

Associative learning or Bayesian inference? Revisiting backwards blocking reasoning in adults

Deon T. Benton^{a,*}, David H. Rakison^b

^a Vanderbilt University, USA

^b Carnegie Mellon University, USA

ARTICLE INFO

Keywords:

Causal reasoning
Causal mechanisms
Computational models
Analytical models
Associative learning
Bayesian inference

ABSTRACT

Causal reasoning is a fundamental cognitive ability that enables humans to learn about the complex interactions in the world around them. However, the cognitive mechanisms that underpin causal reasoning are not well understood. For instance, there is debate over whether Bayesian inference or associative learning best captures causal reasoning in human adults. The two experiments and computational models reported here were designed to examine whether adults engage in one form of causal inference called backwards blocking reasoning, whether the presence of potential distractors affects performance, and how adults' ratings align with the predictions of different computational models. The results revealed that adults engaged in backwards blocking reasoning regardless of whether distractor objects are present and that their causal judgements supported the predictions of a Bayesian model but not the predictions of two different associative learning models. Implications of these results are discussed.

1. Introduction

No ability is more important for learning about how the world works than reasoning about cause-and-effect relations. This is a key cognitive capacity because it enables human learners to encode causal relations, to make predictions and inferences (e.g., Schlottmann & Shanks, 1992, to intervene on those relations to generate new effects (e.g., Gopnik et al., 2001), and to reason counterfactually about causal events to determine what would have happened if a different action were taken (e.g., Glymour, 2000; Harris et al., 1996; Pearl, 2000; Sobel, 2004; Spirtes et al., 2000; Woodward, 2004).

Despite agreement among researchers about the importance of causal reasoning, there is considerably less consensus about how—that is, by what cognitive mechanism—causal reasoning occurs. In the 1980s and 1990s this debate centered on whether human causal learning is best captured by associative models or statistical “multiple linear regression” models (e.g., Chapman, 1991; Dickinson & Shanks, 1985; Shanks, 1985). The focus of this debate has since shifted to whether causal learning and reasoning are best explained by Bayesian inference or by associative learning. Proponents of the Bayesian inference perspective argue that people use a simple form of Bayes rule to reason

about causal events (e.g., Gopnik & Tenenbaum, 2007). The idea is that learners must figure out which hypothesis—among potentially infinitely many hypotheses—best explains the observed causal data. To solve this problem, learners must combine information about the likelihood of each hypothesis—before any data are observed—with information about how likely one would observe some data if a given hypothesis were true (independent of any observed data) by application of Bayes rule. Support for the notion that individuals use Bayesian inference to reason causally has been generated from research with infants (e.g., Sobel & Kirkham, 2006, 2007), children (e.g., Gweon et al., 2010; Schulz & Gopnik, 2004), and adults (e.g., Bonawitz et al., 2014; Griffiths et al., 2011; Schulz et al., 2007; Steyvers et al., 2003).

In contrast, proponents of the associative learning perspective maintain that causal reasoning involves the formation of “associative links” between candidate causes and effects. The idea is that the more a candidate cause is associated with an effect, the greater the strength of the connection (i.e., association) between the cause and effect. One of the earliest models of associative learning used to explain causal learning in both nonhuman animals as well as in people was the Rescorla-Wagner (RW) model (Rescorla & Wagner, 1972; see also Gluck & Bower, 1988, 1990). According to this model, learning an associative

* Corresponding author at: Department of Psychology and Human Development, Vanderbilt University, Peabody College, 230 Appleton Place #552, Nashville, TN 37235, USA.

E-mail address: deon.benton@vanderbilt.edu (D.T. Benton).

<https://doi.org/10.1016/j.cognition.2023.105626>

Received 13 March 2023; Received in revised form 12 September 2023; Accepted 15 September 2023

Available online 26 September 2023

0010-0277/© 2023 Elsevier B.V. All rights reserved.

relation between a cause and an effect depends on whether the cause predicts the effect. If a candidate cause does not predict an outcome to the level that it should (i.e., if there is a discrepancy between what is expected and what is observed), the “weight” or associative strength between that cause and the outcome will be changed until the discrepancy disappears (and the cause is able accurately to predict the effect or outcomes). This perspective has also garnered significant empirical support. For example, it has been used to explain causal perception in infants (e.g., Rakison, 2005, 2006; Rakison, Smith, & Ali, 2016) and causal reasoning in children (Benton et al., 2021; Kloos & Sloutsky, 2013) and adults (e.g., Aitken et al., 2001; Chapman, 1991; Chapman & Robbins, 1990).

What remains unresolved is which of these accounts provides the *best* explanation of causal learning in people. One way that researchers have tried to resolve this debate is through research on a causal learning phenomenon called “backwards blocking reasoning”. This is a form of “retrospective reevaluation” in which the observed causal status of one object—say, object A—causes learners to ‘go back’ in their memory to reevaluate and then discount as a potential cause any other cue that may have been presented in combination with cue A (hence, “retrospectively reevaluate”). Although this phenomenon has been studied in human beings since the early 1980s, only somewhat recently have researchers returned to it to determine whether associative learning or Bayesian inference best explains backwards blocking reasoning.

Consider a classic study by Sobel et al. (2004), for example. This study was designed to examine whether 3- and 4-year-old children could engage in backwards blocking reasoning. Children were introduced to a machine called a “blicket detector” and told that objects called “blickets” are the only ones capable of activating the machine by causing it to light up and play music. Following this introduction, children were shown initially that the machine activated when two novel objects, A₁ and B₁, were placed on the surface of the machine, and then that the machine activated when A₁ alone was placed alone on the machine. If children engaged in backwards blocking reasoning, observing that A₁ activated the machine by itself should cause them to go back in memory to “block” or discount B₁ as a cause of the machine’s activation. This is because the effect could fully be explained by A₁. In a closely related event called the *indirect screening-off* event, children were shown that two new objects, A₂ and B₂, activated the detector together and then that object A₂ failed to activate the machine. This condition is called “indirect screening off” because after observing that object A₂ failed to make the machine go, participants should go back in their memory to “screen off” or remove the causal link that initially children may have assumed existed between A₂ and the machine’s activation. Children in both conditions were then asked which of the two objects were blickets and to make the machine go by placing the appropriate object on it.

Sobel et al. (2004) predicted that if children could engage in backwards blocking reasoning, then object B₂ in the indirect screening-off event should be placed on the machine to make it activate more often than B₁ in the backwards blocking event. The data supported this prediction with the 4-year-olds but not with the 3-year-olds (though see Sobel & Munro, 2009 for evidence of backwards-blocking reasoning with 3-year-olds). Sobel et al. (2004) extended this finding in a separate experiment to show that when the frequency that objects were blickets was manipulated such that blickets were made to be “rare” or “common”, participants’ treatment of object B differed. They were more likely to consider object B a blicket when blickets were common (e.g., when 10 out of 12 objects were shown to activate the machine prior to the critical backwards blocking event) compared to when they were rare (e.g., when 2 out of 12 objects made the machine go).

Griffiths et al. (2011) extended these findings to show that adults are also sensitive to base rates, but across a much wider range than that tested in Sobel et al. (2004). Participants were introduced to a “super lead detector” and told that only “super pencils” activated the detector. The experimenter then picked two pencils (at random) from a cup and showed participants that one of the pencils was a “super pencil” (it

activated the machine when “scanned”); the other pencil was not a super pencil because it failed to activate the machine. Then participants were assigned to one of five base rate conditions: 1/6, 1/3, 1/2, 2/3, and 5/6. For example, in the 1/2 base rate condition, participants removed and then scanned each of 10 more pencils (from the same cup). Five of the 10 pencils activated the machine. In total, participants observed 6 activations and 6 nonactivations. Following the base rate manipulation, all participants were shown the standard backwards blocking event. The super-pencil detector activated when two new pencils, A and B, were scanned and then activated again when A alone was scanned. Participants were instructed to rate how confident they were that both objects were super pencils on a scale from 1 to 7. One indicated that they were maximally not confident that the pencil was a super pencil; 7 indicated that they were maximally confident that it was a super pencil.

