COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 36 (2012) 1178–1203

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2012.01262.x

Non-Bayesian Inference: Causal Structure Trumps Correlation

Bénédicte Bes,[a] Steven Sloman,[b] Christopher G. Lucas,[c] Éric Raufaste[a]

[a]*Laboratoire CLLE-LTC, Université de Toulouse*
[b]*Cognitive, Linguistic, and Psychological Sciences, Brown University*
[c]*Carnegie Mellon University, Pittsburgh*

Received 21 December 2010; received in revised form 29 February 2012; accepted 29 February 2012

Abstract

The study tests the hypothesis that conditional probability judgments can be influenced by causal links between the target event and the evidence even when the statistical relations among variables are held constant. Three experiments varied the causal structure relating three variables and found that (a) the target event was perceived as more probable when it was linked to evidence by a causal chain than when both variables shared a common cause; (b) predictive chains in which evidence is a cause of the hypothesis gave rise to higher judgments than diagnostic chains in which evidence is an effect of the hypothesis; and (c) direct chains gave rise to higher judgments than indirect chains. A Bayesian learning model was applied to our data but failed to explain them. An explanation-based hypothesis stating that statistical information will affect judgments only to the extent that it changes beliefs about causal structure is consistent with the results.

Keywords: Probability judgment; Causal explanations; Bayesian model

1. Introduction

There is wide consensus that causal beliefs and assessments of probability are closely connected in both philosophy (e.g., Spirtes, Glymour, & Scheines, 1993; Suppes, 1970) and cognitive science (e.g., Cheng, 1997; Rehder, 2009; reviewed in Sloman, 2005). Is one given priority when people make judgments? In the study of cognition, at least two views are possible. One is that people make judgments of probability using prior expectations based on a variety of information sources, of which statements about causal structure are

Correspondence should be sent to Steven Sloman, Cognitive, Linguistic, & Psychological Sciences, Brown University, Box 1821 Providence, RI 02912. E-mail: steven_sloman@brown.edu

just one type. We call this the Bayesian view. On this view, people make judgments by combining any available data with whatever prior beliefs they may have, including those about causal structure. One contrasting view is that judgments are derived directly from knowledge of causal structure rather than from a collection of probabilistic information where structure has no special status. We call this the explanation-based view. On this view, judgment processes operate on the assumption that causal structure is what makes the world go round and observable quantities like probability and correlation are merely reflections of it (e.g., Hume, 1976; Pearl, 2000).

The explanation-based view is a staple in the psychological literature (reviewed in Sloman, 2005). An early proponent was Ajzen (1977), who argued that people use a ''causality heuristic'' when making probability judgments, relying on causal knowledge while neglecting noncausal statistical information except to the extent that it changes their causal beliefs. This causality heuristic implies that people ignore quantitative data concerning a causal relation when they already have qualitative information about it.

A causality heuristic, like all heuristics, is often effective but can lead to systematic errors in some situations. For example, Tversky and Kahneman (1983) showed a conjunction fallacy resulting from a causal relation between the two components of a conjunction. Thus, the statement ''a randomly selected male has had one or more heart attacks'' was judged less likely than ''a randomly selected male has had one or more heart attacks and is over 55 years old.'' In general, an event seems more likely when a potential cause is presented in the conjunction, whereas the conjunction rule states that a conjunction cannot be more probable than one of its constituents. Kahneman and Tversky's result can be explained by the existence of an explanatory relation between the constituents: Having a heart attack can be explained by being more than 55 years old (Fabre, Caverni, & Jungermann, 1995, 1997). Indeed, Crisp and Feeney (2009) found that the strength of the causal connection between constituent events directly affected the magnitude of the causal conjunction fallacy.

Causality has also been investigated in the study of subadditivity (Tversky & Koehler, 1994). Implicit subadditivity refers to the fact that a hypothesis A is judged less likely when its components (A_1 and A_2) are not mentioned than when its description is unpacked into components: $P(A) < P(A_1 \cup A_2)$. For instance, Rottenstreich and Tversky (1997) compared the probability of a packed description ''homicide'' with one unpacked either according to the causal agent ''homicide by an acquaintance or by a stranger'' or according to the time of occurrence ''daytime homicide or nighttime homicide.'' Results indicated more implicit subadditivity in the causal partition. Rottenstreich and Tversky conjectured that a causal partition brings to mind more possibilities than a temporal partition. A causal partition also provides an explanation for the occurrence of the event.

These studies support the idea that people rely on causal explanations when they are making judgments. Further support for this view comes from Pennington and Hastie (1993) who showed that an event is given a higher probability if the evidence is presented in chronological order (rather than random order), enabling the construction of an explanatory story. Some studies have shown how causal explanations are generated and how they affect the probability of a focal scenario. Dougherty, Gettys, and Thomas (1997) investigated the role of mental simulation in judgments of likelihood. First, they

showed that simulating a large number of competing causal scenarios for an outcome diminished the probability of the focal scenario. Second, they found that participants appeared to generate several causal scenarios initially and then rejected the less likely causal scenarios before making their probability judgment. Thus, the probability of a scenario depended on both the number and likelihood of causal scenarios imagined by the participant. Further support for the explanation-based view comes from findings that, when interpreting evidence, explanations tend to dominate. Chapman and Chapman (1969) present a classic demonstration that people observe ''illusory correlations'' that are consistent with their prior beliefs but inconsistent with the data. Brem and Rips (2000) show that explanations take priority over data in argument.

So reliance on causal explanation can lead to neglect of data. As another example of this phenomenon, highlighting causal relations affects the extent to which people neglect base rates in probability judgment. Ajzen (1977) found that probability judgments were influenced by base rates of a target outcome in the population only to the extent that the base rates had causal implications for the object of judgment. Tversky and Kahneman (1982) also found less base-rate neglect with causal than with incidental base rates (however, Sloman, 2005, reports a failure to replicate using one of Tversky and Kahneman's items).

Proponents of the contrasting Bayesian view in the study of causality and judgment include Krynski and Tenenbaum (2007), who applied a causal Bayesian net framework to base-rate neglect, arguing that its advantage over a purely statistical framework is that it explains how judgments are made with limited statistical data. The framework states that people process data in three steps: (a) they construct a causal model; (b) they assign parameters to the variables; and (c) they infer probabilities. Parameters are assumed to be estimated from statistical information provided in the task in conjunction with background knowledge. In studies of base-rate neglect, they found that statistics that map onto parameters of a causal model were used appropriately.

2. Current studies

To compare the Bayesian and explanation-based positions, we ran three experiments that obtained conditional probability judgments with various causal structures while holding statistical information constant. We made sure that the statistical information was highly available and salient. Specifically, we provided participants with two pieces of information: the causal links among a set of variables and statistical information about relationships between their values. We then asked them to judge the probability of one variable given the value of another. The statistical information provided was sufficient to calculate the desired conditional probability.

The Bayesian view has some leeway in what it predicts in this situation because it allows that people might have various prior beliefs about the variables and the strengths of the causal links they are given. However, it does impose some constraints on judgment. For instance, judgments should be closer to the statistical information if there is more of it than if there is less.[1] Another constraint that we discuss in detail below is that, given reasonable

assumptions, judgments associated with common-cause structures must be in between judgments associated with forward and backward chains.

The explanation-based hypothesis proposes that people construct explanations of data and these explanations then serve as the basis of judgment without further regard to the statistics on which they are based. The hypothesis assumes that an explanation is constructed from prior knowledge about causal mechanisms that posits some combination of causes, enablers, disablers, and preventers to describe how the data were generated. This explanation serves as a summary representation of the data but can take a life of its own if the data are not entirely consistent with it. This view suggests that people will make judgments based on qualitative causal structure that encodes explanatory relations and will neglect the original data. This view predicts, like the causality heuristic, that people will be directly influenced by causal structure and statistical information will affect their judgments only to the extent that it changes their beliefs about causal structure. Causal structure will mediate the relation between data and judgment so that different causal beliefs could lead to different judgments even when the underlying statistical support is identical.

