mechanism that best explains learners’ inferences in the present task is counting, following the BB control event learners should treat objects A-D equivalently. This is because all four objects would have been paired with the machine’s activation an equal number of times. In other words, object A would have been seen with the machine’s activation exactly once, object B would have been seen with the machine’s activation exactly once, object C would have been seen with the machine’s activation exactly once, and object D would have been seen with the machine’s activation exactly once. As such, the strength of the associative relation (or link) between each object and the effect should be equivalently for all four objects. It should be noted that although children observe trials during which objects are placed on the machine in triplets, learners who rely on a counting mechanism are nonetheless considering each object separately and are counting the frequency with which a given object is paired with the machine’s activation.

It should be clear that although a counting-based mechanism and the traditional RW model are both instances of associative-learning processes, they differ in their specifics. For instance, learners who rely on the traditional RW model to make causal inferences should adjust their beliefs about the causal status of an object based on prediction error. In contrast, if a simple counting-based mechanism underlies learners’ causal inferences in the present context, then they should increase their belief that a given object is causally effective based on the number of pairings of the object and the machine’s activation. Below are the predictions that this account makes for all four conditions.

|  |  |  |  |
| --- | --- | --- | --- |
| The predictions of a simple associative-based counting mechanism | | | |
| BB main – ABC+ A+ | | | |
| Is A a blicket? | Is B a blicket? | Is C a blicket? | Is D a blicket? |
| +2 | +1 | +1 | N/A |
| BB control – ABC+ D+ | | | |
| Is A a blicket? | Is B a blicket? | Is C a blicket? | Is D a blicket? |
| +1 | +1 | +1 | +1 |
| ISO main – ABC+ A- | | | |
| Is A a blicket? | Is B a blicket? | Is C a blicket? | Is D a blicket? |
| 0 | +1 | +1 | N/A |
| ISO control – ABC+ D- | | | |
| Is A a blicket? | Is B a blicket? | Is C a blicket? | Is D a blicket? |
| +1 | +1 | +1 | 0 |

It should be clear from the table above that this account predicts that for the BB main trials children should say that object A is a blicket significantly more often than either B or C but that their treatment of B and C should not differ. This is because objects B and C would have been paired with the machine exactly once, whereas object A would have been paired with the machine twice. Similarly, this account predicts that during the BB control trials participants’ treatment of all four objects should not differ. This prediction results from the fact that all four objects would have been paired with the machine’s activation exactly once. In contrast, this account predicts that participants should not consider object A to be a blicket but should be split in their treatment of objects B and C because B and C would have been paired with the machine’s activation an equal number of times; in other words, the effect of seeing object A paired with the machine’s activation during the ABC+ trials should be exactly canceled out by the effect of seeing object A paired with the machine’s inactivation during the A- trials. Likewise, during the ISO control trials, this account predicts that participants should not consider object D to be a blicket but should be split in their treatment of objects A, B, and C.

**The present investigation**

The present investigation had three goals. The first goal was to determine whether 4-, 5-, and 6-year-olds could engage in BB reasoning when asked to reason about three and four objects. A second goal was to determine whether participants show evidence of BB reasoning when it is operationally defined as greater treatment of object B in the BB control condition compared to the BB experimental condition. A third goal was to gain greater insight into how—that is, by what cognitive mechanism—children reasoned about the present events. Specifically, a goal of this paper was to determine which of the three cognitive mechanisms mentioned above (and outlined more fully in the appendix) accounts for children’s causal inferences in the present context. Finally, given that some previous research operationally defined BB reasoning as greater treatment of object B in the BB condition compared to the ISO condition, participants in the present series of experiments also experienced the ISO condition (in a between-subjects manipulation).

**Experiment 1**

Experiment 1 assessed 4-year-old children’s ability to engage in BB when asked to reason about three objects. Participants were introduced to a computer-animated machine called the “blicket detector” and were told that their task was to determine which objects make the machine activate—and thus represent blickets—and which objects do not make the machine activate. Following this brief introduction phase, participants received two backwards-blocking trials and two backwards-blocking control trials and were asked to indicate whether the objects in each trial were blickets. In this experiment, only a single object was shown during the elemental portion of the BB experimental event.