A crucial distinction between this study and that by Sobel et al. (2004)—in addition to the fact that a wider range of base rates were used—was that participants in Griffiths et al. (2011) provided confidence ratings before the backwards blocking trial began, midway through the trial (i.e., after objects A and B were scanned), and then after the trial had concluded (i.e., after object A was scanned). We maintain that this is a more valid operationalization of backwards blocking than the one used by Sobel et al. (2004) that involved comparing participants’ treatment of object B between the backwards blocking and indirect screening-off events. This way of operationalizing backwards blocking reasoning makes it possible to determine to what degree participants’ treatment of object B (the object that only ever appeared with object A and was never placed on the machine alone) changed over the course of the backwards blocking event. In our view, this captures the essence of what it means to engage in backwards blocking. This also explains why we use this operationalization of backwards blocking reasoning in the current work.

Griffiths et al. (2011) found that as the base rate of super pencils decreased (from more common to rarer), participants’ confidence that object B was a super pencil also decreased. This is a sensible response: If most of the things you observe are not causes, then it is reasonable to conclude that a new thing is also not a cause. Crucially, the authors found these data were best explained by a simple Bayesian model than other competing models. An unanswered question concerns *why* studies on backwards blocking are problematic for certain associative learning accounts such as the traditional RW model. The reason is straightforward: This model predicts that the strength between object B and the machine’s activation should remain unchanged between the backwards blocking (where A is effective) and indirect screening-off (where A is not effective) trials as well as across the different phases of the backwards blocking trial (e.g., Griffiths et al., 2011). This is because B would have been shown with the machine’s activation an equivalent number of times across the backwards blocking and indirect screening-off trials.

However, there are several reasons to exercise caution before accepting the conclusion that Bayesian inference rather than associative learning underlies human causal reasoning and inference. The first is that when one considers the backwards blocking literature in its entirety rather than a subset of studies, the evidence for backwards blocking reasoning is considerably mixed. For example, some studies showed evidence of backwards blocking reasoning in children and adults (e.g., Griffiths et al., 2011; Sobel et al., 2004), whereas other studies either showed no evidence of it (e.g., Larkin et al., 1998) or evidence of backwards blocking reasoning but only under very specific conditions such as when cues are described as causes rather than effects (e.g., Beckers et al., 2005; De Houwer et al., 2002; Kloos & Sloutsky, 2013; Lovibond et al., 2003; Waldmann & Holyoak, 1992). Thus, the robustness of backwards blocking reasoning remains unclear, which in part was one of the motivations for conducting the present research.

A second reason to exercise caution is that by comparing participants’ treatment of object B in the backwards blocking condition to their treatment of object B in the indirect screening-off condition (as was done in Sobel et al., 2004), it is impossible to know whether participants’

differential treatment of object B is due to backwards blocking, indirect screening-off, or some combination of both (although see De Houwer et al., 2002; Larkin et al., 1998; Kruschke & Blair, 2000; Lovibond et al., 2003; Shanks, 1985; Van Hamme & Wasserman, 1994 for alternative operationalizations). It turns out that when a more appropriate measure of backwards blocking reasoning is used—such as when participants' treatment of object B in the standard backwards blocking condition (i.e., AB+ A+) is compared to their treatment of object B in a closely related control event (i.e., AB+ C+)—there is some evidence of backwards blocking in children and adults (Griffiths et al., 2011; McCormack et al., 2009; see also McCormack et al., 2013). Although this alternative operationalization of backwards blocking reasoning represents an improvement over that used by Sobel et al. (2004), it also has its limitations. The main limitation is that it cannot provide insight into whether participants' treatment of object B changes over the course of the backwards blocking event. As we indicated above, backwards blocking is a phenomenon whose name implies a change in the treatment of object B at different points of the backwards blocking event. Studies that measure backwards blocking reasoning only after the event has concluded—which, as it turns out, is all studies dating back to the early 1980s except for the Griffiths et al. (2011) study—cannot capture this changing dynamic.

A final reason to exercise caution is that it remains largely unknown how the *absence* of base rate manipulations impacts adults' causal inferences when a more appropriate measure of backwards blocking reasoning is used. Of course, there are principled reasons to manipulate base rates. One reason is to determine to what extent participants' causal judgements match the predictions of a Bayesian model. The benefit of this approach is that it can shed light on whether cognitive mechanisms that implement or approximate Bayesian inference are in operation. However, the difficulty is that this approach has somewhat limited ecological validity. For example, when a child or an adult first encounters a novel causal system in the real world—which may be unlike any they have ever encountered (e.g., getting a new gadget from the local electronics store)—very rarely, if ever, are they provided with the base rates of causes in that system to aid their causal inferences. Instead, they must figure out which things are causes and which things are not through trial and error. This means that studies that manipulate base rates, although useful, tell us very little about whether Bayesian inference (or some mechanism that approximates it) is in operation in contexts in which base rate information is not explicitly provided such as in the real world.

2. The present research

With these concerns in mind, the present research had two goals. The first was to determine whether human adults would engage in backwards blocking reasoning in a blinket detector context but adopting Griffiths et al.'s operationalization of backwards blocking. The first experiment tested adults' backwards blocking inferences in the context of two objects and across three rating phases. Experiment 2 was essentially a replication of Experiment 1 with one key exception. We included a third and fourth “distractor” object to determine whether the presence of such objects would impact whether adults engaged in backwards blocking reasoning. To our knowledge, Sobel and Kirkham (2006) represent the only work previously to examine the role of a distractor object on toddler's backwards blocking inferences. In this study, an experimenter placed three objects (i.e., objects A, B, and C) on a table and then showed 19- and 24-month-old children that objects A and B together made the detector activate when they were placed on it. Object A was then placed alone on the detector, which caused the detector to activate. Object C remained on the table throughout the trial and was never demonstrated on the machine; it merely served as a distractor. This research revealed that although the 24-month-olds were unaffected by this distractor object, the backwards blocking performance of the 19-month-olds was impacted. However, it remains unknown whether

the same is true with adults but for 3 (or more) distractor objects.

The reason we chose to test adults rather than children was to provide the cleanest possible test for the presence of Bayesian inference. If we conducted a version of the current study with children and found that their backwards blocking performance was better explained by an associative model than by a Bayesian model, it would be difficult to know whether this was due to the operation of an associative learning mechanism or because of insufficient information processing abilities. In particular, asking children to provide three separate ratings of objects A and B across the three rating phases of the backwards blocking event could inadvertently impose a tax on children's information-processing abilities that in turn causes their performance to suffer. We can eschew this scenario altogether by testing adults given that they presumably possess more robust information-processing abilities than children. This means that we can be confident that if adults' inferences in Experiments 1 and 2 were best captured by an associative model than by a Bayesian model, it was not due to insufficient information-processing abilities. We used the blinket detector paradigm rather than the super-pencil detector paradigm because many contemporary studies on causal reasoning have used this paradigm. This has the benefit of allowing for cross-study comparisons between children and adults.

The second goal of the present work was to investigate the cognitive mechanism by which adults processed the present backwards blocking events (see the “Computational Models” section). We were specifically interested in determining whether a simple, two-layer connectionist neural network trained with the Delta rule—which has been shown to be formally equivalent to the traditional RW model (e.g., Danks, 2003; Gluck & Bower, 1988)—the modified Rescorla-Wagner model (Van Hamme & Wasserman, 1994)—which has been presented as an improved alternative to the traditional RW model—or a simple Bayesian model best explained adults' current backwards blocking performance. If participants' causal ratings were consistent with one of these models, this would suggest that adults' causal performance was underpinned by the mechanism implemented by that model.

3. Experiment 1

Experiment 1 was designed to determine whether adults would engage in backwards blocking reasoning in a blinket detector context. As we mentioned in the Introduction, this is a paradigm that has been used extensively to study causal reasoning in human children but surprisingly never in adults. Participants were introduced to the “blinket detector” and were told that their task was to determine which objects made the machine activate (i.e., which objects caused it to light up and play music). Participants were told that the machine activates by lighting up and playing music only when objects called blinkets are placed on its surface. Participants then received three trials: backwards blocking, one-cause, and two-cause. The one- and two-cause trials were included to ensure that participants understood the task and could reason about blinkets. For the critical backwards blocking trial, participants rated how confident they were that each object was a blinket before the backwards blocking trial began, midway through the backwards blocking trial (i.e., after A and B were placed on the machine [AB+]), and after the trial ended (i.e., after the A+ trial).