We will compare situations where the target event and the conditioning event (hereafter referred to as the evidence) are linked by a causal chain (one is a cause of the other) and situations where the target event and the evidence are not directly linked (they are effects of a common cause). The explanation-based view suggests that the easier it is to construct an explanation, the more influence the explanation will have on judgment. Events that are causally related by the explanation will be perceived as more highly correlated. Explanations of a target event are easier to generate when the event is a cause or an effect of the evidence than when they are both effects of a common cause. When a chain of causation relates the target event and evidence, the target can be explained by a single mechanism that leads from the evidence. But when they share a common cause, two mechanisms are necessary, one from the common cause to the target and the other from the common cause to the evidence. Even if the details of the two mechanisms are identical, each must be considered separately. Because of this difference in ease of explanation, the explanation-based hypothesis predicts that judgments of the conditional probability of the event given the evidence will be higher in the case of a causal chain than in the case of a common cause.

When the evidence and the target event are linked by a causal chain, judging the probability of the event requires an inference from the evidence to the event. This inference can be in a predictive direction, where the evidence is a cause of the target, or in a diagnostic direction, the evidence is an effect of the target. For example, judging the probability that a woman is physically fit given that she participates in a sport would be a predictive inference, whereas judging the probability that a woman participates in a sport given that she is fit would be diagnostic. These two types of inferences are asymmetric: Inferences from effect to cause tend to use more information about alternative causes than inferences from cause to effect (Fernbach, Darlow, & Sloman, 2010, 2011). In that sense, predictive inferences are easier than diagnostic ones. They also take less time (Fernbach et al., 2010). White (2006) described causal asymmetry as the general tendency to overestimate the force exerted by a cause on an effect and to underestimate the corresponding force exerted by an effect on its cause. Tversky and Kahneman (1982) also provide evidence that the probability of an event

is higher when the inference is predictive than when it is diagnostic. Although Fernbach et al. (2011) failed to replicate Tversky and Kahneman's specific effect, they did find that predictive inferences were higher than diagnostic ones in the presence of strong alternative causes of the effect. They also showed that even a normative analysis will more often than not predict that predictive inferences will be higher than diagnostic inferences. Roughly speaking, to the extent that effects have alternative causes, a cause will provide more evidence for its effect than vice versa. For these reasons, we expect predictive questions to lead to higher judgments than diagnostic ones.

In our experiments, we held the correlations among the variables constant. In Experiments 1 and 2, we did so by providing participants with a summary description of the correlation. In Experiment 3, statistical information was implicit by presenting a series of observed events.

3. Experiment 1

This experiment aims to investigate whether causal models can affect probability judgments while statistical information is held constant. More precisely, we will vary the causal models connecting the variables in a scenario. Drawing from the explanation-based account of the role of causal structure, our hypotheses are twofold:

1. The judged probability of one event given another will be higher if there is a causal path from one to the other (whether this path is in a diagnostic or predictive direction). We will compare the case in which the evidence and the target event are linked by a causal chain with the case in which they are both effects of a common cause. Our hypothesis is that the probability of the target event will be lower in the second case because of the absence of a direct causal path between the elements. A third variable will be used to build the causal models but it will not be mentioned in the judgment task.
2. Our second hypothesis deals with the nature of the causal chains. When assessing the probability of the target event, two different types of inferences can be defined depending on the direction of the causal chain. If the evidence is a cause of the event to be judged, the inference is predictive: People have to judge the probability of an effect knowing a cause. But if the evidence is an effect of the event, the inference is diagnostic: People have to judge the probability of a cause knowing an effect. We expect predictive chains to give rise to higher probability judgments than diagnostic chains.

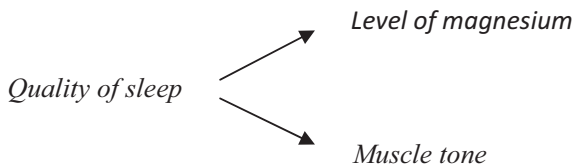
3.1. Participants

A total of 144 students of the University of Toulouse le Mirail participated in this experiment. They were recruited on a voluntary basis in the university library.

3.2. Materials

Participants were tested in French. They were presented with a questionnaire divided into three parts: a presentation of the task, a training scenario, and the experimental scenarios. Table 1 illustrates the organization of a scenario. First, we presented three variables: *A*, *B*, and *C*. We indicated their statistical correlation by saying that in 40% of cases, *A*, *B*, and *C*

Table 1
Organization of the common-cause scenario, Experiment 1

Presentation of 3 variables	Recently, some researchers have revealed the existence of a statistical relation between muscle tone, level of magnesium, and the quality of sleep.
Statistical correlation	In 40% of people, muscle tone, level of magnesium, and the quality of sleep are all high. In 40% of people, muscle tone, level of magnesium, and the quality of sleep are all low. In 20% of people, those variables have different levels: some are high whereas others are low.
Causal model	The researchers found an explanation of the existence of this statistical relation: An increase in the level of magnesium leads to an increase in the quality of sleep. An increase in the level of magnesium leads to an increase in muscle tone.
Diagram	This explanation can be represented by the following diagram: 
Comprehension Questions	With the help of the previous information, please respond to the following questions: The level of magnesium has a direct effect on: ☐ The quality of sleep ☐Muscle tone ☐None of them The quality of sleep has a direct effect on: ☐ Level of magnesium ☐ Muscle tone ☐ None of them Muscle tone has a direct effect on: ☐ The quality of sleep ☐ Level of magnesium ☐ None of them
Evidence	Mary, 35 years-old, has good quality of sleep.
Probability judgment	According to you, what is the probability that Mary has good muscle tone? o—o—o—o—o—o—o—o—o—o—o—o—o—o—o—o—o—o—o—o—o—o—o 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

are low, in 40% of cases, *A*, *B*, and *C* are high, and in 20% of cases, the variables have different levels: Some are high whereas others are low. Then we presented the causal model relating the three variables. We illustrated this model with a diagram and asked some questions to check the understanding of the model. Finally, we presented a situation where one variable was present (such as *A*) and we asked for the probability of another (such as *B*) without saying anything about *C*. The participants had to make their judgments on graduated scales going from 0% to 100%. Materials translated from French to English appear in Appendix A.

3.3. Design

We manipulated the causal model presented. In two conditions, the evidence *A* and the target *B* were related by a causal chain. In one case, *A* was a cause of *B* (a predictive chain) and, in the other, *A* was an effect of *B* (a diagnostic chain). In another condition, *A* and *B* were effects of a common cause *C* (common-cause condition). We also tested a condition with no causal model (control condition). Table 2 displays the four conditions.

The questionnaire was composed of eight different scenarios. Each participant saw each condition two times, with each presentation involving a different scenario. The scenarios presented real-world variables so that people could represent them easily. As we wanted people to believe the causal models, we presented the models as scientific findings. Also, the variables were chosen to minimize participants' prior knowledge about the existence of causal links between them.

3.4. Results

Judgments in the common-cause condition were the lowest followed by the control and diagnostic chains. Judgments in the predictive chain condition were highest (see Fig. 1). A repeated-measures analysis of variance showed a significant effect of condition, $F(3, 429) = 16.79$, $MSE = 212$, $\eta^2_p = .11$, $p < .001$. Planned comparisons were used to test our specific hypotheses. As expected, causal chains gave rise to higher probability judgments than common cause models, $t(143) = 31.02$, $d = 0.56$, $p < .001$. Within causal chains, predictive chains gave rise to higher probability judgments than diagnostic chains, $t(143) = -2.79$, $d = 0.23$, $p = .01$. The control condition gave higher judgments than the common-cause condition $t(143) = 4.49$, $d = 0.37$, $p < .00$, but lower than predictive chains $t(143) = -2.45$, $d = 0.20$, $p = .02$.