**Materials.** The “device” used in the experiments presented here was a computer-animated version of the blicket detector. 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 became ocean blue. If the device was “off”, the white region remained white. In addition, a maximum of 4 differently colored circles were used, and each circle measured 2.67 cm × 2.67 cm (INCLUDE FIGURE OF MACHINE AND TOYS). The machine was designed such that it activated immediately when a circle that was predetermined to be a blicket contacted it. At the start of any given trial, three (for the BB experimental trials) or four equally spaced (for the BB control trials) circles appeared above the blicket machine. Finally, the videos contained a built-in script, which experimenters were instructed to read to ensure that all participants were given exactly same instructions and received the same text throughout the experiment (INCLUDE FIGURE OF MACHINE AND TOYS). All video events were created in Microsoft PowerPoint.

**Procedure.** Participants were either tested in a quiet room on campus or in quiet rooms in local children’s science museums. 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 that were used during the main portion of the experiment. The pretraining phase began with the triangle (object A) and pentagon (object B), which were located side-by-side and above the machine. Object A then descended until it contacted and immediately activated the machine (i.e., the white region changed from white to ocean blue). Object A then ascended until it 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 the machine, which immediately activated (ostensibly because object A contacted it). Participants were then asked whether each object was a blicket. This event was identical to the “one-cause” event in Gopnik, Sobel, Schulz, and Glymour (2001) and was included to ensure that participants could reason about blicket objects.

Following the pretraining phase, participants were given four test trials—two BB experimental trials and 2 BB control trials—in counterbalanced order using a Latin square. It should be noted that differently colored objects were used across all trials. This meant that no two events overlapped in the colors (for the objects) that were used.

The two BB 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 ascended to their starting positions. The left- or right-most (counterbalanced) object (i.e., 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; that is, the text, “Is this one a blicket?” with a downward-facing arrow then appeared above each object, and participants were asked whether each object was a blicket. The first and second BB experimental trials were identical except that different colors were used for the objects.

The two BB 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 then descended by itself until it contacted and activated the machine. Children were then asked whether each object was a blicket. Note that object A descended with the remaining two objects in the BB experimental trials, whereas object D did not descend with the remaining three objects in the BB control trials. This means that D’s causal status should have no bearing on participants’ treatment of objects A-C. Note also that the BB control trials used the same text as the BB experimental trials. The first and second BB control trials were identical except that different colors were used for the objects.

Finally, the ISO experimental and control conditions were identical to the BB experimental and control conditions except that objects A and D failed to activate the machine in the ISO experimental and control trials, respectively. The schematic for this experiment is shown below in Table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Schematic of Experiment 1 | | | |
|  | Compound | Elemental | Test |
| 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 (-)

**Results**

All analyses were conducted in R (R Development Core Team, 2008). All *p* values were supplemented with Bayes Factors. *BF*H1 quantifies support for the alternative hypothesis compared with the null hypothesis. *BF*s close to 1 indicate equal support for both hypotheses, whereas *BFH0*s≥ 3 and < 10 indicate moderate and strong evidence for the alternative hypothesis, respectively (Lee & Wagenmakers, 2014).

**[Add Exp. 1 results here]**

**Discussion**

[INSERT EXPERIMENT 1 DISCUSSION HERE]

**Experiment 2**

Experiment 2 was similar to Experiment 1 except that 5- and 6-year-old children were tested.

**Method**

**Participants.** Participants were X 5-year-olds (X boys and X girls; *Mage* =X months, range = X-Y) and X 6-year-olds (X boys and X girls; *Mage* =X months, range = X-Y). Although most children were from white, middle-class backgrounds, a range of ethnicities that resembled the diversity in the population were represented. All children were tested in a quiet room at a Children’s museum in Providence, RI.

**Materials & Procedure.** The materials and procedure for Experiment 2 was identical to that for Experiment 1.

**Results**

Figure X shows the results for this experiment. The data were entered into a five-way linear model with Age (5-year-olds vs. 6-year-olds) as the between-subjects factor and Condition (BB vs. ISO), Trial Type (experimental vs. control), and Objects (A vs. B vs. C vs. D) as the within-subjects factors. This analysis revealed a main effect of Condition, *F*(1, 548) = 12.68, *p* < .001, a main effect of Objects, *F*(3, 548) = 5.70, *p* < .001, a main effect of Event Type, *F*(1, 548) = 13.05, *p* < .001, and a significant interaction between Condition and Objects, *F*(3, 548) = 9.28, *p* < .001. This significant two-way interaction was qualified by a significant three-way interaction among Condition, Event Type, and Objects, *F*(2, 548) = 9.67, *p* < .001.