3.1. Method

Participants, stimuli, and design. Twenty-four college students were recruited for Experiment 1 and were given course credit for their participation (11 males; 13 females). Demographic information about the participants was not collected. Sample size was determined based on power analysis using “pwr2” package in R software (Lu et al., 2017). We aimed to achieve a power of .80 with an alpha level of 0.05 to detect a large effect size (Cohen's $f = 0.80$) for the interaction effect between Phase (pre, mid, or post) and Object Type (A vs. B). Based on these parameters, the power analysis indicated that 16 participants were

needed.

Stimuli and Design. A blicket detector was used (e.g., Gopnik et al., 2001; Sobel et al., 2004). The box was 5" x 7" x 3", was made of wood that was painted black, and had a white semi-clear top. The machine operated via a remote control attached to the end of an electric wire, which itself was attached to the side of the box. The remote control contained a single, circular red button. When the experimenter pressed this button and simultaneously placed the object predetermined to be a blicket on the machine's surface, the music played, and the lights flashed. Experimenters were instructed to press the button precisely at the moment the predetermined blicket was placed on the machine to ensure that there was no temporal delay between when the object was placed on the machine and when the machine activated. If an object that was not predetermined to be a blicket was placed on the detector, the button was not pressed, and the music and lights did not play or flash. If one or more objects caused the machine to activate (because one or more of them were blickets), experimenters were instructed to say, "Look, these objects/this one made the machine go". If one or more objects failed to cause the machine to activate, the experimenters were instructed to say, "Look, these objects/this one failed the make the machine go." Crucially, this text was only read aloud during the one- and two-cause trials. This meant that participants were not read this text during the critical backwards blocking event to prevent the possibility that they interpreted the experimenter's comments as explicit information about the causal efficacy of objects.

The stimuli for this experiment were eight cube and cylinder objects (four of each). Each object was approximately 1" in diameter and differed in color. Within each trial, the objects were chosen to be different colors and shapes and the designated blicket object was counterbalanced across participants. Thus, across trials new objects were used, which meant that participants never saw the same object in different trials to prevent potential carryover effects.

Procedure. Participants were tested in a quiet room on campus. At the beginning of the experiment, participants were introduced to the machine and told it was called a blicket detector that activated when blickets were placed on it but not when objects that were not given the blicket label were placed on it. Following this introduction, participants were given three test trials in counterbalanced order: one-cause, two-cause, and backwards blocking (e.g., Gopnik et al., 2001; Sobel et al., 2004). Following each trial, participants were asked to provide causal ratings for objects A and B at three different points: before a trial, midway through a trial, and after a trial. Specifically, participants were asked, "How confident are you that each object is a blicket on a scale that ranges from 0 to 100. A 0 on this scale indicates that you are absolutely confident that a given object is not a blicket. A 100 on this scale indicates that you are absolutely confident that a given object is a blicket." Participants were encouraged to use the entire range to avoid anchoring on either the low or high ends of the scale.

During the one-cause trial, participants were shown two novel objects (A and B). Object A was then placed on the detector, which caused it to activate. Object B failed to activate the detector when it was placed on it. Objects A and B were then placed together on the machine twice. This caused the detector to activate both times (i.e., A+; B-; AB++). During the two-cause trial, a different A object was placed on the machine three times, and the machine activated all three times. Object B—which was different from that used in previous trials—was then placed on the machine three times. The machine failed to activate the first time that object B was placed on the machine, but it activated the remaining two times that object B was placed on it (i.e., A+++; B-; B++). The one- and two-cause trials have been used in previous research and were included here to ensure that participants understood how the machine worked and recognized what distinguished blickets from non-blickets. In the critical backwards blocking trial, new A and B objects together were placed on the machine, which caused the machine to activate. This was performed twice. Object A was then placed on the machine by itself, which again caused the machine to activate (i.e.,

AB++; A+).

In all three trials, participants were asked to rate each object using the previously mentioned scale, before, midway through, and following each trial. The exact points at which participants provided these ratings are shown below in Table 1.

3.2. Results

All analyses were conducted in R (R Core Team, 2021). Fig. 2 shows the mean causal ratings of objects A and B across the three ratings phases in the backwards blocking condition. Given evidence of non-normality and unequal variance in the causal-rating data, all analyses used non-parametric analyses with 4000 replications. We relied principally on two non-parametric approaches. First, we used bootstrapping to estimate confidence intervals. Second, statistical significance was evaluated using permutation tests. *P*-values represent the probability of obtaining the observed effect size (or larger) if the null hypothesis of zero difference between phases is true. Given that the models (discussed in more detail in the Computational Models section) differ only in their predictions for how participants should respond during the mid- and post-rating phases, the analyses reported here focused on participants' mid- and post-ratings.

A linear mixed-effects model was fit to participants' causal ratings in the backwards blocking condition. Participants were included as a random-effect factor and phase (pre vs. mid vs. post) and object type (A vs. B) were included as within-participants fixed-effects factors. This analysis yielded main effects for Phase, $F(2,115) = 17.21, p < .0001$ and Object Type, $F(1,115) = 21.61, p < .0001$, which were qualified by a Phase x Object Type interaction, $F(2,115) = 29.29, p < .0001$. Planned comparisons for object A revealed that although participants' A ratings increased between the mid- and post-rating phases, $p < .0001$, participants' ratings of object B did not differ across the rating phases, all p 's $> .2$.

3.3. Discussion

Experiment 1 was designed to determine whether human adults engage in backwards blocking reasoning in a blicket detector task. The results indicated that participants' causal ratings of object A increased across the pre-, mid-, and post-rating phases, whereas their ratings of object B remained the same across all three rating phases. This was a surprising result principally because—as we will show in the Computational Models section—the simple Bayesian model and simple neural network model differ in their predictions for participants' mid- and post-ratings of object B. The traditional RW model predicts that participants' post-ratings of B should be equivalent to their mid-ratings of it; the Bayesian model predicts that participants' post-ratings of B should be less than their mid-ratings of it, with diminishing differences as the base rate of blickets increases.

Given the implications that this finding—that is, that participants' B ratings did not change over the course of the backwards blocking trial—could have on theories of backwards blocking reasoning, we attempted to replicate it in Experiment 2. The design of this experiment

Table 1
The three points at which participants were asked to provide pre-, mid-, and post-ratings in the across the four conditions. An * corresponds to the point at which participants were asked to rate objects A and B. The number of +/- signs corresponds to the number of times a particular sequence was demonstrated and whether the effect happened (+) or did not happen (-).

Schematic of Experiment 1					
	PRE		Mid		Post
One-Cause	*	A+; B-	*	AB+	*
Two-Cause	*	A+++	*	B-; B++	*
Backwards-Blocking	*	AB++	*	A+	*

was identical to Experiment 1 with one exception: Participants were asked to provide causal ratings for three objects (i.e., objects A-C) or four objects (i.e., objects A-D). Crucially, in both conditions (i.e., the three- and four-object conditions) the only objects to participate on the machine were objects A and B—object C in the three-object condition and objects C and D in the four-object condition remained motionless on the table throughout the entire backwards blocking event. By asking participants to provide causal ratings for three and four objects—instead of two objects—we could determine whether the mere presence of additional objects (beyond the ones that participated on the machine) impacted adults' backwards blocking performance. As we mentioned earlier, the only study to examine whether backwards blocking reasoning was impacted by the presence of other objects was that by Sobel and Kirkham (2006, Exp. 1) with 19- and 24-month-old infants. Thus, it remains unknown whether backwards blocking reasoning is affected by the presence of other objects in human adults, even if those other objects do not participate on the machine. As in Experiment 1, participants were asked to rate how confident they were that each object was a blicket before, midway through, and following the backwards blocking event.

4. Experiment 2

4.1. Method

Participants. Twenty-four college students were tested in the B three- object condition (13 females, 11 males) and 20 students were tested in the four-object condition (13 males, 7 females). Participants received course credit for their participation. These participants were different from those tested in Experiment 1.

