Title

Evaluating Causal Hypotheses: The Curious Case of Correlated Cues

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 38(0)

Authors

Rehder, Bob
Davis, Zachary

Publication Date

2016

Peer reviewed

Evaluating Causal Hypotheses: The Curious Case of Correlated Cues

Bob Rehder (bob.rehder@nyu.edu)
Zachary Davis (zach.davis@nyu.edu)
Department of Psychology, New York University
6 Washington Place, New York, NY 10003 USA

Abstract

Although the causal graphical model framework has achieved considerable success accounting for causal learning data, application of that formalism to multi-cause situations assumes that people are insensitive to the statistical properties of the causes themselves. The present experiment tests this assumption by first instructing subjects on a causal model consisting of two independent and generative causes and then requesting them to make *data likelihood judgments*, that is, to estimate the probability of some data given the model. The correlation between the causes in the data was either positive, zero, or negative. The data was judged as most likely in the positive condition and least likely in the negative condition, a finding that obtained even though all other statistical properties of the data (e.g., causal strengths, outcome density) were controlled. These results pose a problem for current models of causal learning.

Hypothesis testing occupies a central role in learning theory. On this view, learners use observed data to update their beliefs about different possible models of the world. A critical component of this process are learners' judgments regarding how probable, or improbable, it is that the observed data were generated by each of the hypotheses. In this paper we consider what factors affect learner's judgments regarding the likelihood that data was generated a particular *causal* hypothesis.

For example, consider the situation where there are two potential causes (C_1 and C_2) of an effect E. Fig. 1 shows the four hypotheses, or *graphs* (G), formed by crossing the presence/absence of $C_1 \rightarrow E$ with the presence/absence of $C_2 \rightarrow E$. Evaluating the posterior probability of these graphs involves calculating the probability of the observations, or the data (D), were generated by each graph, $p(D|G_i)$, and then applying Bayes' law. Indeed, Carroll, Cheng, and Lu (2013) adopted the hypothesis testing framework shown in Fig. 1 to account for learning data from some traditional associative learning paradigms involving two cues. While Griffiths and Tenenbaum (2005) initially developed the hypothesis testing methodology to account for learning data from simpler situations involving just *one* potential cause (also see Lu et al., 2008; Meder et al., 2014), it has since been extended to multi-cause situations (e.g., three potential causes in Powell et al., 2016).

Our purpose in this article is to highlight what we find to be an interesting property of how these models calculate the probability that data D were generated by a particular causal graph G. To do so we will present a modified version of the notation presented in Carroll et al. (2013). For generative causes, the likelihood of the data D under a particular parameterization of graph G was defined as

$$p(D|\boldsymbol{w}, G) = \prod_{e,\boldsymbol{c}} p(e|\boldsymbol{c}, \boldsymbol{w}, G)^{N(e,\boldsymbol{c})} \quad (1)$$

where $\boldsymbol{c}$ is a vector denoting the presence ($c_i = 1$) or absence ($c_i = 0$) of the cues and $N(e, \boldsymbol{c})$ gives the frequency counts for each combination of the presence/absence of the effect and the cues in D. By denoting the causal strength of the causal relations, the vector $\boldsymbol{w}$ represents the parameterization of G. The probability of the effect was defined as

$$p(e = 1|\mathbf{c}, \boldsymbol{w}, G) = 1 - \prod_{i \in I(G)} (1 - w_i)^{c_i} \quad (2)$$

where $I(G)$ is the set of indices corresponding to the causal links that are present in G (i.e., $I(G_0) = \{\}$, $I(G_1) = \{1\}$, $I(G_2) = \{2\}$, $I(G_3) = \{1,2\}$). Eq. 2 codifies the common assumption that each causal relation operates independently and that multiple causal influences combine according to a noisy-or integration rule (Cheng, 1997). Note that it is also common to include an additional causal strength parameter, w_0, that represents the probability that E might be caused by something other than C_1 and C_2.

Calculating $p(D|G_i)$ from Eqs. 1 and 2 requires integrating over possible causal strengths (the ws), which Carroll et al. assumed were uniformly distributed. Finally, calculating the posterior probability of the graphs, $p(G_i|D)$, requires stipulating their prior probability, which were assumed to be equal.