We followed up this three-way interaction with separate one-way linear models for the main and control trials within the BB and ISO conditions. The Objects factor was treated as the sole within-subjects factor in these follow-up analyses. The first one-way linear model for the control trials within the BB condition did not reveal a significant effect of Objects, *F*(3, 217) = 0.63, *p* = .59. This means that participants treated the objects similarly during the control trials of the BB condition. In contrast, the second one-way linear model for the main trials within the BB condition revealed a significant main effect of Objects, *F*(2, 159) = 3.63, *p* = .03. This main effect reflected the fact that participants considered object A marginally to be more of a blicket (*M* = 1.75, *SD* = 0.64) than object B (*M* = 1.47, *SD* = 0.77), *t*(52) = 1.92, *p* = .06, and significantly more of a blicket than object C (*M* = 1.37, *SD* = 0.83), *t*(52) = 2.5, *p* = .02.

The third and fourth one-way linear models for the main and control trials within the ISO condition both revealed a significant main effect of Objects, both *F*’s > 11.43, both *p*’s < .0001. This reflected the fact that participants considered object A (*M* = 0.57, *SD* = 0.89) to be significantly less of a blicket than objects B (*M* = 1.58, *SD* = 0.78) and C (*M* = 1.62, *SD* = 0.71) during the main condition, both *t*’s > -4.14, both *p*’s < .0001, and object D (*M* = 0.83, *SD* = 0.75) to be less of a blicket than objects A (*M* = 1.73, *SD* = 0.67), B (*M* = 1.66, *SD* = 0.67), and C (*M* = 1.63, *SD* = 0.61) during the control trials, all *t*’s > -4.74, all *p*’s < .0001.

**Assessing BB under the new operationalization of BB reasoning**

To examine whether there was evidence of BB reasoning under the new operationalization—in which participants’ treatment of some of the redundant causes is compared to their treatment of other redundant causes *within* the BB condition—data for the redundant causes within the BB experimental and control conditions were entered into a two-way linear model with Objects (A, B, and C) and Trial Type (main vs. control) as the within-subjects factors. This analysis revealed only a main effect of Trial Type, *F*(1, 267) = 5.26, *p* = .02, which reflected the fact that participants’ blicket ratings were higher during the control trials (*M* = 1.67, *SD* = 0.63) than during the main trials (*M* = 1.42, *SD* = 0.80). Crucially, neither the main effect of Objects, *F*(2, 267) = 0.44, *p* = .64, nor the interaction between Objects and Trial Type*, F*(2, 267) = 0.44, *p* = .64, was significant. This reflected the fact that participants treated the redundant causes equally within the BB main and control conditions.

**Assessing BB under the old operationalization of BB reasoning**

To examine whether there was evidence of BB reasoning under the old operationalization—in which participants’ treatment of some of the redundant causes is compared to their treatment of other redundant causes *between* the BB and ISO conditions—data for the redundant causes between the BB and ISO conditions were entered into a two-way linear model with Objects (A, B, and C) and Trial Type (main vs. control) as the within-subjects factors. Similar to the analysis above, the only main effect was Trial Type, *F*(1, 400) = 4.47, *p* = .04, which reflected the fact that participants’ blicket ratings were higher during the control trials (*M* = 1.67, *SD* = 0.63) than during the main trials (*M* = 1.42, *SD* = 0.80). In contrast and crucially, neither the main effect of Objects, *F*(2, 400) = 0.53, *p* = .59, nor the interaction between Objects and Trial Type*, F*(1, 400) = 0.04, *p* = .83 was significant. These latter resulted indicated that participants treated the redundant causes equivalently between the BB and ISO main and control trials. Thus, these results indicate that participants neither engaged in BB reasoning under either the new or old operationalizations of it.