Table 2
Causal models presented in each experimental condition, Experiment 1

Experimental condition	Causal model
Predictive chain	$A \rightarrow C \rightarrow B$
Diagnostic chain	$B \rightarrow C \rightarrow A$
Common cause	$A \leftarrow C \rightarrow B$
Control condition	No causal model

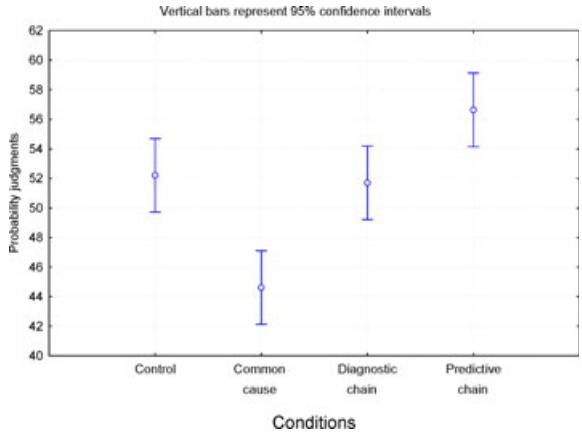


Fig. 1. Mean probability judgments with 95% confidence intervals, as a function of causal models, Experiment 1.

Further analyses were conducted in order to investigate individual differences. We used SPSS18's (IBM SPSS Inc., Chicago, IL, USA) two-step classification procedure to sort participants according to the deviations from their own base-level judgment (computed by intraindividual averaging over the four conditions). In the preclustering step of the procedure, individuals are arranged into subclusters using a clustering feature tree (Zhang, Ramakrishnon, & Livny, 1996). In the second step, subclusters are grouped into clusters using a hierarchical method. The target number of clusters is automatically selected using the Bayesian Information Criterion computed over the models (see SPSS, Inc., 2001, for more details on the algorithms). Three groups emerged, hereafter referred to as Cluster 1 (n = 58), Cluster 2 (n = 57), and Cluster 3 (n = 29). Overall, the three clusters did not significantly differ in the size of the deviation but did in the patterns of the deviations, as shown by the significant Cluster × Condition interaction, $F(6, 423) = 40.91$, $MSE = 136.1$, $\eta^2_p = .37$ (see Fig. 2). Within each cluster, the conditions were rated differently,

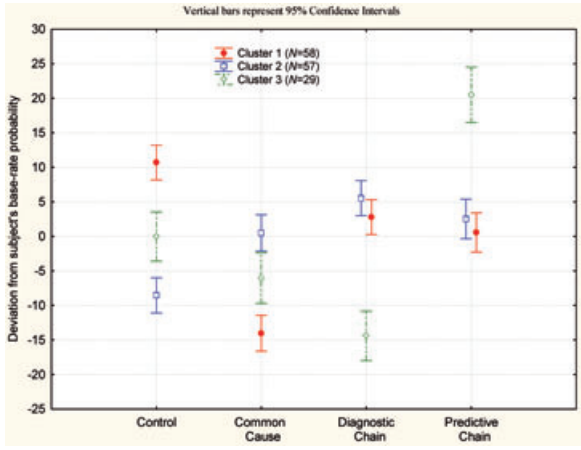


Fig. 2. Mean probability judgments with 95% confidence intervals, as a function of experimental conditions and clusters, Experiment 1.

$F(3, 171) = 43.5$, $MSE = 141.8$, $\eta^2_p = .43$; $F(3, 168) = 16.4$, $MSE = 127.7$, $\eta^2_p = .23$; $F(3, 84) = 45.5$, $MSE = 141.3$, $\eta^2_p = .62$; all $p < .01$. Bonferroni post hoc analyses were used to characterize the differences among patterns. Cluster 1 participants rated the common cause lower than any other condition (all $p < .01$, all Cohen's $d > 0.62$), but the two chain types were not significantly different ($d = 0.18$). In Cluster 2, the three target structures were not rated differently (all pairwise comparisons *ns*, all $d < 0.31$), but the control condition was rated below all other conditions ($p < .01$, all $d > 0.53$). In Cluster 3, predictive chains were rated higher than all other conditions ($p < .01$, all $d > 0.95$), and diagnostic chains were rated the lowest, significantly lower than the control ($p < .01$, $d = 1.15$), marginally significantly lower than the common cause ($p = .053$, $d = 0.57$). Overall, probability judgments were influenced by causal models in 65% of the participants (Clusters 1 and 3).

3.5. Conclusion

Experiment 1 showed that participants could be influenced by the causal relations between the evidence and the event to be judged despite a constant statistical relation between them. Results are consistent with the explanation-based hypotheses. Indeed, probability judgments seemed to depend on the ease of constructing causal explanations. The common-cause condition gave the lowest judgments. This may have occurred because an event is perceived to provide less evidential support in the absence of a direct causal path to the target. Predictive chains gave higher judgments than diagnostic chains, providing more evidence of a causal asymmetry in judgment. One possibility is that judgments in the control condition were derived directly from correlations because no causal explanation was available. If so, the conditional probability judgments are surprisingly low. The most reasonable estimate of the conditional probabilities from the data is around 87%,[2] whereas the mean judgment was only 52%. Another possibility is that people inferred their own causal models from the cover stories. In that case, the mean judgment could reflect the average of a variety of different assumed causal models.

Our clustering solution indicates that a plurality of participants conform to this general pattern (Cluster 1). Another large group was not affected by causal structure (Cluster 2) and could have focused entirely on the correlational information. A third smaller group (Cluster 3) conformed to Bayesian prescriptions on the assumption that they consistently treated predictive links as stronger than diagnostic links.

4. Experiment 2

In this experiment, we attempted to replicate the results of Experiment 1 and extend them by investigating the effect of causal proximity in judgment. Does the presence or absence of an intermediate variable between the evidence and target events affect judgment?

In Experiment 1, when the two variables were related by a causal chain, a third variable was introduced as a mediating variable. This mediating variable was only used to describe the causal model but nothing was said about it in the judgment task. We do not know whether people thought it was present or absent. If they apply a principle of indifference

rather than inferring its value based on the evidence event, they may conclude that it has a 50% of chance of being present. They may also have thought it was absent because nothing was said about it. Either way, this may have increased their judgments and so, putting this variable C at the end of the chain ($A \rightarrow B \rightarrow C$) rather than between the evidence and the uncertain event ($A \rightarrow C \rightarrow B$) may significantly enhance the perceived probability of the uncertain event given the evidence $P(B|A)$. Experiment 2 thus included the conditions of Experiment 1 and also contrasted probability judgments of indirect and direct causal chains. As in Experiment 1, we predict higher judgments in the chain conditions (because the explanations are easier to generate) and we predict the causal asymmetry effect.

4.1. Participants

A total of 180 students of the University of Toulouse le Mirail participated in this experiment. They were recruited on a voluntary basis in the university library.

4.2. Materials and procedure

We used exactly the same procedure as in Experiment 1. The questionnaires had the same structure.

4.3. Design

In addition to the four conditions of Experiment 1, two conditions were created: diagnostic and predictive direct chains. Table 3 summarizes the six conditions. The questionnaires presented six different scenarios so that each participant was in all conditions, each with a different scenario.