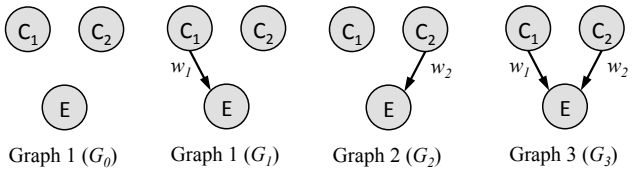

Figure 1

Although Eqs. 1-2 may appear to compute "the likelihood of the data given a model," in fact there is what might be considered an omission—no consideration is given to whether the data is consistent with the *base rate* of the causes stipulated by the models. This omission manifests itself in the absence of a parameter representing the base rates of the causes. Moreover, not only are the base rates of the causes not considered, neither is any *correlation* between C_1 and C_2. Stated more generally, in all these cases the computation of $p(D|G_i)$ is insensitive to any assumptions that learners might have regarding the statistical properties of the causes themselves. Such models are referred to as *conditional Bayesian networks* because they encode the distribution of a subset of variables (in our case, E) given their par-

ents (the Cs) but not the distribution of the parents themselves (Koller & Friedman, 2009).

Of course, a theorist might reasonably assume that when evaluating causal hypotheses it is natural for learners to focus on the manner in which causes co-vary with an effect rather than on the statistical properties of the causes themselves. Here we present data suggesting that this assumption is false. We do this by manipulating, in a two-cue learning scenario like that in Fig. 1, a statistical property of the two causes while holding their relationship with the effect E constant. In particular, we show that people's explicit judgments of $p(D|G_3)$ vary depending on whether the correlation between the two causes is positive, negative, or, zero.

Why Cue Correlations Might Matter

What reason might there be to think that causal learners expect cues to be correlated? Evidence for this view comes from a study of causal *reasoning* (rather than learning) reported by Rehder (2014a). Subjects were taught artificial causal relations intended to be plausible. For example, the causal structure represented by G_3 in Fig. 1 was instantiated in the domain of economics by telling subjects that "low interest rates causes high retirement savings" ($C_1 \rightarrow E$) and that "small trade deficits causes high retirement savings" ($C_2 \rightarrow E$). After learning these materials subjects were presented with 3AFC trials in which they were asked to decide which of two economies were more likely to possess a particular variable (a third "equally likely" alternative was also available). For instance, Fig. 2 depicts two economies in which the subject is being asked to predict which is more likely to have C_2 (in the example, small trade deficits). In the economy on the left of Fig. 2, $C_1 = 1$ (interest rates are low) is given; in the economy on the right, $C_1 = 0$ (interest rates are high). No information about the state of E was provided in either economy. Subjects were significantly more likely to choose the economy on the left, that is, they behaved as if they believed the two causes were positively correlated (also see Perales et al., 2004).

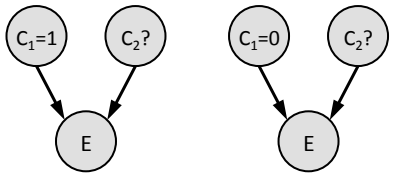

Figure 2

Formally speaking, this result represents a violation of the independence relations entailed by G_3. This is so because although a causal graphical model may have exogenous influences that are not included in the graph, those influences are constrained to be uncorrelated (ruling out, e.g., all unobserved common causes whose values are not constant)—a fact that entails the unconditional independence of C_1 and C_2 (Pearl, 1988; 2000; Spirtes et al., 2000). Of course, one might question whether people had prior beliefs that the causes were correlated. For instance, they may believe that low interest rates are more likely when trade deficits are small than they would be otherwise, despite the instructions during the task. However, this interpretation is defeated by the counterbalancing of the senses of the variables (e.g., the role of C_1 was sometimes played by *high* instead of low interest rates; the role of C_2 was sometimes played by *large* instead of small trade deficits) and by the fact that the same pattern of results obtained in multiple domains (meteorology and sociology in addition to economics).