Discussion

Taken together, these results suggest that when participants are asked to reason about three and four objects—which corresponds to hypothesis spaces that consist of 8 and 16 candidate causal hypotheses, respectively—they do not engage in BB reasoning. These results are also inconsistent with the predictions of theCrucially, if BB reasoning is used as an indirect measure of the operation of a Bayesian-inference mechanism, then these findings are inconsistent with the notion that children use Bayesian inference to reason about three objects. It will be recalled that this perspective predicts that human reasoners should engage in BB reasoning irrespective of the number of candidate causes about which they are asked to reason. However, these data are consistent with the traditional RW model. As we discussed at the outset of this paper, this model does predict that participants should treat the redundant objects—that is, the objects that are not presented alone on the machine—equivalently. This is because the strength of the association between each redundant cue and the causal effect is equivalent for each cue. Crucially, these findings could not be explained by the fact that participants were insensitive to the causal status of each object and thus considered all objects to be blickets: Participants were more confident that object A was a blicket during the BB experimental trials than during the ISO experimental trials. Likewise, participants were more confident that object D was a blicket during the BB control trials than during the ISO control trials.

General Discussion

This study had three aims. The first aim was to examine whether 4-, 5-, and 6-year-olds would engage in BB reasoning when asked to reason about 3 and 4 objects. This study departs from previous research on BB reasoning in human children. In these studies, typically children were asked to reason about two potential causes (e.g., Beckers et al., 2009; Griffiths et al., 2011; McCormack et al., 2013; Sobel et al., 2004). The second aim was to determine whether participants would engage in BB reasoning either under the older operationalization of it or under the newer operationalization of it. The third aim was to clarify the debate on *how* children reason about causal events in a BB context by assessing whether their performance aligned with the predictions of a simple Bayesian model, the traditional RW model, or an associative-based counting mechanism.

In terms of the first two aims, there was no evidence that children engaged in BB reasoning when asked to reason about three or four objects. Crucially, this was true regardless of how BB reasoning was operationalized. Specifically, there was no evidence of BB reasoning when we compared participants’ treatment of the redundant causes *between* the BB and ISO conditions. Similarly, there was no evidence of BB reasoning when we compared participants’ treatment of the redundant causes *within* the BB condition itself. This finding extends previous research to show that when participants are asked to reason about three or more objects, they do not engage in BB reasoning (see below for a potential explanation for this incongruity).

In terms of the third aim, the present results neither provide support for the notion that children’s causal inferences in the present context were subserved by a Bayesian-inference mechanism nor do they provide support for the notion that associative learning—based on the traditional RW model—underpinned their performance. This is because participants’ treatment of the objects did not conform with the full set of model predictions for either model. For example, both the traditional RW model and the simple Bayesian model predict that participants should be maximally confident that object D is a blicket after the BB control trials but should treat objects A-C equivalently. However, this is not what the data showed. Instead, participants treated all four objects equivalently during the BB main trials. Children’s performance was consistent with the predictions of both models for the remaining trials.

Although children’s performance neither aligned with the predictions of a simple Bayesian model or the predictions of the traditional RW model, these findings are entirely compatible with the idea that children’s causal performance in the present context originated from the operation of a simple associative-based counting mechanism. In other words, these results suggest that children’s willingness to say that an object was a blicket depended on the frequency with which that object was paired with the machine’s activation; the more frequently that an object is paired with the machine’s activation, the more likely children should be to say that the object is a blicket, especially in relation to objects that are paired with the machine’s activation less often.

One potential criticism of this study is that it should be interpreted with caution because the results are inconsistent with the findings from previous studies on BB reasoning in human children. Such previous research showed that children do engage in BB reasoning when asked to reason about two objects; the current study showed that children do not engage in BB reasoning when asked to reason about three objects. However, we believe that the present results are entirely compatible rather than at odds with this previous research.

The present findings suggest that when children’s information-processing capacities are stretched, they may deploy simpler associative mechanisms in causal contexts like the present one. Indeed, although at the level of individual objects the difference between three and four objects is miniscule, by contrast the corresponding increase in the underlying psychological hypothesis space is substantial, and this increase may have important ramifications on the cognitive mechanism that gets deployed by children. For example, children who are asked to reason about two candidate causes—which is the approach that has been taken in most contemporary studies on BB reasoning in human children (e.g., Beckers et al., 2009; Griffiths et al., 2011; Kloos & Sloutsky, 2013; McCormack et al., 2009; McCormack et al., 2013; Sobel & Kirkham. 2006; Sobel et al., 2004)—need only to represent and choose among *four* candidate causal hypotheses (i.e., 2n, where *n* is the number of potential causes). This may be within the information-processing capacities of 3- and 4-year-olds.