4.4. Results

Results show an effect of causal structure, $F(5, 895) = 7.51$, $MSE = 320$, $\eta^2_p = .04$, $p < .001$ (see Fig. 3). Planned comparisons were used to test our specific hypotheses. The findings of Experiment 1 fully replicated. Causal chains led to higher judgments than common-cause models, $t(179) = -3.10$, $d = 0.23$, $p=.002$, and predictive chains led to

Table 3
Causal models presented in each experimental condition, Experiment 2

Experimental condition	Causal model
Predictive direct chain	$A \rightarrow B \rightarrow C$
Diagnostic direct chain	$B \rightarrow A \rightarrow C$
Predictive indirect chain	$A \rightarrow C \rightarrow B$
Diagnostic indirect chain	$B \rightarrow C \rightarrow A$
Common cause	$A \leftarrow C \rightarrow B$
Control	No causal model

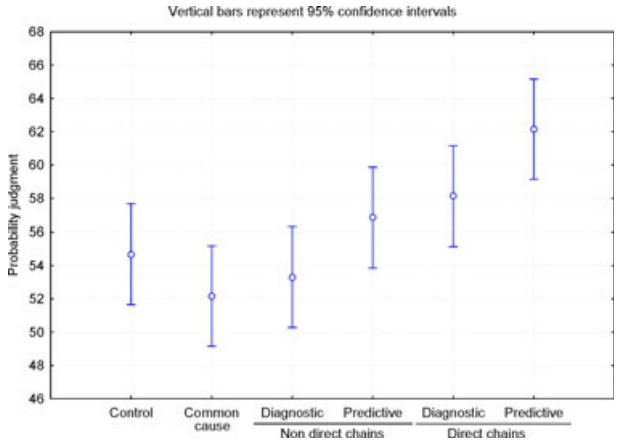


Fig. 3. Mean probability judgments with 95% confidence intervals, as a function of causal models, Experiment 2.

higher judgments than diagnostic chains, $t(179) = -2.87$, $d = 0.21$, $p = .01$. As expected, direct chains led to higher judgments than nondirect chains, $t(179) = -3.87$, $d = 0.29$, $p < .001$. The control condition differed significantly from just one condition, the direct predictive chain, $t(179) = -4.19$, $d = 0.31$, $p = .000$.

The existence of individual differences in participants was assessed using a two-step classification according to the deviations from their mean judgments over the six conditions as in Experiment 1. The procedure resulted in two groups, Clusters 1 and 2, that did not differ in their means but in the pattern of judgments over conditions, as shown by the significant Cluster × Condition interaction, $F(5, 890) = 32.5$, $MSE = 272$, $\eta^2_p = .15$ (see Fig. 4). Within each cluster, the conditions were rated differently, $F(5, 320) = 24.1$, $MSE = 434$,

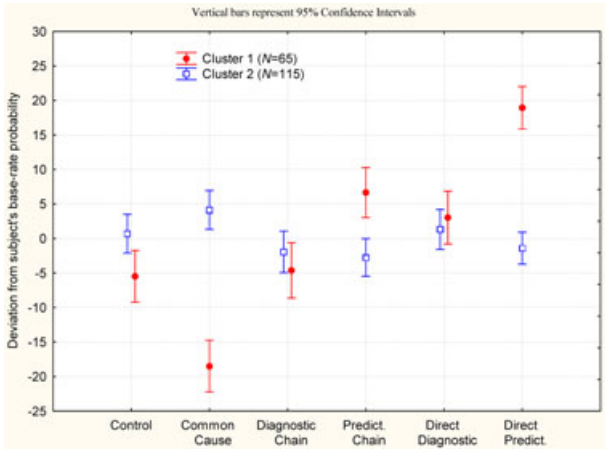


Fig. 4. Mean probability judgments with 95% confidence intervals, as a function of experimental conditions and clusters, Experiment 2.

$\eta^2_p = .27$; $F(5, 570) = 4.16$, $MSE = 180$, $\eta^2_p = .04$; all $p < .01$. As in Experiment 1, Bonferroni post hoc tests showed that the predictive and diagnostic directions of inference were significantly different in Cluster 1—predictive inference led to higher judgments both when comparing chains ($p = .03$, $d = 0.58$) or direct causes ($p < .01$, $d = 0.71$). In contrast, in Cluster 2 participants, diagnostic and predictive inferences were not significantly different, whether direct or chain inferences were considered (all $d < 0.17$).

Thus, as in Experiment 1, one group of participants (Cluster 1, 36% of all participants) was influenced by the direction of the inference. However, in Experiment 2, the other group was still sensitive to the global structure, that is, common cause above causal chains ($p < .01$, all $d > 0.34$) and common cause above direct predictive ($p = .027$, $d = 0.41$). The third rating pattern observed in Experiment 1 did not appear in Experiment 2.

4.5. Conclusion

In addition to the replication of previous findings, this experiment revealed that the absence of an intermediate variable between the target event and the evidence led to an increase in perceived probability. However, one could argue that our way of presenting the statistical information was not precise enough. Indeed, the statement ''In 20% of cases, some variables are high whereas some are low'' refers to six different cases and we did not specify that each case has the same probability. When learning the causal models, people may have overestimated the probability of some of the six cases. Another explanation for the results is that if people combined statistical evidence with their prior beliefs, the participants in Experiment 1 might have treated the statements about the evidence as very sparse data and thereby changed their beliefs very little. Experiment 3 addresses these issues.

5. Experiment 3

In this experiment, to eliminate any ambiguity about the data being presented, we used a different mode of presentation of the correlations among the three variables. Instead of expressing them verbally, we presented a sample of observations in which the level of each variable was indicated. In 40% of observations, the three variables were low; in 40% of observations, they were high; in 20% of cases some were high and others low. Therefore, the correlations were identical to those in the previous experiments. The advantage of this method is that all cases are explicitly presented so that there cannot be any misinterpretation. The disadvantage is that it greatly increases demands on memory.

For half of the participants, the series of observations was long (60 observations), and for the other half, the series was short (5 observations). By comparing these conditions, we can measure how much use is made of the observations to update beliefs. To the degree that participants did update their beliefs by observing the data, the long series should have more influence than the short series. Except for how the data were presented, this experiment was identical to Experiment 2. Table 4 indicates how long and short series were constructed (''+'' = high level/''−'' = low level). In the long series, 60 observations were presented:

Table 4
Composition of series of observations, Experiment 3

Types of observations	Number in long series	Number in short series	% of all cases
A+ *B*+ *C*+	24	2	40%
A− *B*− *C*−	24	2	40%
A+ *B*+ *C*−	2	1 at random	20%
A+ *B*− *C*+	2		
A+ *B*− *C*−	2		
A− *B*+ *C*+	2		
A− *B*+ *C*−	2		
A− *B*− *C*+	2		
Total	60	5	100%

on 24 (40%) the three variables had high levels, on 24 (40%) the three variables had low levels, and on 12 (20%) some variables were low whereas others were high (the six possibilities were presented two times each). In short series, five observations were presented: on two (40%) the three variables had high levels, on two (40%) the three variables had low levels, and on one (20%) some variables were low whereas others were high (one of the six cases was presented at random).

5.1. Participants

A total of 120 students of the University of Toulouse le Mirail participated in this experiment. They were recruited on a voluntary basis in the university library. One of them was chosen by drawing lots and received a gift.

5.2. Materials and procedure

This experiment consisted of a questionnaire displayed on a computer screen. The scenarios had the same organization as previously. The correlations were presented via a series of observations. Each observation was displayed on a separate screen. For each one, the names of the variables and their levels (high/low) were displayed. People could watch the observation as long as they wanted and then had to click on a button to go to the next observation. Fig. 5 presents an example observation. The program also enabled us to check automatically whether people understood the causal models, by asking questions about how the variables are causally related. Participants were not allowed to move on to the following screen if their answers did not fit the causal models previously presented.

5.3. Design

This experiment consisted of the same six conditions as Experiment 2. Each participant was exposed to each of the six conditions using a different scenario.

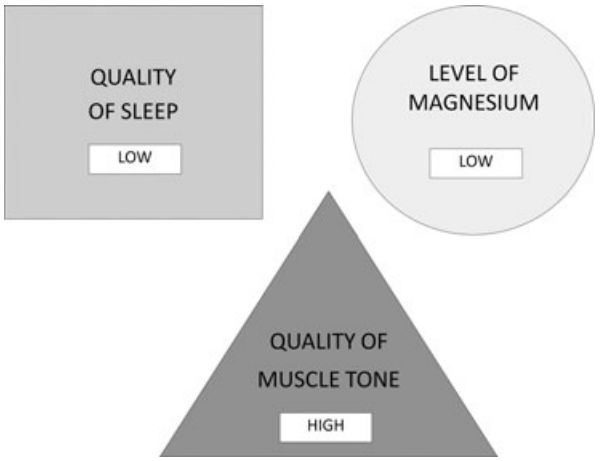


Fig. 5. Example of an observation, Experiment 3.