It has been established that people violate independence relations (a.k.a., commit "Markov violations") with network topologies other than G_3. Rationalizations of these errors all appeal to the possibility that subjects reason with knowledge in addition to that assumed by the experimenters (see Rottman & Hastie, 2013 for a review). Yet none explain the errors found by Rehder (2014a) with G_3. For instance, Park & Sloman (2003) demonstrated that the Markov violations that arose with their *common cause networks* (i.e., $E_1 \leftarrow C \rightarrow E_2$)—namely, subjects incorrectly treated E_1 and E_2 conditioned on C as dependent—were partly due to subjects' beliefs that the two causal links could be disabled by a common factor (also see Burnett, 2004; Lagnado & Sloman, 2004; Mayrhofer & Waldmann, 2015; Walsh & Sloman, 2008). However, this account does not explain the violations that occur with G_3. Relatedly, Rehder & Burnett (2005) explained the large variety of Markov violations they observed by assuming that all variables were related by an underlying common cause (an assumption justified on the basis of the fact that the variables were features of the same category; also see Rehder, 2014b). However, this account also fails to explain the results from Rehder (2014a) because it tested materials that were not features of categories.

The aim of this study is to assess whether the independence violations observed in reasoning tasks generalize to a *learning* context. One explanation for the independence violations observed during reasoning is that they are a side effect of the particular causal reasoning processes subjects invoke to render conditional probability judgments. If this is the case, we would not expect those violations to generalize to a learning task. If, on the other hand, people's understanding of the statistics implied by causal networks really differs from those assumed by formal models, we would expect independence violations to be reflected during both reasoning *and* learning. In particular, if people's beliefs about the statistics of common effect models is such that they think that the causes are positively correlated, then a data sample in which the cues are positively correlated will be viewed as more likely to be generated by a common effect model than samples in which the cues exhibit a zero or negative correlation.

Overview of Experiment

We present an experiment that evaluates the effect of intercue correlations on the evaluation of a causal hypothesis. Note that we adopted a novel experimental paradigm in which we didn't ask subjects to evaluate the relative likelihood of two hypotheses. Instead, we cut out the middle man, so to speak, by (a) presenting a candidate causal theory (one that took the form of G_3), (b) presenting a set of observations, and (c) asking subjects to rate how likely it is that the data was generated by the theory assuming that the theo-

ry is true. That is, we asked subject to *directly* evaluate $p(D|G_3)$.

Our key manipulation is to vary the degree of cue covariation across three levels: negative, zero, and positive. To generate sets of observations (hereafter, "samples"), we first identified a number of target parameters that we wanted the samples to match. Those parameters were, in order of importance, the strength of the causal relationships (.50), the marginal probability of the effect (.50), the log odds ratio (LOD) between the cues (–3, 0, 3 in the negative, zero, and positive conditions), the probability that E was caused by something other than C_1 and C_2 (0), and the marginal probability of each of the causes (.50). Sets of observations that closely instantiates these target parameters are presented in Table 1. The first eight rows of Table 1 correspond to the eight states that can be formed by three binary variables (again, the presence of a variable is denoted by 1 and its absence by 0). Each of these rows presents the number of instances of that type of observation in each of the three conditions. As described in detail later, the presentation of these observations was simultaneous, that is, all observations were visible on the computer screen at the same time.

Round off error induced by the finite-sized samples means that they didn't perfectly match the target parameters and so the bottom portion of Table 1 presents the statistics computed from the actual samples. As can be seen, the LODs between the cues were –3.51, 0, and 3.18, the causal strengths are all very close to .50, the marginal probability of E is always .46, and the probability that E occurs in the absence of C_1 and C_2 is always 0.

How might subjects estimate the likelihood of each of the three data sets given G_3? We first treated G_3 as a conditional Bayesian network in which the probability of the effect conditioned on the causes is computed without making any assumptions about the statistical properties of the causes. We refer to the probability of D given G_3 in this case as $p^{C_1C_2}(D|G_3)$ to emphasize that the network is conditioned on C_1 and C_2. The parameter space of G_3 is thus $\boldsymbol{\theta} = (w_0, w_1, w_2)$ and we sampled (10,000 times) over that space in the manner described in the Appendix. The averaged values of $log(p^{C_1C_2}(D|G_3))$ for each of the three conditions are shown in Table 1, which shows that the sample with a negative inter-cue correlation is most likely to be generated by G_3 (log likelihood of –15.70) followed by the samples with the zero (–16.05) and positive (–16.37) cue correlations. In particular, because this analysis fails to predict a preference for the sample with positively correlated cues, a finding that subjects in fact exhibit such a preference will bolster our claim that the cues of a common effect model are expected to be positively correlated.