Stimuli, Design, and Procedure. All aspects of Experiment 2 were identical to Experiment 1 with one exception: For participants who were asked to provide causal ratings for three objects (i.e., objects A-C), nine objects were used. These included cube-, cylinder-, and triangular-prism-shaped objects, each differently colored and approximately 1" in diameter. For participants asked to provide causal ratings for four objects (i.e., objects A-D), twelve objects were used. These included cube-, cylinder-, triangular-prism-, and rectangular-prism-shaped objects, each differently colored and approximately 1" in diameter. For both groups, objects that were used on one trial were not used on another trial to prevent potential carryover effects: within each trial, the objects were different shapes and colors and the designated blicket was counterbalanced across participants. Crucially, for both groups, only objects A and B participated on the machine. This meant that object C for the three-object condition and objects C and D for the four-object condition remained motionless on the table in plain sight throughout the entire backwards blocking event. This mirrored the approach taken in Sobel and Kirkham (2006). We did not include the one- and two-cause trials.

4.2. Results

Given that participants treated objects A and B equivalently across the three- and four-object groups and experimental phases (i.e., the pre-, mid-, and post-phases), $F(2, 171) = 0.22, p = .80$, and that the four-object group was insufficiently powered to detect a large effect with a power of .80, we combined the data across both groups. Fig. 2 shows the mean pre and post causal ratings of objects A, B, C, and D across the three rating phases of the backwards blocking event. An additional linear mixed-effects model was fit to participants' causal ratings of objects A-D with the same random and fixed effects as those used in Experiment 1. This analysis yielded main effects for Phase, $F(2, 421) = 18.36, p < .0001$ and Object Type, $F(3, 421) = 38.55, p < .0001$, which were qualified by a Phase x Object Type interaction, $F(6, 421) = 21.03, p < .0001$. Planned (permutation) comparisons for object A revealed that participants provided higher post-ratings of object A ($M = 96.05$,

Bootstrapped 95% CI[91.09, 101]) than mid-ratings ($M = 65.36$, Bootstrapped 95% CI[59.55, 71.18]) of it, $p < .0001$. In contrast, participants provided reliably higher mid-ratings of object B ($M = 60.29$, Bootstrapped 95% CI[53.61, 66.98]) than post-ratings ($M = 47.50$, Bootstrapped 95% CI[39.22, 55.78]) of it, $p < .01$. The final set of analyses showed that participants' ratings of objects C and D did not differ across the three rating phases. These results are inconsistent with those from Experiment 1 and demonstrate clearly that participants engaged in backwards blocking reasoning: Participants' causal ratings of object B decreased between the mid- and post-rating phases. This suggests that the results from Experiment 1 likely reflected an issue of insufficient statistical power and thus should be interpreted cautiously.

4.3. Discussion

In contrast to the results from Experiment 1, data from Experiment 2 provided clear evidence that adults engaged in backwards blocking reasoning. This conclusion is based on the fact that when the data for participants in the three- and four-object conditions were combined, participants' causal ratings of object B decreased between the mid- and post-rating phases of the backwards blocking event. This suggests that the null effect in Experiment 1 was due to a lack of statistical power rather than to the fact that adults did not engage in backwards blocking reasoning. The fact that adults' causal ratings seem to be consistent with a Bayesian inference account but not with an associative learning account based on the RW model suggest that adults used Bayesian inference to reason about the present backwards blocking events. Nonetheless, it is worth keeping in mind that although these results are consistent with a Bayesian inference account, this is a *qualitative* conclusion that requires converging, *quantitative* support. As such, the goal of the Computational Models section below was to determine which model provided the best *quantitative* fit to the behavioral data. If our conclusion is correct that adults' causal inferences are consistent with a Bayesian inference account, then we would expect that a model based on Bayesian inference would provide a better quantitative fit to the data than a connectionist model based on the traditional RW model. We also explored the possibility that the current data are consistent with a different associative account which, to our knowledge, has not been explored as well as with a modified version of the traditional RW model called the "modified Rescorla-Wagner model" (Van Hamme & Wasserman, 1994). We then compare the quantitative fit of this modified associative account to the data with the quantitative fit of the Bayesian model to the data.

5. Computational models

We fit two separate computational models to the behavioral data in Experiments 1 and 2: A model based on Bayesian inference and a simple connectionist model that implemented a version of the traditional RW model. The Bayesian model was based on that presented in and Griffiths et al. (2011).

Bayesian Model. A fundamental assumption of Bayesian inference is that the causal induction process involves determining which causal hypothesis—within a space that can consist of potentially infinitely many such causal hypotheses—is most consistent with the observed data by application of Bayes rule. Typically, these models are expressed as parameterized causal graphical models with nodes that are connected by edges that encode the Markov condition (or the idea that two cues are probabilistically independent of one another conditioned on their parents). A key assumption of the Bayesian inference approach is that at the beginning of a learning episode, an ideal Bayesian reasoner represents all possible candidate hypotheses, H , whereby each hypothesis, $h \in H$, is assigned some prior probability, $p(h)$. This prior probability represents the learners' confidence that a given hypothesis generated the causal data that is ultimately observed but prior to the actual presentation of that data. Fig. 1 shows the possible hypotheses for Experiment 1 in

which adults were asked to reason about two objects. Note that the causal hypothesis space for the three- and four-object conditions in Experiment 2 consists of 8 and 16 causal hypotheses, respectively. Following observations of data, d , the learner uses Bayes' rule to compute and assign posterior probabilities to each hypothesis, $p(h|d)$,

$$p(d) = \frac{p(h)p(d|h)}{\sum_h p(h)p(d|h)},$$

where $p(d|h)$ represents the likelihood or the probability of the data d under a given hypothesis h . The denominator serves as a normalizing term to ensure that the sum of the posterior probabilities of each hypothesis is 1 (see Fig. 3).

Because the hypotheses in all of the experiments presented here are assumed to be deterministic (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 (Fig. 1) and is consistent with the observed data; otherwise, they are set to 0. If participants observed a negative effect (e.g., if the machine failed to activate when object A was placed on it), the likelihood is set to 1 if a link also does not exist between that object (in this case, object A) and the machine's activation (otherwise it is set to 0) in the model. Once we have determined whether such a link exists for a particular object, we can compute the likelihood that any of the objects in the experiments are blinkets by taking the product of the likelihood that object A (or B) activated the detector under each hypothesis (according to the previously mentioned procedure) and the posterior probability of each hypothesis, after which we simply sum this product. For example, we can determine the probability that object B is a blanket by computing the following equation:

$$p(d) = \sum_h p(h)p(d|h), \quad (2)$$

where $p(B \rightarrow E | h)$ is 1 if there is an edge between that object and the detector in a given hypothesis h , and 0 otherwise.

Given that that we did not explicitly manipulate prior probabilities in any of the experiments (which is one of the key parameters in a Bayesian model), it was important to determine which prior led to the best fit to the human data. We carried out different runs of the model with the following prior probabilities: .1, .2, .3, .4, .5, .6, .7, .8, .9, .95. We found that the prior probability that provided the best fit to the data across both experiments was .50. Thus, the model fit indices that are reported below are based on this probability.

Connectionist model. The connectionist model we built was a simple two-layer network. Shown below in Fig. 4 is the architecture used to simulate the four-object condition in Experiment 2. The architectures used in the simulations of Experiment 1 and the three-object condition in Experiment 2 were identical except that for the Experiment 1 simulation the input layer consisted only of two units (which corresponded to two objects), whereas the input layer consisted of three units (corresponding to three objects) for the simulation of the three-object condition in Experiment 2.