5.4. Results

As shown in Fig. 6, causal models had an effect on probability judgment, $F(5, 590) = 16.98$, $MSE = 403$, $\eta^2_p = .12$, $p < .001$, and did not interact with the length of the series, $F(5, 590) = 1.0$, $MSE = 403$, $\eta^2_p = .01$, $p = .41$. Results indicate no effect of the length of the series of observations at all $F(1, 118) = .05$, $MSE = 939$, $\eta^2_p = .00$, $p = .83$. Planned comparisons were used to test our specific hypotheses. Causal chain models led to higher judgments than common-cause models, $t(119) = -6.38$, $d = 0.58$, $p < .001$. Predictive chains led to higher judgments than diagnostic chains, $t(119) = 26.86$, $d = 2.45$, $p < .001$. Direct chains led to a higher probability than indirect chains, $t(119) = -3.47$, $d = 0.32$, $p < .001$. The control condition showed higher ratings than the common-cause condition, $t(119) = 3.03$, $d = 0.28$, $p = .003$, but lower than predictive direct, $t(119) = -5.95$, $d = 0.54$, $p < .001$, or indirect chains, $t(119) = -5.90$, $d = 0.54$, $p = .01$.

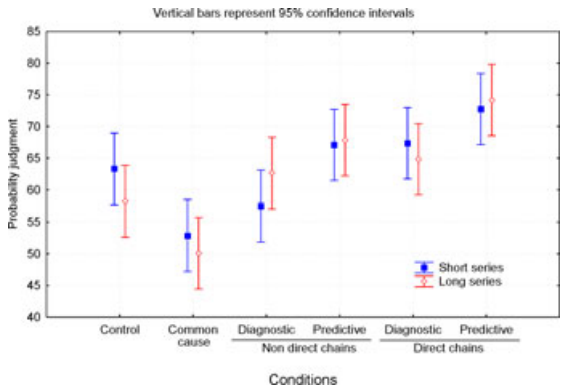


Fig. 6. Mean probability judgments with 95% confidence intervals, as a function of causal models, Experiment 3.

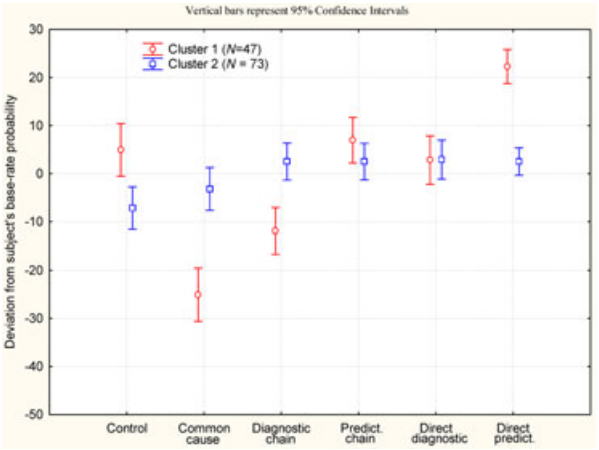


Fig. 7. Mean probability judgments with 95% confidence intervals, as a function of experimental conditions and clusters, Experiment 3.

As in previous experiments, analyses confirmed the existence of individual differences. Using the same two-step classification procedure as before, two different groups emerged, Cluster 1 (n = 47) and Cluster 2 (n = 73). The two clusters were again not differentiated by their mean values but by the patterns of judgments over the conditions, as shown by the Condition × Cluster interaction, $F(5, 590) = 20.5$, $MSE = 346$, $\eta^2_p = .15$, $p < .01$ (see Fig. 7).

Within each cluster, the conditions were rated differently, $F(5, 230) = 24.1$, $MSE = 525$, $\eta^2_p = .15$; $F(5, 360) = 5.57$, $MSE = 232$, $\eta^2_p = .07$; all $p < .01$. As in Experiment 1, Bonferroni post hoc tests showed that predictive inference led to higher judgments both when comparing direct and indirect chains in Cluster 1 ($p < .01$, $d > 0.72$). In contrast, in Cluster 2 participants, diagnostic and predictive inferences were not significantly different whether direct or indirect ($d < 0.03$).

Thus, as in Experiment 1 and 2, one group of participants was influenced by the direction of the inference (39% of the participants), whereas another group was sensitive only to the global structure, common cause versus causal chain (61% of all participants). The third rating pattern observed in Experiment 1 did not appear in Experiment 3.

5.5. Discussion

Experiment 3 replicated the pattern of results of Experiments 1 and 2 implying that the effect of causal structure is not sensitive to the presentation format of the data: verbal summary (Experiments 1 and 2) or a series of events (Experiment 3). Moreover, the number of data points presented did not have a significant effect on probability judgments.

In Experiments 1–3, judgments were lower than we expected based on the data presented. One explanation for this is that participants assumed that C was low when it was not

mentioned in a statement saying A was high. In this case, the probability that B was high given that A was high and C low was 50%. Similarly, in some cases, people may have inferred that C was high. For example, if nothing was said about the quality of sleep, people may have inferred that it was normal (high). To examine these possibilities, we replicated Experiment 3 with a group of 30 people in which we mentioned that C was unknown and could be either low or high. Results indicated no significant difference with judgments obtained in Experiment 3.

These three experiments support the idea that probability judgments are strongly influenced by causal models. More specifically, the easier causal explanations are to construct, the more they reduce uncertainty in the relations among their constituent events. In addition, they replicate previous findings of causal asymmetry (Fernbach et al., 2011).

A question that remains open is to determine whether these judgments can be fit by a rational model. A common normative standard relating causal structure and covariational data uses causal graphical models to describe the structure of causal relationships (Pearl, 2000; Spirtes et al., 1993). Under this account, causal structure is understood to explain statistical relationships between causes and their effects, so that evidence from a learner's observations and actions can be used along with other kinds of information to recover underlying causal relationships. In cases where explicit information about causal structure is available, it is integrated with covariation evidence to give a detailed picture of the underlying structure, which can be used to provide conditional probabilities.

There are several ways by which causal graphical models can be used to understand causal learning. We will focus on the Bayesian approach as it provides clear prescriptions for how prior knowledge and evidence should be combined and has been used extensively to understand causal learning in humans. The Bayesian perspective on causal learning posits that learners use prior knowledge or beliefs combined with evidence—typically observations of events or the results of interventions—to make inferences about variables that are not directly observable, such as what causal relations are present, or what parameters or causal strengths govern a causal relationship that is known to exist (e.g., Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008). We will examine the predictions that follow given only weak constraints, specifically that causal relations are generative, as is implied by our experimental cover stories.

Each of the three causal structures can be expressed as a causal graphical model, in which edges run from causes to their direct effects (see Fig. 8).

Given priors—assumptions about the probable values of parameters w determining marginal and conditional probabilities before any data are observed—such a model yields predictions corresponding to the judgments participants were asked to make, including the conditional probabilities $P(A|B)$ and $P(B|A)$ for the chain structure $A \rightarrow C \rightarrow B$ and $P(B|A)$ for the common-cause structure $A \leftarrow C \rightarrow B$.