We also computed $p(D|G_3)$ treating G_3 as *unconditional* Bayesian network, that is, taking the base rates of C_1 and C_2 into account. In this case the parameter space of G_3 is $\boldsymbol{\theta} = (b_1, b_2, w_0, w_1, w_2)$ where b_1 and b_2 represent the base rates of C_1 and C_2, respectively. Table 1 reveals that for this model, $log(p(D|G_3))$ is highest in the zero correlation condition. These analyses reveal that subjects should not favour the sample with positively correlated cues regardless of whether G_3 is treated as a conditional or unconditional network.

Whereas these analyses sampled over a uniformly distributed parameter space, subjects might think that some parameters values are more likely than others. Using beta distributions we introduced a prior in which $E(b_1) = E(b_2) = .5$, (i.e., C_1 and C_2 are each expected to occur about half the time), $E(w_1) = E(w_2) = .9$, (i.e., $C_1 \rightarrow E$ and $C_2 \rightarrow E$ are expected to be strong causal relations), and $E(w_0) = .1$ (i.e., causes of E other than C_1 and C_2 are expected to be weak). In this prior each expectation was worth 20 "observations." In fact, the introduction of this prior left the qualitative pattern of $p(D|G_3)$ and $p^{C_1C_2}(D|G)$ unchanged.

Table 1.

			Condition		
			Negative	Zero	Positive
C_1	C_2	E	Count	Count	Count
0	0	0	3	6	8
0	0	1	0	0	0
0	1	0	5	3	1
0	1	1	5	3	1
1	0	0	5	3	1
1	0	1	5	3	1
1	1	0	0	1	3
1	1	1	1	5	9
	$LOD(C_1, C_2)$		-3.51	0	3.18
	$p(E=1\|C_i=1, C_j=0)$		0.52	0.53	0.50
	$p(E)$		0.46	0.46	0.46
	$p(E=1\|C_1=0, C_2=0)$		0	0	0
	$p(C_i)$		0.46	0.50	0.58
	$\log(p(D\|G_3))$		-40.98	-35.52	-41.85
	$\log(p^{C_1C_2}(D\|G_3))$		-15.70	-16.05	-16.37

Note that the process via which the samples in Table 1 were generated was chosen to avoid confounds that might complicate the interpretation of the results. One possibility that concerned us is that subjects might interpret the $p(D|G_3)$ query as one requesting the *strengths* of the causal relations that inhered in each sample. This is why all three samples were constrained to have almost equal causal strengths. Moreover, because causal strength judgments themselves are known to be affected by the marginal probability of the effect (the "outcome density bias"), the samples were matched on that factor as well; they were also matched on the probability of E occurring in the absence of C_1 and C_2 (it was 0). Note that one factor on which the samples *weren't* matched is the marginal probability of the causes (which were .46, .50, and .58 in the negative, zero, and positive correlation conditions). Because in G_3 the marginal probability of E will, all else being equal, increase as the cues become negatively correlated and decrease as they become positively correlated, equating $p(E)$ across conditions entails that $p(C_i)$ is smaller in the negative condition and larger in the positive condition. Equating $p(E)$ was deemed

more important because of the documented outcome density bias.

In summary, if subjects' judgments concerning $p(D|G_3)$ reflect an expectation that cues are uncorrelated, then the ratings should be highest in the zero-correlation condition. If it ignores the statistical properties of the cues altogether, then ratings should be highest in the negative-correlation condition. Finally, if it reflects an expectation that cues will be positively correlated, then ratings should be highest in the positive-correlation condition and lowest in the negative-correlation condition. The following experiment tested these predictions.