For all simulations, the model consisted of an input layer and an output layer—there were no hidden layers in this model. The rationale for building a simple two-layer model rather than a more complex model that consisted of many layers and units within layers was to explore whether a simple connectionist model trained with the Delta rule (Widrow & Hoff, 1960) could explain the current data. Given that the Delta rule has been shown to be equivalent to the RW model (Danks, 2003; Gluck & Bower, 1988), the results of this model could therefore be taken as an indication of how well the traditional RW model could

explain the current data.¹

The output layer consisted of a single unit for all simulations. Each input unit corresponded to each of up to four total objects used in the experiment. Whenever an object was present, the input unit that corresponded to it had an activation value of "1". Input units of objects that were absent were set to "0". If an object that was predetermined to be a blanket was presented to the model, the target activation of the single output unit in the network was set to "1". Given that the observed activity of the single output unit was unlikely to match the target at the beginning of training, the weights in the network were adjusted—via application of the Delta rule—to reduce the difference between the observed activation and the target activation. If a nonblanket were presented to the model, the target activation of the single output unit was set to "0". To simulate the fact that object C and objects C and D in the three- and four-object conditions, respectively, in Experiment 2 did not participate on the blanket machine, the weights between the input units that stood in for these objects and the single output unit were disconnected. Thus, any weight adjustments that were made between any of the input units and the single output unit did not apply to these objects (i.e., units). We used a similar approach to that above for the Bayesian model to determine which value for the number of training epochs (from the following list: 500, 1000, 1500, 2000, 2500, 3000) yielded the best fit to the data. We found that training the model for 500 epochs yielded the best model fit to the data across both experiments. Thus, the model fit indices reported below are based on this value for the number of training epochs. It is important to note that because the number of training epochs was the only "free parameter" in the model—all other parameters were held constant—this model had no more degrees of freedom than the simple Bayesian model. This means that any difference between the connectionist model and the Bayesian model in their fits to the human data cannot be attributed to differences in model parameters.

All simulations were initialized with small random values (distribution range = ± 0.1), used a fixed learning rate of 0.025, and had no momentum. In addition, the activations of the input unit had a small amount of uniform noise (range = 0.1) applied to them to reflect a small amount of noise in participants' encoding of an object. Finally, the activation of the output unit was determined by the logistic or sigmoidal activation function. The value of the activation of the single output unit was interpreted as the model's confidence that a given object was a blanket and could range between 0 and 1. Thus, if object A was presented to the model (i.e., its input unit was set to 1) and the model produced an output activation of 0.50, this indicated that the model was uncertain about A's status as a blanket.

The procedure for training the models mirrored that for adults across the two experiments. For example, during the portion of the backwards blocking event in which objects A and B were placed on the machine (thereby causing the machine to activate; i.e., the AB+ portion), for the corresponding portion of this event in the simulation, the first two input units were turned on (i.e., the activation of each input node was set to a value of 1), and the model's task was to learn to activate the single output unit (i.e., to set the activation of the single output unit to match the target of 1). During the subsequent A+ portion, the network's task was again to learn to activate the single output node. Finally, because the best fitting model required 500 epochs of training, this meant that the network experienced the AB+ phase for 500 epochs and the A+ phase for 500 epochs for a total of 1000 epochs of training.

5.1. Results

To assess the quantitative fit of the predictions of the connectionist

¹ For completeness, it is worth noting that the Delta rule (and by extension, the RW model) can be thought of as representing a Bayesian algorithm in that it models agents as doing a (very simple form of) Bayesian inference except with different constraints on the prior and conditional probabilities.

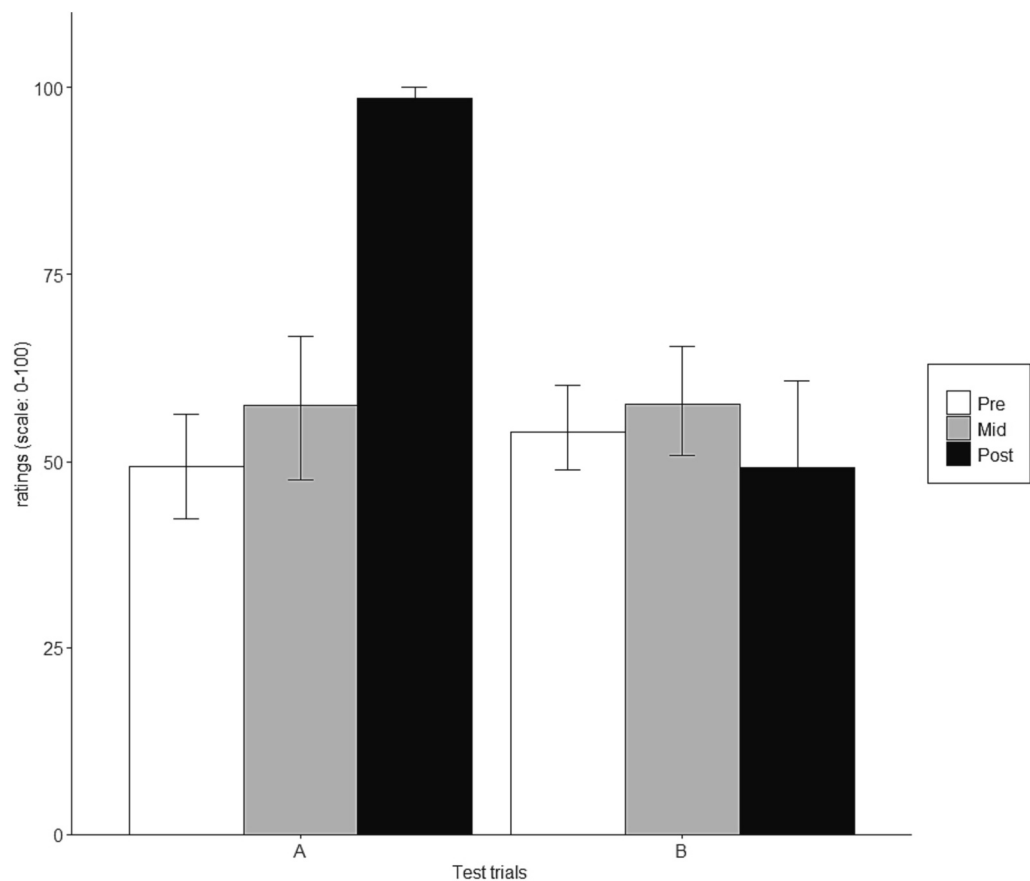


Fig. 1. Subjects' causal ratings of objects *A* and *B* before each of the four test event sequences were shown, midway through each event sequence, and after the entire event sequence had been demonstrated.

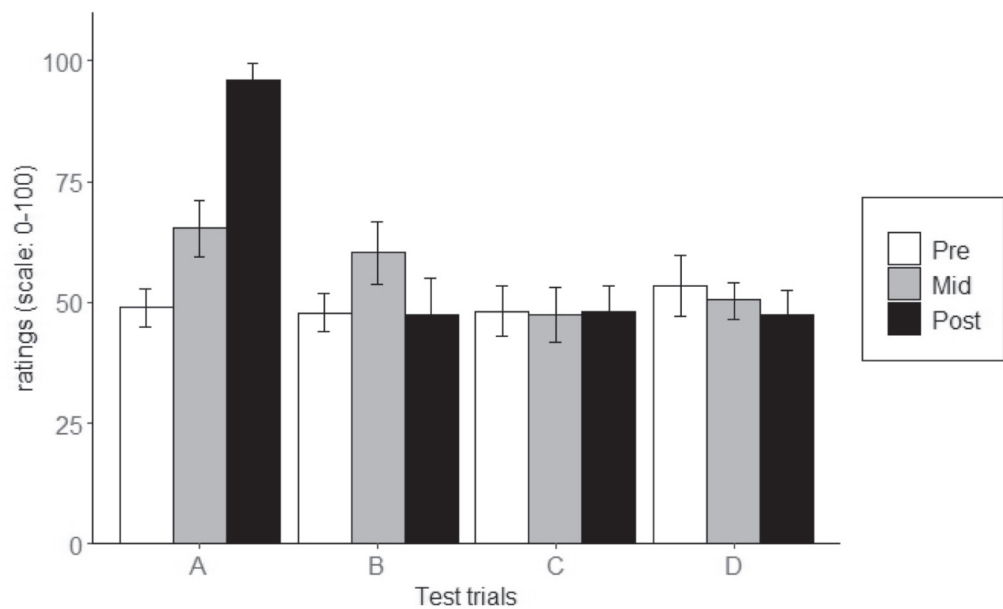


Fig. 2. Participants' causal ratings of objects *A*, *B*, *C*, and *D* at each of the same time points.