More precisely, given a data set $\boldsymbol{d}$ composed of observations of the values of the three variables A, B, and C and a causal structure s, the probability of B given A is

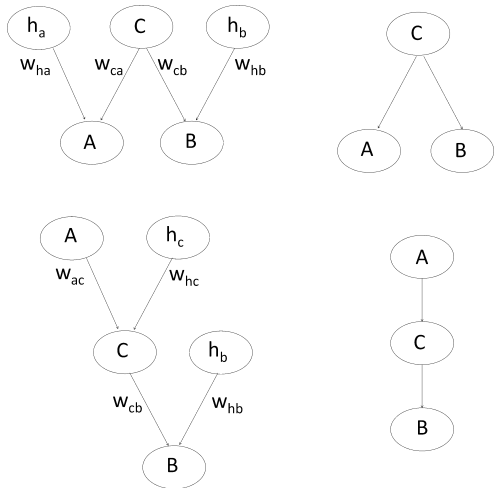


Fig. 8. Chain and common-cause structures.

$$P(B = 1|A = 1, \boldsymbol{d}, s) = \int_w P(B = 1|\boldsymbol{w}, A = 1, s)p(\boldsymbol{w}|d, s)d\boldsymbol{w}.$$

P(B = 1|w, A = 1,s) takes one of three simple forms, depending on the causal structure, as described later. The posterior distribution of the parameters, $p(\boldsymbol{w}|\boldsymbol{d},s)$, can be recovered using Bayes' rule:

$$p(\boldsymbol{w}|\boldsymbol{d}, s) \propto P(\boldsymbol{d}, s|\boldsymbol{w})p(\boldsymbol{w}, s),$$

where $P(\boldsymbol{d},s|\boldsymbol{w}) = P(\boldsymbol{d}|\boldsymbol{w},s)P(s|\boldsymbol{w})$, and $P(s|\boldsymbol{w}) = P(s)$, so likelihood is determined by $P(\boldsymbol{d}|\boldsymbol{w},s)$, or the probability of the observed data given a causal graphical model with parameters $\boldsymbol{w}$.

5.5.1. Evaluating causal graphical models

In this section, we will demonstrate that, given the data and causal structures that participants saw in Experiments 1–3, causal graphical models with generative parameterizations—that is, in which effects are more likely in the presence of their causes—make predictions that are systematically inconsistent with the judgments obtained in our experiments.

Recall that our experiments elicited three kinds of probability judgments: predictive judgments in chains, or $P(B|A)$ when A causes C and C causes B; diagnostic judgments in chains, or $P(A|B)$ under the same relation; and common-cause judgments, or $P(B|A)$ when C causes both A and B. In general, people offered lower probabilities in the common-cause cases than in the other two, and we will show that a causal graphical model with generative causes cannot fit this pattern given the data that participants saw.

We will show first that common-cause judgments always fall between diagnostic and causal chain judgments for the same parameters in three-node causal graphical models,

and that the parameters are the same given the data our participants saw, subject to weak assumptions. More formally, we will show that if P_{chain} is the causal chain probability, P_{common} is the common-cause probability, and $P_{\text{diagnostic}}$ is the diagnostic probability estimate, then for the data that participants saw, regardless of what prior one chooses, we can express the relevant probabilities as follows: the probability that an effect occurs given its cause does not occur is π_0, the probability that an effect occurs given its cause occurs is π_1, and the probability that a variable without an observed cause occurs is r. These probabilities cover all possible parameterizations for causal graphical models when each effect has at most one cause, as is the case here. Using this notation, the conditional probabilities of interest are

$$P_{\text{chain}} = \pi_0 - \pi_0\pi_1 + \pi_1^2,$$

$$P_{\text{diagnostic}} = \frac{r(\pi_1^2 + \pi_0 - \pi_0\pi_1)}{r(\pi_1 - \pi_0)^2 + \pi_0(\pi_1 - \pi_0 + 1)},$$

$$P_{\text{common}} = \frac{r(\pi_1^2 - \pi_0^2) + \pi_0^2}{r(\pi_1 - \pi_0) + \pi_0},$$

We show in Appendix B that this relationship between probabilities is true whenever $\pi_0 < \pi_1$. In other words, it is true if an effect is more likely given its cause than in the absence of its cause. This result is only applicable to our experiments if the distributions of r, π_0, and π_1 are the same for all three structures given the same data and priors. This equality does hold, as can be shown by noting that the posterior distributions of r, π_0, and π_1, for all variables and causal relations depend only on a prior and a likelihood term. The likelihood term is driven by data that are identical across the three causal structures and all pairs of variables, and the prior should be insensitive to the causal structure. Specifically, $P(r|\text{data})$ is determined by the number of times the corresponding variable takes high and low values, which is equal across the different structures, and π_0 and π_1 depend on the rates of different values for cause and effect pairs, which is also equal across different structures. See Appendix B for details. The consequence of these results is that Bayesian inference on generative causal graphs cannot explain the human tendency to assign low probabilities to P_{common} in our experiments.

5.6. Conclusion

This mismatch poses a challenge to the idea that Bayesian inference applied to causal graphical models constitutes a complete model of causal inference in humans. One answer to that challenge is that the participants' expectations about the rates of failure and hidden

causes depend on the cover stories in such a way that they are not independent of the causal structure. Although this possibility cannot be entirely excluded, a hypothesis of that form could explain a wide range of different results, and given the robustness of the effect across cover stories, it does not seem likely. Another answer is that participants are making inferences at a more abstract level than causal graphical models, and learning, for instance, about the category membership of the variables at hand. This may be true, but at least one version of that explanation—that variables that share some causal roles are likely to share others (as might be suggested by the infinite relational block model, e.g., Kemp, Tenenbaum, Griffiths, Yamada, & Ueda, 2006)—would predict *higher* conditional probability judgments in the common-cause scenario. A third answer is that the parameterizations used here were not appropriate. Given that our analytical results apply to any strictly generative parameterization, and that our cover stories indicated that all causes were generative, this seems unlikely. Absent a plausible and parsimonious computational-level explanation for the judgments in Experiment 3, it may be necessary to turn more attention to the time and memory constraints under which people operate when making causal inferences, and, by extension, revisit models that emphasize the processes and representations that people use. This is what the explanation-based approach tries to do.

6. General discussion

In this study, three experiments showed that causal models had a direct effect on probability judgment. More precisely, changing the causal links between hypotheses and evidence changed the perceived probability of a target event. Despite identical correlations between the variables, results indicated higher conditional probability judgments for causal chains than for common-cause structures, higher for predictive than diagnostic chains, and higher for direct than indirect chains. These results obtained whether data were presented verbally or by showing a series of observations. We conclude that probability judgment in our paradigm was largely determined by causal explanations.

In order to evaluate the possibility that the role of the causal models was to introduce prior beliefs into the inference process, we considered Bayesian inference applied to causal graphical models, the implicit or explicit foundation of many normative models of causal inference (Cheng, 1997; Griffiths & Tenenbaum, 2005; Lu et al., 2008), and which incorporated both hypothetical prior knowledge and the data to generate predictions. We showed that such a model predicts that probability judgments in our common-cause condition should have been between those in our predictive and diagnostic conditions. This is not what we observed. Instead, judgments were consistently lower in the common-cause condition. Although the possibility exists that some other rational analysis could explain our findings, we believe they suggest that, when both data and causal beliefs are available, data may well influence causal beliefs, but it is causal beliefs that determine judgment, and data play no further role.

The fact that judgments in the common-cause case were so low was not anticipated by any previous account of causal inference. This effect seems fairly reliable in the sense that the cluster where this effect was most salient—namely Cluster 1, represented 40% of the

participants in Experiment 1, 36% in Experiment 2, and 39% in Experiment 3. When the two variables share a common cause, participants have to make two types of inference: one diagnostic and one predictive. When the variables are linked by a nondirect diagnostic chain, participants have to make two diagnostic inferences. The fact that predictive inferences are easier to draw than diagnostic inferences (Fernbach et al., 2011; White, 2006) would lead to the expectation of higher judgments with a common cause than with indirect diagnostic chains. However, our results indicate the opposite pattern. Perhaps the absence of a direct causal path from evidence to hypothesis in either the predictive or diagnostic direction made it difficult for participants to imagine how to update belief from evidence to hypothesis and the resulting confusion led to lower judged probabilities. Even if the correlation between the two variables was high, it may have been neglected because it did not signify a causal pathway. Another possibility is that an extra cognitive cost is imposed for changing the inference direction while following the path from the evidence to the target. When making the judgment, such increased difficulty could lower the final estimated probability.