Method

Materials. Participants were presented with causal hypotheses in three domains: meteorology, sociology, and economics. Each domain had three variables: interest rates, trade deficits, and retirement savings in economics; degree of urbanization, interest in religion, and socio-economic mobility in sociology; and ozone level, air pressure, and humidity in meteorology. Each variable could take on two possible values. One of these values was described as "Normal" and the other was either "High" (or "Large" for some variables) or "Low" (or "Small") according to a ran-

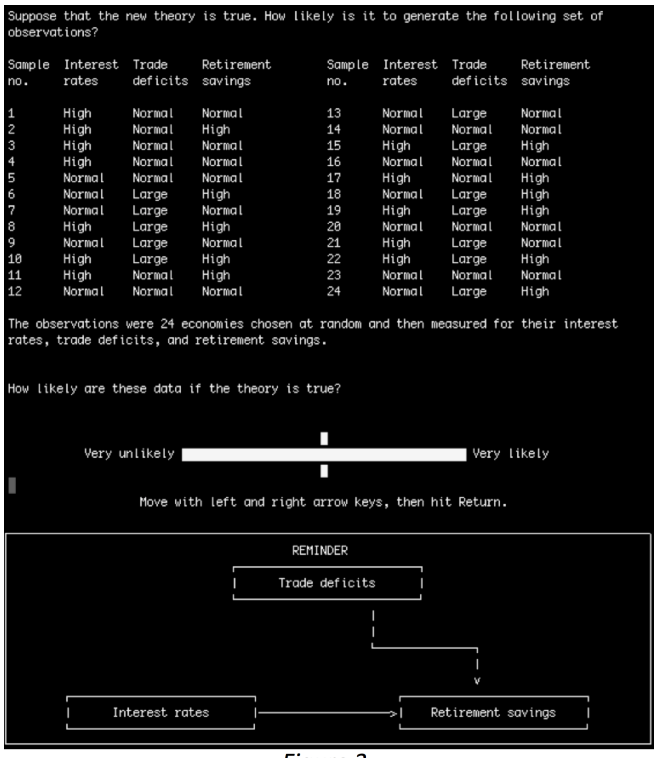

Figure 3

domization scheme described below.

Each hypothesis was of the form G_3 in Fig. 1. The description of each of the two causal relationships consisted of two sentences, one that stated the cause and effect and a second that provided information about the causal mechanism. For example, "High interest rates cause high retirement savings. The high interest rates result in high yields on government bonds, which are especially attractive for retirement savings because they are such a safe investment."

Procedure. Participants first learned a hypothesis in each domain. For each hypothesis, they then judged the degree of fit between it and the three samples in Table 1. Participants thus made a total of nine judgments.

In each domain, participants were first told that researchers (a "research group of economists" in the case of economics) had proposed a "new theory" consisting of two paragraphs describing the two causal relationships. A diagram of those causal relationships analogous to G_3 in Fig. 1 was also presented. Subjects were verbally instructed in two ways that the causal relationships operated independently. Firstly, they were explicitly told in two ways that the causes were independent (e.g., "note that high interest rates and large trade deficits each bring about high retirement savings on its own"; "high interest rates can produce high retirement savings by itself, and large trade deficits can independently produce high retirement savings by itself as well"). Secondly, as described in Materials, they were provided with separate causal mechanisms for each of the causes, providing an account by which each cause could independently bring about the effect.

Participants were then presented with the three samples. Each was presented on one screen that listed the 24 observations (i.e., presentation was simultaneous). See Fig. 3 for an example. Each sample was described as being "chosen at random" and measured on the three variables that made up the hypothesis (interest rates, trade deficits, and retirement savings in the case of economics). Subjects were asked to judge the likelihood of observing this sample assuming that the theory was true. The response scale ranged from "Very unlikely" to "Very likely". So that subjects did not have to rely on their memory of the theory, the diagram of the causal relationships was re-presented at the bottom of this screen (see Figure 3). The next two screens presented the second and third sample. The presentation order of the domains (economies, societies, and weather systems) and the three samples (Negative, Zero, Positive) was randomized as described below.

Design and participants. There were two between-subject factors. A Latin square determined the order of presentation of the three domains (meteorology, sociology, economics). Subjects were randomly assigned to these three cells subject to the constraint that an equal number appeared in each. The senses of the variables' non-normal values were randomized such that they were either all high, all low, or all high in the first, second, and third domain, respectively (HLH) or exhibited the reverse pattern (LHL). Within each domain the presentation order of the three samples was randomized for each subject. 21 New York University undergraduates received course credit for participating.