and Bayesian models to the data, we computed the root mean square (RMSE) and mean R-squared (R^2) between each model's predictions (for the connectionist model these were the average activation of the single output unit in response to each object; for the Bayesian model these were point estimates) and participants' mean responses to objects A-B in

Experiment 1, objects A-C in the three-object condition in Experiment 2, and objects A-D in the four-object condition also in Experiment 2. We chose both metrics because they have been used in previous simulation studies to assess the quantitative fit of a model to human data (e.g., [Bhat et al., 2022](#); [Buss & Spencer, 2014](#); [Spencer et al., 2022](#); [Steyvers et al.,](#)

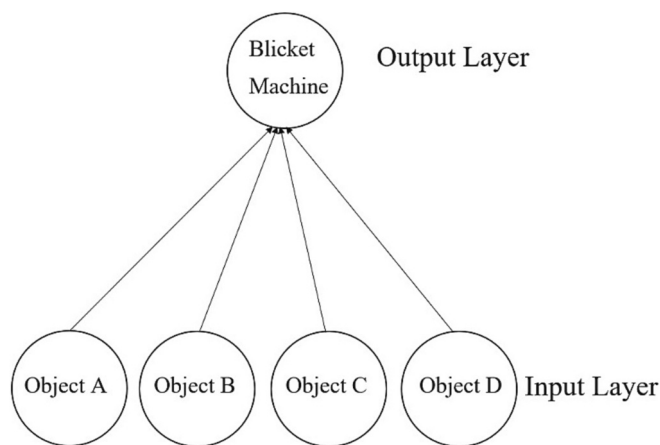


Fig. 4. The connectionist model used to simulate the four-object condition in Experiment 2.

2003; Stojnic et al., 2023). Lower and higher values for RMSE and R^2 , respectively, indicate better model fit. Table 2 above shows model fits.

It should be clear that the simple Bayesian model yielded a lower RMSE value and a higher R^2 value than the simple connectionist model. This was true across both experiments. This indicates that the Bayesian model provided the best quantitative fit to the data. This finding suggests that the traditional RW model is insufficient to explain adults' backwards blocking inferences in the present study. This finding extends that by Griffiths et al. (2011) to show that adults will continue to engage in backwards blocking reasoning in a blinket-detector context and regardless of whether or not distractor objects are present. Although the Bayesian model did outperform the connectionist model, it is possible that a more ecologically valid reformulation of the traditional RW model (or the putative “modified” RW model; Van Hamme & Wasserman, 1994) could better explain participants' causal inferences than the present connectionist model as well as potentially provide a better quantitative fit to the data than the simple Bayesian model.

An improved connectionist model. The fact that the simple Bayesian model provided a better quantitative fit to the data than did the connectionist model suggests that the form of associative learning stipulated by the traditional RW model was insufficient to explain adults' causal inferences in the present experiments. However, it is possible that participants relied on a form of associative learning that is based on, but crucially not identical to, that stipulated by the traditional RW model. This new account assumes that, at the beginning of the backwards blocking trial before objects A and B together are placed on the machine, adults began this trial with some amount of uncertainty about the causal statuses of objects A and B. This may be a learned assumption based on the fact that candidate causes in the real world sometimes produce effects and sometimes do not, but we remain agnostic about the origins of

this assumption (and do not attempt to model it here). However, once A and B are placed on the machine, adults' beliefs about the causal status of objects A and B are increased; that is, the individual associations between each object and the machine's activation increases in strength following the AB+ event. This increase in strength is easily predicted by the traditional RW model (as well as by the Bayesian model, although for this model this increase is increasingly smaller as the probability that any one object is a blinket increases), although the cognitive mechanisms that underlie both predictions differ. On the traditional RW model, this increase in strength is due to weight adjustments made between each object and the effect that are proportional to the difference between what the network expected and what it observed. In contrast, on the Bayesian inference account, this increase (at least for certain base rates) is due to fact that the data are more consistent with models in which A and B are causes than models in which neither are causes or models in which one but not the other is a cause.

However, when adults subsequently observe that A can activate the machine by itself, their confidence that B is a cause returns to its baseline levels. This decrease may be due to a learned assumption that single effects tend to be produced by single causes. This itself may also be based on extensive real-world experience with events in which single effects tend to be produced by single causes (e.g., single light switches cause single lamps to come on; a single ‘on’ buttons cause single television sets to turn on, etc.). It turns out that there is some empirical support for this idea. For example, Lucas et al. (2014) showed that adults were less likely than children to endorse disjunctive causal rules in which multiple cues can produce a single effect than conjunctive causal rules in which single effects are produced by single causes.

Crucially, this idea contrasts sharply with the Bayesian inference account. This alternative mechanism suggests that instead of relying on Bayes' rule, human reasoners supplement traditional associative learning, as defined by the Rescorla-Wagner model, with pre-existing, empirically supported assumptions about how causes tend to behave in the real world. We instantiated this account in the connectionist model presented above by applying a small bit of weight decay to object B following the AB+ event. This is a parameter that allows the connections in the network to shrink by some amount; weight decay can be applied to all weights in the network or, as is the case here, to specific

Table 2
Model fit indices for the connectionist, Bayesian, improved connectionist, and modified RW models.

	Experiment 1		Experiment 2	
	RMSE	R^2	RMSE	R^2
Connectionist model	14.05	.46	9.10	.63
Bayesian Model	5.53	.93	2.89	.98
Improved Connectionist Model	10.45	.65	5.57	.85
Modified RW Model	22.76	.38	31.97	.39

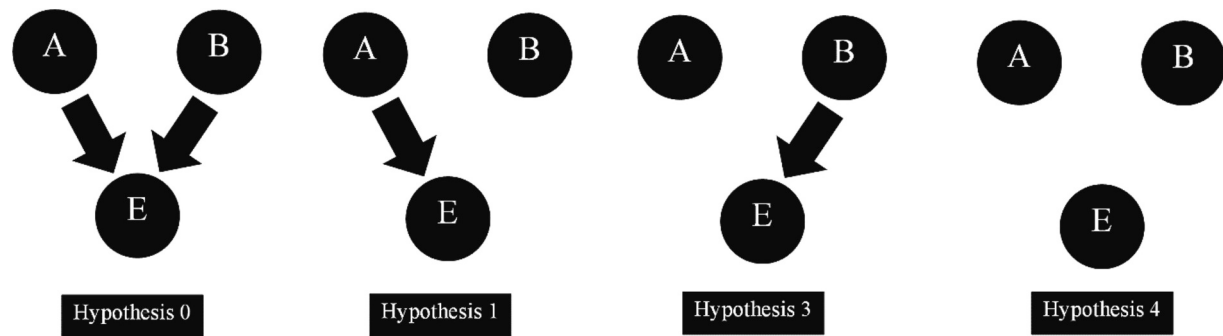


Fig. 3. The four different causal graphical models indicating the possible causal relations for a causal event that involves two objects and one blinket detector. A and B correspond to the two objects that were used on the machine and E indicates the activation of the machine.

connections. In connectionist models, the process works as follows: When the network determines by how much to adjust each weight in order to minimize error (i.e., the difference between what the network sees and what it expects), a fraction of this proposed adjustment is deducted from the weight adjustment itself. This is what enables an associative learning mechanism to capture the “return to baseline” of object B (i.e., the putative backwards blocking effect) in people without the additional assumption that learners are searching through a potentially infinitely large psychological hypothesis space. We found that a weight decay of 0.025 yielded the best model fit to the data across both experiments. The model fit results for this model, as well as the previous models, are shown above in Table 2.