On the basis of three experiments, we conclude that people rely on causal explanations to make their judgments. A Bayesian learning model was tested to try to explain how people updated their beliefs, but its predictions were inconsistent with our pattern of results. People are known to be sensitive to causal structure when making decisions. This has long been known by philosophers like Nozick (1993) who proposed a causally based decision theory that inspired a more psychological proposal by Hagmayer and Sloman (2009; Sloman & Hagmayer, 2006). Hagmayer and Sloman propose that people use causal structure along with a representation of intervention to work out the likelihood of outcomes when considering options. They report multiple supportive experiments, although most of the experiments focus on qualitative predictions of the theory. Overall, there is good reason to believe that people excel at working out consequences of actions and events using qualitative causal reasoning. People's ability to update their causal beliefs and work out likelihoods with quantitative precision is more suspect.

According to Ajzen (1977), people focus on qualitative data and neglect quantitative data. Indeed, causal data are simple and easy to use. This idea is supported by the main results of our experiments and we agree with Ajzen that people rely on causal explanations when judging the probability of an event. However, he suggests that statistical data can be used if no qualitative data are present. In our experiments, this occurred in the control conditions in which participants were not presented with a causal model. Results indicated that people underestimated the conditional probabilities either because they neglected the data or misused them. Ajzen proposed that statistical data will be considered if they have a causal frame. This condition is satisfied when the correlation is high and variables are linked by a direct causal chain. In this condition, we found that people underestimated raw probabilities too but less than in the other conditions. In sum, we did not find cases where statistical data were used properly but judged probabilities were close to raw probabilities when the data presented were easily explained by the causal structure.

Probability judgments in our experiments were consistently low. In the first three experiments, the probability of *B* knowing *A* was 87% based on the statistical information alone, but judgments were around 50–60%. People might have focused on presented cases where *A* and *B* had different values and therefore perceived correlations as being lower than they actually were. Another possibility is that people may have found it hard to compute a conditional probability. For example, it is possible that they estimated the ratio of the number of cases where *A* and *B* were high to the total number of cases. Participants in these experiments may also lack mathematical skills. Indeed, our experimental participants were students in the humanities, who may privilege qualitative over quantitative reasoning.

7. Conclusion

The simplest explanation for our results is that people rely on causal explanations to make their judgments and, under the conditions of our experiments, neglect covariational data. Neglect of covariational data has been reported in the literature on psychodiagnosis (Chapman & Chapman, 1969) and on argument strength in discourse (Brem & Rips, 2000). People also neglect covariational data when learning causal structure (Lagnado, Waldmann, Hagmayer, & Sloman, 2007) perhaps because they focus too much on local computations, neglecting global properties of the distribution of data (Fernbach & Sloman, 2009). In learning, several other cues to causal structure are available: temporal order and timing, spatial contact, instruction, and so on. In judgment, the alternative to appealing to covariation is to appeal to knowledge about the antecedents or consequents of the object of judgment, a causal explanation.

Notes

1. Strictly speaking, it is possible to construct a published Bayesian model that is insensitive to the number of data points, but for most patterns of data—including those that we give in our experiments—this requires unusual priors that are not used by any Bayesian account of causal inference.
2. The correlation data indicate the following: $P(A \cup B \cup C) = 40\%$, $P(\neg A \cup \neg B \cup \neg C) = 40\%$, $P(A \cup B \cup \neg C) = 3.33\%$, $P(A \cup \neg B \cup \neg C) = 3.33\%$, $P(A \cup \neg B \cup C) = 3.33\%$, $P(\neg A \cup B \cup \neg C) = 3.33\%$, $P(\neg A \cup \neg B \cup \neg C) = 3.33\%$, $P(\neg A \cup \neg B \cup C) = 3.33\%$. Therefore, *A* has a probability of 50% and the co-occurence of *A* and *B* has a probability of 43.33%. Thus, the conditional probability of *B* knowing *A* is $43.33/50 = 86.66\%$.

References

Ajzen, I. (1977). Intuitive theories of events and effects of base-rate information on prediction. *Journal of Personality and Social Psychology*, *35*, 303–314.

Brem, S. K., & Rips, L. J. (2000). Explanation and evidence in informal argument. *Cognitive Science*, *24*, 573–604.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, *74*, 271–280.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367–405.

Crisp, A. K., & Feeney, A. (2009). Causal conjunction fallacies: The roles of causal strength and mental resources. *Quarterly Journal of Experimental Psychology*, *12*, 2320–2337.

Dougherty, M. R. P., Gettys, C. F., & Thomas, R. P. (1997). The role of mental simulation in judgments of likelihood. *Organizational Behavior and Human Decision Processes*, *70*, 135–148.

Fabre, J.-M., Caverni, J.-P., & Jungermann, H. (1995). Causality does influence conjunctive probability judgments if context and design allow for it. *Organizational Behavior and Human Decision Processes*, *63*, 1–5.

Fabre, J.-M., Caverni, J.-P., & Jungermann, H. (1997). Effects of event probability and causality on the conjunction fallacy. *Swiss Journal of Psychology*, *56*, 106–111.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science.*, *21*(3), 329–336.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experiment Psychology: General*, *140*, 168–185.

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 678–693.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354–384.

Hagmayer, Y., & Sloman, S. A. (2009). Decision makers conceive of themselves as interveners, not observers. *Journal of Experimental Psychology: General*, *138*, 22–38.

Hume, D. (1976). *An enquiry concerning human understanding*. Indianapolis, IN: Hackett Publishing Company.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006, July 16-20). Learning systems of concepts with an infinite relational model. 21st National Conference on Artificial Intelligence, Boston, MA.

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, *136*, 430–450.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154–172). Oxford, England: Oxford University Press.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*, 955–984.

Nozick, R. (1993). *The nature of rationality*. Princeton, NJ: Princeton University Press.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, England: Cambridge University Press.

Pennington, N., & Hastie, R. (1993). Reasoning in explanation-based decision-making. *Cognition*, *49*, 123–163.

Rehder, B. (2009). Causal-based property generalization. *Cognitive Science*, *33*, 301–343.

Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review*, *104*, 406–415.

Sloman, S. A. (2005). *Causal models: How we think about the world and its alternatives*. New York: Oxford University Press.

Sloman, S. A., & Hagmayer, Y. (2006). The causal psycho-logic of choice. *Trends in Cognitive Sciences*, *10*, 407–412.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. New York: Springer-Verlag.

SPSS, Inc. (2001). The SPSS TwoStep Cluster Component: A scalable component enabling more efficient customer segmentation. Technical Report. Available at: http://www.spss.ch/upload/1122644952_The%20SPSS%20TwoStep%20Cluster%20Component.pdf. [accessed on may 23, 2012].

Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: North-Holland Publishing Company.

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153–160). Cambridge, UK: Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *9*, 293–315.

Tversky, A., & Koehler, D. (1994). Support Theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*, 547–567.

White, P. A. (2006). The causal asymmetry. *Psychological Review*, *113*, 132–147.

Zhang, T., Ramakrishnon, R., & Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases. In *Proceedings of the ACM SIGMOD conference on management of data* (pp. 103–114). Montreal, Canada: ACM.