Results

Initial analyses revealed no effect of domain (economics, sociology, meteorology) or sense (HLH, LHL), so the results presented are collapsed over these factors

As hypothesized, there was a main effect of inter-cue correlation on ratings of ($p(D|G_3)$, $F_{2,32} = 33.82$, MSE = 53.58, $p < .001$). Ratings from 0-100 for each of these correlations

are displayed in Fig. 4. Additionally, participants consistently preferred positive correlation, with positively correlated cues having higher ratings than non-correlated cues ($p < .001$), which in turn had higher ratings than negatively correlated cues ($p < .001$).

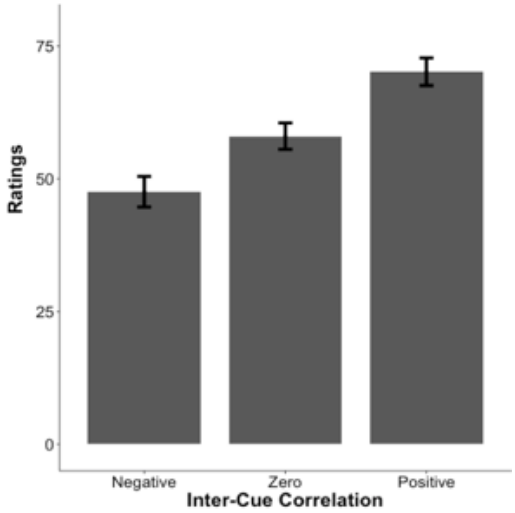

Figure 4. Mean Ratings. Error bars represent standard errors

Discussion

This article presents evidence that people are sensitive to inter-cue correlations in a way not predicted by current causal learning models. In particular, they expect that the two cues in a common effect structure will be positively correlated, resulting in a preference for data that manifests such a correlation. Because the current findings involved a new sort of task, they generalize the violations of the independence constraints stipulated by causal Bayes nets that obtain in reasoning tasks in which subjects render conditional probability judgments. Participants' non-normative expectations of the statistical properties of data generated from causal structure poses problems for models (e.g., Carroll et al., 2013; Griffiths & Tenenbaum, 2005; Lu et al., 2008; Meder et al., 2014; Powell et al., 2016) that assume that learners estimate $p(D|G)$ in a veridical manner.

There are important methodological differences between the present work and traditional causal learning studies. Whereas we asked participants to judge the probability that a particular causal structure generated some data set (i.e. judge $p(D|G_3)$), participants are usually asked to estimate causal strength (e.g., Cheng, 1997, and many others) or to select the causal structure most consistent with the observed data (e.g., Lu et al. 2008). We have also conducted a straightforward extension of our paradigm by asking subjects to instead estimate the posterior probability of the four hypotheses in Fig. 1—that is to estimate $p(G_i|D)$ for each G_i instead of $p(D|G_3)$. Consistent with the results above, we found that $p(G_3|D)$ was highest in the positively-correlated condition and lowest in the negatively-correlated one. We have also asked subjects to generate hypothetical data from G_3 (having them, in effect, estimate G_3's joint distribution). In fact, that joint distribution reflected their expectation that the two causes were positively correlated.

Our paradigm can reveal the non-normative expectations about statistical structure that people have for other network topologies. For instance, in a common cause structure ($E_1 \leftarrow C \rightarrow E_2$), the fact that E_1 and E_2 should be independent conditioned on their common cause C is often violated (e.g. Rehder, 2014a). It would be straightforward to generate samples of data in which, conditioned on C, the correlation between E_1 and E_2 is either negative, zero, or positive. We predict that the sample with the positive correlations will be treated as more consistent with $E_1 \leftarrow C \rightarrow E_2$.

Most modern models of causal learning fall under the umbrella of conditional Bayesian networks, which are invariant over statistical properties of parent nodes (Carroll et al., 2013; Griffiths & Tenenbaum, 2005; Lu et al., 2008; Meder et al., 2014; Powell et al., 2016). The current study provides evidence that such assumptions contrast with actual causal-based judgments. A fully descriptive formal model of causal learning behaviour, then, must also be able to account for participant expectations of the statistical properties of cues.

