Is Bayesian inference even necessary? Revisiting backwards-blocking reasoning in human children

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

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 children an associative-learning mechanism or a Bayesian-inference mechanism to reason about causal events. 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 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 is an inadequate model of human causal reasoning. 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. **The present investigation**

The present investigation had five broad goals. The first goal of the experiments presented here was to determine whether 4-, 5-, and 6-year-olds could engage in BB reasoning when asked to reason about three objects (Experiments 1 and 2) and when the elemental phase consists of two rather than one objects. The 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. The third goal was to gain greater insight into how—that is, by what cognitive mechanism—children reasoned about the present events. The fourth goal was to determine whether BB reasoning depends on whether one or two objects are shown during the elemental phase in the BB condition. Traditionally, the elemental phase in the BB condition has consisted of only a single object (i.e., object A). This means that it remains to be seen whether the BB effect is greater for object C (the object that is not shown when A and B are placed on the machine; object C is analogous to object B in the BB condition when two objects are used) when one object is shown during the elemental phase (i.e., object A) compared to when two objects are shown (i.e., objects A and B). Finally, because BB reasoning in previous research was operationally defined as greater treatment of object B in the BB condition compared to the ISO condition, here participants also experienced the ISO condition. It was important to demonstrate that participants who failed to show evidence of BB reasoning under the new operationalization of BB reasoning would nonetheless show evidence of it under the old operationalization. Failing to show BB reasoning under the new operationalization of BB as well as under the older operationalization of it would suggest that there were issues with the study rather than indicate a lack of BB reasoning in children. **Experiment 1**

Experiment 1 assessed whether 5- and 6-year-olds will engage in BB reasoning for 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.

**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.

**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 | ABCD+ | D+ | Is A/B/C/D a blicket? |
| ISO experimental trial | ABC+ | A- | Is A/B/C a blicket? |
| ISO control trial | ABCD+ | 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). Separate logistic mixed-effects models were fit to participants’ responses for the 5- and 6-year-olds as well as for the BB and ISO conditions, with participant as the random-effect factor in all models.

**5-year-olds.** Given that there was no effect of 5-year-olds’ pretest performance on their choices for any of the objects, we collapsed across this variable. The analysis revealed that they were equally likely to say that object A was a blicket during the BB experimental trials compared to the BB control trials, odds ratio (OR) = 2.6, *p* = .20, *BF*H1= .25, bootstrapped 95% CI[-4.34, 9.48]. Likewise, participants were as likely to say that objects B, and C were blickets during the BB experimental trials compared to the BB control trials, both ORs < .73, both *p*’s > .20, both *BF*H1’s = .25. Crucially, this result was not due to participants choosing to respond equally to all objects across the BB and ISO experimental and control conditions. In fact, participants were much less likely to respond that object A was a blicket in the ISO experimental condition than in the BB experimental condition, OR = .03, *p* < .0001, *BF*H1= .2249.40. [ADD THE DATA FOR THE ISO CONDITION. NOTE THAT THE R SCRIPT IS READY TO BE USED]

**6-year-olds.** Similar to the analysis with the 5-year-olds, an initial analysis that examined whether there was an effect of 6-year-olds’ pretest performance on their choices for any of the objects revealed that there was no such effect. Thus, we collapsed across this variable.