Appendix A

Scenarios used in Experiment 1 (in the judgment task, *A* is known, *B* is the uncertain event, and *C* is not mentioned)

	Variables		
Scenarios	*A*	*B*	*C*
1	Quality of sleep	Quality of muscle tone	Level of magnesium in the blood
2	Quality of products	Number of employees	Sales objectives
3	Activation of a valve	Activation of a wheel	Activation of a piston
4	Sweating	Impulsiveness	Hormone activity
5	Competence in English	Negotiation skills	Frequency of international missions
6	Level of iron in the blood	Irritability	Level of teranin (neurotransmitter)
7	Employment	Housing construction	Number of inhabitants
8	Well-being at work	Bonus	Efficiency

Scenarios used in Experiments 2 and 3

Scenarios	*A*	*B*	*C*
1	Quality of sleep	Muscle tone	Level of magnesium
2	Flow of water	Ground conductivity	Water pressure
3	Nerve conduction	Level of potassium	Quantity of astrocytes
4	Self-efficacy feeling	Perseverance	Interns attributions of success
5	Heat capacity	Temperature	Pressure
6	Level of iron in the blood	Irritability	Level of teranin (neurotransmitter)

Scenarios used in Experiment 4

Scenarios	*A*	*B*	*C*
1	Level of deldrin (substance in organisms)	Liver necrosis	Posopathy (disease)
2	Anxiety	Thermotaxis	Level of serotonin
3	Depressive symptoms	Overweight	Diabetes
4	Level of xéroxin (substance in organisms)	Pupillary diameter	Heartbeat

Appendix B

This appendix contains a proof that for the causal structures and data we have considered, Bayesian inference on causal graphical models cannot reproduce the pattern of conditional probabilities offered by experimental participants.

Specifically, we show that if P_{chain}, P_{common}, and $P_{\text{diagnostic}}$ correspond to probabilities solicited from our participants for causal chains, common causes, and diagnostic chains, respectively, then causal Bayes nets with generative (e.g., noisy-OR and additive linear) parameterizations require that

$$P_{\text{chain}} \leq P_{\text{common}} \leq P_{\text{diagnostic}} \vee P_{\text{chain}} \geq P_{\text{common}} \geq P_{\text{diagnostic}} \tag{1}$$

in contrast to our data, in which people judge P_{common} to be lower than $P_{\text{diagnostic}}$ and P_{chain}. Our demonstration has two parts. In the first, we show that when different structures share the same probabilities of exogenous causes and effects given their causes, Expression 1 is true. In the second, we show that the data given to participants have the same likelihoods across all edges in the three structures as a function of those probabilities, which implies that Expression 1 holds in general as long as we assume a prior that is indifferent to the identities of the specific variables.

We will represent ''low'' and ''high'' values for variables with 0 and 1, respectively. Let π_1 denote $P(Y = 1|X = 1)$ where X is a cause of Y, let π_0 denote $P(Y = 1|X = 0)$, and let r denote the rate at which exogenous variables—those lacking an observable cause—are equal to 1. We assume that π_0 and π_1 are the same across different edges of the chain and common-cause structures, an assumption we return to later.

If S is a causal structure that can be either $A \rightarrow C \rightarrow B$ (''ACB-chain''), or $A \leftarrow C \rightarrow B$ (''C-cause''), then

$$\begin{aligned} P_{\text{chain}} =& P(B = 1|A = 1, S = \text{ACB} - \text{chain}) \\ =& \frac{\sum_c P(B = 1|C = c)P(C = c|A = 1)P(A = 1)}{\sum_c \sum_b P(B|C = c)P(C = c|A = 1)P(A = 1)} \\ =& \pi_0 - \pi_0\pi_1 + \pi_1^2 \end{aligned}$$

$$
\begin{aligned}
P_{\text{diagnostic}} &= P(A = 1 | B = 1, S = \text{ACB} - \text{chain}) \\
&= \frac{\sum_c P(B = 1 | C = c) P(C = c | A = 1) P(A = 1)}{\sum_c \sum_a P(B = 1 | C = c) P(C = c | A) P(A)} \\
&= \frac{r(\pi_1^2 + \pi_0 - \pi_0 \pi_1)}{r(\pi_1 - \pi_0)^2 + \pi_0(\pi_1 - \pi_0 + 1)}
\end{aligned}
$$

$$
\begin{aligned}
P_{\text{common}} &= P(B = 1 | A = 1, S = \text{C} - \text{cause}) \\
&= \frac{\sum_c P(A = 1 | C = c) P(B = 1 | C = c) P(C = c)}{\sum_c \sum_b P(A = 1 | C = c) P(B | C = c) P(C = c)} \\
&= \frac{r(\pi_1^2 - \pi_0^2) + \pi_0^2}{r(\pi_1 - \pi_0) + \pi_0}
\end{aligned}
$$

Part 1: Expression 1 is true for any fixed π_0, π_1, and r

Our approach will be to take arbitrary valid values for π_0 and π_1 and show that Expression 1 is true for all valid values of r.

The terms P_{chain}, $P_{\text{diagnostic}}$, and P_{common} are all continuous functions of r if $\pi_1 > \pi_0$, so if $P_{\text{common}} = P_{\text{chain}}$ at exactly one value of r given by r', then any inequality that holds between P_{common} and P_{chain} for some $r > r'$ is stable, in that it must hold for all $r > r'$. Similarly, inequalities between P_{common} and P_{chain} are stable for $r < r'$. If $P_{\text{common}} = P_{\text{diagnostic}}$ only at that same r', then the same stability relationships hold for those variables. We will show that for some $r > r'$ and for some $r < r'$, Expression 1 is true, so that total order of all three terms is stable: P_{common} does not change its ordering relative to P_{chain} or $P_{\text{diagnostic}}$; and P_{chain} and $P_{\text{diagnostic}}$ cannot change their relative ordering without doing so relative to P_{common}. As a result, Expression 1 is true for $r > r'$, $r < r'$, and $r = r'$ (in which all three terms are equal), for arbitrary π_0 and π_1.

If we solve $P_{\text{common}} = P_{\text{chain}}$ for r, we obtain $r' = \pi_0/(1 - \pi_0 + \pi_1)$. Solving $P_{\text{common}} = P_{\text{diagnostic}}$ yields the same unique solution, assuming that r, π_0, and π_1 are valid probabilities and $\pi_0 < \pi_1$. This shows that orderings between P_{common} and P_{chain} and $P_{\text{diagnostic}}$ are stable for all $r < r'$ and $r > r'$. Letting $r = 1$, we find that $P_{\text{common}} = \pi_1$, $P_{chain} = \pi_0 - \pi_0 \pi_1 + \pi_1^2$, and $P_{\text{diagnostic}} = 1$, implying that $P_{\text{diagnostic}} \geq P_{\text{common}} \geq P_{\text{chain}}$. Letting $r = 0$, $P_{\text{common}} = \pi_0$, $P_{chain} = \pi_0 - \pi_0 \pi_1 + \pi_1^2$, and $P_{\text{diagnostic}} = 0$, implying that $P_{\text{diagnostic}} \leq P_{\text{common}} \leq P_{\text{chain}}$. Thus, for arbitrary valid π_0 and π_1, and r, the common-cause probability falls between the causal and diagnostic chain probabilities.

Part 2: The distributions of π_0, π_1, and r are identical across graphs and edges

The above result would not support our central claim if π_0, π_1, and r differed in certain systematic ways between the different experimental conditions or across edges in a single causal graphical model, which is why we must also show that the distributions of π_0, π_1, and r can reasonably be expected to be the same in the three conditions and across all edges.

We have no reason to believe that people's priors over π_0, π_1, and r systematically vary with causal structure or the identities of the given variables as we are considering several different cover stories. Consequently, we assume that differences in the posterior distributions for the parameters depend only on the data, via likelihoods.

We focus here on the case where participants were given event data, which is less subject to ambiguity in its interpretation than statements about correlations. For all three structures, the influence of r on the likelihood of the events depends only on the value of the exogenous variable, which takes a high value in half of the cases regardless of the structure, meaning that the posterior distribution of r does not vary across the three different structures. Similarly, π_0 and π_1 influence the likelihood of the data only via the values of the two variables on the corresponding edge, which have an identical pattern between structures and between edges of a specific structure. Consequently, the posterior distributions are the same for all edges and exogenous variables across all structures, implying that the result in Part 1 is generally applicable to our data.