In contrast, children who are asked to reason about three or four candidate causes must now contend with *eight* or *sixteen* candidate causal hypotheses, respectively. For the developing child, this may well be outside the limits of their restricted information-processing capacities. A recent study by Kenderia and Kibbe (2023) provided data that were in line with this general thesis. Eight- and 10-year-old children were told that they would play a memory game that involved clicking on a single square in a 3 x 3 array of squares to reveal the identity of a virtual card hidden at that location. They were then told their task was to find three objects that shared some number of features. For example, they might be told to “find three objects that share two features”., Kenderia and Kibbe (2023) found that when children’s working-memory abilities were taxed—such as when children were asked to find three objects that shared one feature and differed on another feature—they relied less on working memory and manual exploration, which ostensibly is less cognitively effortful. It should be noted that finding—in which children use less cognitively effortful strategies when their information-processing abilities are taxed—is consistent with has been replicated —that is, that general showed that in a task in which children were asked to find matching objects in an array of objects that varied in a number of surface features, children relied less on working memory when the task complexity increased. For example, when the complexity of a task is increased, children rely less on

This proposal is broadly consistent ewith the Information-Processing view that was put forward by Leslie Cohen and others (Cohen, 1998; Cohen, Chaput, & Cashon, 2002). On this view, on any given task, infants and children are biased to process information at the highest possible level. However, when the information that infants or children encounter in a task exceeds their processing abilities, there will be a tendency for children to process information at the next highest level. This may well explain why, in the present context, children used a counting strategy rather than one based on Bayesian inference or the traditional RW model to make causal inferences.

A second potential criticism is that we have not expliclty shown that participants are sensitive to base rates. In other words, an open question that this study leaves unaddressed concerns what effect, if any, establishing the base rate of blickets would have on participants’ BB performance in this setting. For example, it is possible that participants would engage in BB reasoning—and thus show evidence of the use of a Bayesian-inference mechanism—if the base rate of blickets is established to be low. In contrast, if the base-rate of blickets is established to be high, it is possible that participants’ performance would mirror those of participants in the current study. Such a study, in combination with the results of the present study, would clarify what base rate, if any, participants default to when base rate is not explicitly manipulated. Although previous research has shown that children are sensitive to base rates and can integrate that information into their causal judgements about two potential causes (e.g., Griffiths et al., 2011; Sobel et al., 2004), it remains unknown whether participants would be sensitive to base-rate information in the present context. Although we have not expliclty manipulated base rates here, we can speculate on the effects of such manipulations on children’s causal inferences.

Nonetheless, by examining whether participants are sensitive to base rate information in a context with multiple candidate causes, we can provide still further insight into the underlying causal mechanism that supports causal judgements in human children. For instance, if children’s causal judgements are shown to be affected by base-rate information, such that their BB reasoning performance changes as a function of changes to the base rates of blickets, then this would suggest that participants may use Bayesian inference to reason about multiple candidate cause after all. This is because Bayesian inference requires that learners combine the current data with our prior beliefs about how likely a given object is to be blicket to choose the causal hypothesis that is generating the data. Thus, participants who are insensitive to the base-rates of blickets cannot be said to be using Bayesian inference. Crucially, if participants continued not to engage in BB reasoning despite manipulations to the base rate of blickets—as evidenced by equivalent treatment of the redundant candidate causes within and between conditions—then this would further suggest that associative learning provides a better account of causal reasoning in human children.

Along these lines, one potential criticism of the present study is that it cannot be ruled out that participants were relying on Bayesian inference. For example, if participants assumed a priori that blickets were common in the present context—which is plausible given that the detector activated much more frequently in the present study than, say, in Sobel et al. (2004)—then participants should be less likely to block redundant causes. We are disinclined to accept this explanation for two reasons. First, the performance of the 4-year-olds and the 5- and 6-year-olds was equivalent. If possessing sufficient information-processing capacities and showing sensitivity to base-rate information are important prerequisites for using Bayesian inference, then the 4- and 5-year-olds might be expected to perform differently than the 6-year-olds. The results from Experiment 2 seem to support this supposition: The 5-year-olds’ performance suggested that they were more confident that object A was a blicket than the other redundant causes in the BB experimental condition and that object D was a blicket than the other redundant causes in the BB control condition.

Future research will need to manipulate the base rates of blickets, similar to what was done in Sobel et al. (2004), to determine whether the present results reflect the operation of a Bayesian-inference mechanism or a RW-model-like associative-learning mechanism