The modified Rescorla-Wagner model. In terms of one final associative learning model, it is possible that these data are best explained by the modified Rescorla-Wagner model (Van Hamme & Wasserman, 1994). Similar to the connectionist model that implemented a version of the traditional RW model, this model assumes that learners should adjust their beliefs about the causal status of a candidate cue by an amount that is proportional—the proportion here is determined by two salience parameters α and β (which capture the salience of the candidate cue and the effect [e.g., the machine's activation], respectively)—to the difference between the outcome that participants observe on a given trial and participants' beliefs about whether a given cue is causal. More formally, on this model, error correction is computed according to the following equations:

$$\Delta V_i = \alpha_i \beta_i (\lambda - \Sigma V_k),$$

and

$$V^{n+1} = V_i + \Delta V_i,$$

where ΔV_i is the change in the associative weight between the candidate cause i and effect e , α is the salience of the candidate cause i , β is the salience of the effect e , V is the current associative strength between the candidate cause i and effect e , and ΣV_k is the summed associative strength between all candidate causes i to k and the effect e . Crucially, in this model α is set to a nonzero and negative value. This is because absent cues are assumed to be negatively correlated with the outcome in the modified RW model. The traditional RW model assumes that absent cues have no salience, and thus α is set to 0. As such, rather than object B maintaining its causal strength between the mid- and post-rating phases, the modified RW model predicts that participants' causal ratings of object B between the mid- and post-rating phases should drop to (or around) 0. In contrast, the traditional RW model predicts that participants' mid- and post-ratings of object B should remain unchanged between the mid- and post-rating phases given that absent objects have no salience. This is the key difference between the modified and traditional RW models.

To be consistent with the connectionist model and to better facilitate model comparison, we applied a sigmoid activation function to the associations of individual objects, and we trained the models for a total of 1000 epochs (i.e., 500 epochs during the AB+ portion of the backwards blocking event, and 500 epochs during the A+ portion of the event). We found that the α and β salience parameters that provided the best fit to the data in Experiment 1 were, respectively, 0.1 and 0.1, whereas the α and β salience parameters that provided the best fit to the data in Experiment 2 were, respectively, 0.2 and 0.2. These model fit results are shown in Table 2.

As shown in Table 2, although the improved connectionist model outperformed the original connectionist model as well as the modified RW model, all three associative learning models were outperformed by the simple Bayesian model. This was true across both experiments. This suggests that Bayesian inference, but not associative learning, underpinned adults' backwards blocking performance.

6. General discussion

The goals of the current study were twofold. The first was to determine whether adults, like children, would engage in backwards blocking reasoning in a task that used the blinket detector design. Specifically, we determined whether participants' ratings of object B—the object that only participated on the machine in combination with object A—changed over the course of the backwards blocking event, and whether this change was impacted by the presence of other “distractor” objects. Although variants of the blinket detector design have been used to evaluate causal reasoning in human adults (e.g., Griffiths et al., 2011), to our knowledge this is the first study to use the blinket detector to examine causal reasoning in human adults. We maintain that assessing backwards blocking in human adults with the blinket detector paradigm can enable cross-study comparisons. Indeed, when different designs are used, it becomes difficult, if not impossible, to isolate the true source of any observed differences. Such differences could be due to the presence or absence of backwards blocking, but they can just as well be due to differences in age, methodology, or some combination of both. The second goal was to determine *how* or by what cognitive mechanism adults reasoned in the backward blocking events in the present experiment. Our specific questions concerned whether Bayesian inference or associative learning underpinned adults' backwards blocking performance.

In this study, adults were introduced to a blinket detector that activated when blinkets (but not when non-blinkets) were placed on its surface. The adults were then given a backwards blocking trial (as well as two other trials, but these trials were of less importance). In this trial, two novel objects, A and B, were placed on the machine, which caused it to activate. Then object A was placed on the machine by itself, which again caused the machine to activate. Participants were then asked to rate whether each object was a blinket before the backwards blocking event, midway through the event, and following the event. Crucially, we were interested in whether participants' beliefs about object B changed across the three rating phases within the backwards blocking event, and whether such beliefs were affected by the presence of distractor objects.

We reasoned that if Bayesian inference underpinned adults' causal ratings of object B, then they should provide reliably lower ratings of object B after the backwards blocking event compared to midway through the event (but this difference should decrease as the probability that any one object is a blinket increases). In contrast, if associative learning based on the traditional RW model subserved adults' causal ratings of object B, then their mid- and post-ratings of object B should not differ. The results were consistent with the predictions of the Bayesian computational model but not the connectionist model. We also explored whether a connectionist model that instantiated adults' evolving beliefs about object B could better explain the data than either the Bayesian model or the original connectionist model, but here too the Bayesian model outperformed both models. These data, in conjunction, indicate that adults continue to engage in backwards blocking reasoning in a blinket detector context and suggest that a Bayesian inference mechanism rather than two different associative learning mechanisms underpinned how adults processed the current causal events.

Although the present data suggest that adults may have relied on Bayesian inference to reason about the present causal events, several potential criticisms are worth addressing. First, one may object that although our results provide clear support for the Bayesian inference account, it is unknown *how* adults arrived at causal judgements that aligned with the predictions of the Bayesian model. One possibility is that participants used a simple form of Bayes rule to determine the hypothesis that was most consistent with the data, but this seems unlikely. This issue relates to what has been referred to as the “search problem” (e.g., Gopnik & Wellman, 2012). This refers to the idea that to determine the “winning” causal hypothesis (i.e., the hypothesis that is most consistent with the observed data) learners must check all possible hypotheses (within a space of potentially infinitely many hypotheses)

against the observed data. However, as the number of candidate causes increases, so too does the size of the underlying psychological hypothesis space; this space grows exponentially with the number of candidate causes. Thus, it seems unlikely that adults are engaging in Bayesian inference. The question remains, then: How did adults arrive at causal judgements that aligned with the Bayesian model predictions? One possibility is that adults initially hold onto a single hypothesis or a small number of hypotheses until the data compel them to consider alternative hypotheses. This shift in thinking can be described as a “win-stay-lose shift” algorithm (Bonawitz et al., 2014). Other possibilities include adults employing a psychological version of simulated annealing (Gopnik et al., 2015), where learners gradually narrow down the hypothesis space, or an algorithm based on the Neurath's ship analogy (Bramley et al., 2017). The idea here is that learners progressively reconstruct a single hypothesis until it is consistent with the data. An alternative possibility—which we adopt here—is that adults engaged in a variant of associative learning, similar to the one employed by the enhanced connectionist model. Although this model was outperformed by the Bayesian model, its overall performance more closely resembled that of the Bayesian model compared to the connectionist model implementing the traditional RW model. As such, such a psychologically tractable mechanism could provide a means by which human adults can approximate Bayesian inference without the need to represent *any* hypothesis, much less an extensive hypothesis space. This proposal should be explored further in future research.

Second, it may be argued that although the current results support the Bayesian model, they are nonetheless of limited usefulness because we did not manipulate base rates. This approach contrasts with some prior studies that did manipulate base rate information (e.g., Griffiths et al., 2011; Sobel et al., 2004). The rationale for not manipulating base rates here was twofold. First, we wanted to determine whether human reasoners defaulted to one or more modes of reasoning in the absence of such manipulations. Second and relatedly, as we argued in the Introduction, when a learner encounters a new causal scenario in the real world—from which existing knowledge cannot be generalized—they often have to reason about such a system anew without knowledge about underlying base rates. In other words, rarely is it the case in the real world that learners are explicitly provided with base rate information from which to make causal decisions. In this way, it can be argued that our decision not to manipulate base rates is more ecologically valid than previous studies that manipulated base rates to determine its effect on backwards blocking reasoning. Critically, our results indicate that even when base rates are not explicitly manipulated, adults' causal inferences continue to be consistent with Bayesian inference. However, we also acknowledge that because we did not explicitly manipulate base rates, we cannot know just how Bayesian adults were in their causal inferences (though again see Griffiths et al., 2011 for just such information).

A final criticism concerns the fact that the present experiments only used three trials per subject. This means that it was impossible to estimate within-subject variability; it cannot be known based on the current results whether participants' causal ratings of object B would change from one three-phase backwards blocking event to the next three-phase backwards blocking event. Although we acknowledge that this is a criticism of the current experiments, it is worth noting that this criticism is not unique to our study. For example, in the study by Sobel and Kirkham (2006)—on which Experiment 2 was loosely based—infants experienced each of three trials exactly once. Likewise, in studies by McCormack et al. (2013), Sobel et al. (2004), and Griffiths et al. (2011), children and adults were asked to provide one causal response for a small number of trials. This means that we were justified in using three trials per participant. Still, it will be important for future research to determine whether the current results change when multiple trials are used for each subject.

7. Concluding remarks

The current experiments were designed to test whether human adults would engage in backwards blocking reasoning for multiple objects when a more accurate measure of backwards blocking reasoning was used—by evaluating whether participants' ratings of object B changed across the backwards blocking trial—and in the context of the widely used (at least with children) blinket detector paradigm. Moreover, these experiments were designed to determine how or by what cognitive mechanism adults reasoned about the current events. The results from the Computational Models section suggested that adults relied on Bayesian inference rather than on two different forms of associative learning. Although it remains a possibility that more sophisticated forms of associative learning could explain the present data, at minimum the current data suggest that Bayesian inference can explain adult causal reasoning.

A promising future direction of this research will be to examine whether human children's backwards blocking performance would match adults' current backwards blocking performance. If one is ultimately interested in providing a developmental account of backwards blocking reasoning for multiple objects, then it will be critical to test children with the current design in future studies. Fortunately, given the simplicity of our study and the fact that participants were provided with one of each trial type, it should be straightforward to use the current design to test backwards blocking reasoning in the presence of distractor objects in children. The simplicity of our design should also allow direct comparisons to be made between this study and future studies on backwards blocking reasoning in human children. Such direct comparisons would have been made much more difficult if the current participants received multiple trials over a much lengthier experimental session. Thus, our simple design will allow these future studies with human children to achieve a balance between task complexity (both in terms of number of trials per subject and the length of the experiment) and accurate measures of backwards blocking reasoning.

In summary, these experiments constitute one of the first systematic attempts to examine backwards blocking reasoning in the context of distractor objects over a series of rating phases. It has been proposed that people use Bayes' rule to reason about causal events and that the putative backwards blocking finding derives from a Bayesian inference mechanism (e.g., Gopnik et al., 2004; Sobel et al., 2004). The experiments reported here are consistent with this conclusion: Not only did adults engage in backwards blocking reasoning, but their inferences were also consistent with the predictions of a very simple Bayesian model.

CRedit authorship contribution statement

Deon T. Benton: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft. **David H. Rakison:** Supervision, Funding acquisition.

Data availability

Data will be made available on request.

Acknowledgement

We thank Kaily Bruch and Marina Selenica for their help with data collection and participant recruitment. We also thank the participants who kindly agreed to participate in the research. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

- Aitken, M. R., Larkin, M. J., & Dickinson, A. (2001). Re-examination of the role of within-compound associations in the retrospective revaluation of causal judgements. *The Quarterly Journal of Experimental Psychology Section B*, 54(1b), 27–51.
- Beckers, T., De Houwer, J., Pineno, O., & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(2), 238.
- Benton, D. T., Rakison, D. H., & Sobel, D. M. (2021). When correlation equals causation: A behavioral and computational account of second-order correlation learning in children. *Journal of Experimental Child Psychology*, 202, 105008.
- Bhat, A. A., Spencer, J. P., & Samuelson, L. K. (2022). Word-object learning via visual exploration in space (WOLVES): A neural process model of cross-situational word learning. *Psychological Review*, 129(4), 640–695.
- Bonawitz, E., Denison, S., Griffiths, T. L., & Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: Sampling in cognitive development. *Trends in Cognitive Sciences*, 18(10), 497–500.
- Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, 124(3), 301–338.
- Buss, A. T., & Spencer, J. P. (2014). The emergent executive: A dynamic field theory of the development of executive function. *Monographs of the Society for Research in Child Development*, 79(2), vii.
- Chapman, G. B. (1991). Trial-order affects cue interaction in contingency judgement. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17, 837–854.
- Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory & Cognition*, 18(5), 537–545.
- Danks, D. (2003). Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology*, 47(2), 109–121.
- De 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.
- Dickinson, A., & Shanks, D. (1985). Animal conditioning and human causality judgment. In L. Nilsson, & T. Archer (Eds.), *Perspectives on learning and memory*. Hillsdale, NJ: Erlbaum.
- Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, 117(3), 227–247.
- Gluck, M. A., & Bower, G. H. (1990). Component and pattern information in adaptive networks. *Journal of Experimental Psychology: General*, 119(1), 105–109.
- Glymour, C. (2000). Bayes nets as psychological models. In F. C. Keil, & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 169–197). The MIT Press.
- Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111(1), 3–32.
- Gopnik, A., Griffiths, T. L., & Lucas, C. G. (2015). When younger learners can be better (or at least more open-minded) than older ones. *Current Directions in Psychological Science*, 24(2), 87–92.
- Gopnik, A., Sobel, D. M., 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., & Tenenbaum, J. B. (2007). Bayesian networks, Bayesian learning and cognitive development. *Developmental science*, 10(3), 281–287.
- Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, 138(6), 1085–1108.
- 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.
- Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, 107(20), 9066–9071.
- Harris, P. L., German, T., & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition*, 61(3), 233–259.
- Kloos, H., & Sloutsky, V. M. (2013). Blocking a redundant cue: What does it say about preschoolers' causal competence? *Developmental Science*, 16(5), 713–727.
- Kloos, H., & Sloutsky, V. M. (2013). Blocking a redundant cue: what does it say about preschoolers' causal competence? *Developmental science*, 16(5), 713–727.
- Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & 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.
- Lovibond, P. F., Been, S., Mitchell, S. J., Bouton, M. E., & Frondhardt, R. (2003). Forward and backward blocking of causal judgment is enhanced by additivity of effect magnitude. *Memory and Cognition*, 29(1), 97.
- Lu, P., Liu, J., & Koestler, D. (2017). pwr2: Power and sample size analysis for one-way and two-way ANOVA Models. R package version 1.0 <https://CRAN.R-project.org/package=pwr2>.
- Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, 131(2), 284–299.
- McCormack, T., Butterfill, S., Hoerl, C., & Burns, P. (2009). Cue competition effects and young children's causal and counterfactual inferences. *Developmental Psychology*, 45(6), 1563–1575.
- McCormack, T., Simms, V., McGourty, J., & Beckers, T. (2013). Blocking in children's causal learning depends on working memory and reasoning abilities. *Journal of Experimental Child Psychology*, 115(3), 562–569.
- Pearl, J. (2000). Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association*, 95(450), 428–431.
- R Core Team. (2021). R: A language and environment for statistical computing. In *R foundation for statistical computing, Vienna, Austria*. URL <https://www.R-project.org/>.
- Rakison, D. H. (2005). A secret agent? How infants learn about the identity of objects in a causal scene. *Journal of Experimental Child Psychology*, 91(4), 271–296.
- Rakison, D. H. (2006). Make the first move: How infants learn about self-propelled objects. *Developmental Psychology*, 42(5), 900.
- Rakison, D. H., Smith, G. T., & Ali, A. (2016). Who is the dynamic duo? How infants learn about the identity of objects in a causal chain. *Developmental Psychology*, 52(3), 355–363.
- 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.
- Schlotmann, A., & Shanks, D. R. (1992). Evidence for a distinction between judged and perceived causality. *The Quarterly Journal of Experimental Psychology*, 44(2), 321–342.
- Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology*, 40(2), 162–176.
- Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science*, 10(3), 322–332.
- 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), 1103–1115.
- Sobel, D. M., & Kirkham, N. Z. (2007). Bayes nets and babies: Infants' developing statistical reasoning abilities and their representation of causal knowledge. *Developmental science*, 10(3), 298–306.
- Sobel, D. M., & Munro, S. E. (2009). Domain generality and specificity in children's causal inference about ambiguous data. *Developmental Psychology*, 45(2), 511–524.
- Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive science*, 28(3), 303–333.
- Spencer, J. P., Ross-Sheehy, S., & Eschman, B. (2022). Testing predictions of a neural process model of visual attention in infancy across competitive and non-competitive contexts. *Infancy*, 27(2), 389–411.
- Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search*. MIT press.
- Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, 27(3), 453–489.
- Stojnić, G., Gandhi, K., Yasuda, S., Lake, B. M., & Dillon, M. R. (2023). Commonsense psychology in human infants and machines. *Cognition*, 235, 105406.
- Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation*, 25(2), 127–151.
- Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, 121(2), 222.
- Widrow, B., & Hoff, M. E. (1960). *Adaptive switching circuits*. Stanford Univ Ca Stanford Electronics Labs.
- Woodward, J. (2004). Counterfactuals and causal explanation. *International Studies in the Philosophy of Science*, 18(1), 41–72.