References

Burnett, R. C. (2004). Inference from complex causal models (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI No. 3156566)

Carroll, C. D., Cheng, P. W., & Lu, H. (2013). Inferential dependencies in causal inference: A comparison of belief-distribution and associative approaches. *Journal of Experimental Psychology: General, 142*, 845-863.

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

Griffiths, T. L., & Tenebaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*, 334-384.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-454.

Koller, D. & Friedman, L. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press.

Lagnado, D.A. & Sloman S.A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30*:856–76

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.

Maloney, L. T., Dal Martello, M. F., Sahm, C. & Spillmann, L. (2005), Past trials influence perception of ambiguous motion quartets through pattern completion. *Proceedings of the National Academy of Sciences*, **102**, 3164-3169.

Mayrhofer, R. & Waldmann, M. R. (2015). Agents and causes: Dispositional intuitions as a guide to causal structure. *Cognitive Science, 39*, 65-95.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review, 121*, 277-301.

Park, J. & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. *Cognitive Psychology, 67*, 186-216.

Pearl, J. (2000). Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association*, *95*(450),

428-431.
Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.
Perales, J., Catena, A., & Maldonado, A. (2004). Inferring non-observed correlations from causal scenarios: The role of causal knowledge. Learning and Motivation, 35, 115–135.
Powell, D., Merrick, M. A., Lu, H., & Holyoak, K. J. (2016). Causal competition based on generic priors. *Cognitive Psychology, 86*, 62–86.
Rehder, B. (2014a). Independence and dependence in human causal reasoning. *Cognitive Psychology, 72*, 54-107.
Rehder, B. (2014b). The role of functional form in causal-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*
Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of object categories. *Cognitive Psychology, 50*, 264-314.
Rottman, B., & Hastie, R. (2013). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*.
Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. New York: Springer-Verlag.
Walsh, C. R., & Sloman, S. A. (2008). Updating beliefs with causal models: Violations of screening off. In G. H. Bower, M. A. Gluck, J. R. Anderson & S. M. Kosslyn (Eds.), *Memory and Mind: A Festschrift for Gordon Bower* (pp. 345-358).

Appendix

Computing $p^{C_1C_2}(D|G_3)$ when G_3 is interpreted as a conditional network yields,

$$p^{C_1C_2}(D|G_3) = \int_0^1\int_0^1\int_0^1 p^{C_1C_2}(D|\boldsymbol{\theta}, G_3)\, p(\boldsymbol{\theta}|G_3)\mathrm{d}w_0\mathrm{d}w_1\mathrm{d}w_2 \quad (3)$$

The computation of $p^{C_1C_2}(D|\boldsymbol{\theta}, G_3)$ raises well know issues regarding how people estimate the likelihood of sequences of events. For example, if the likelihood of the effect E in some context is .50, then presumably people will judge a sequence in which E is always present as less probable than one in which E is present about half the time, in the same way that a fair coin is viewed as unlikely to yield a long run of all heads or all tails as compared to a mixed sequence (Kahneman & Tversky, 1972; also see Maloney et al., 2004). This concern is especially relevant in the current experiments in which observations are presented simultaneously rather than sequentially, making non-representative samples easier to detect. Accordingly, here we use the binomial distribution as a first order approximation of $p^{C_1C_2}(D|\boldsymbol{\theta}, G_3)$. In particular,

$$p^{C_1C_2}(D|\boldsymbol{\theta}, G_3) = \prod_{\boldsymbol{c}} B(N(e=1,\boldsymbol{c}); N(\boldsymbol{c}), p(e=1|\boldsymbol{c})) \quad (4)$$

$$p(e=1|\boldsymbol{c}) = 1-(1-w_0)(1-w_1)^{c_1}(1-w_2)^{c_2} \quad (5)$$

where the product ranges over the four distinct settings of C_1 and C_2 in D.[1] B is the binomial distribution which returns the probability of $N(e=1,\boldsymbol{c})$ "successes" (number of times E is present in context $\boldsymbol{c}$) in $N(\boldsymbol{c})$ "trials," where the probability of a success is $p(e=1|\boldsymbol{c})$.

Computing $p(D|G_3)$ when G_3 is interpreted as an unconditional network yields yields

$$p(D|G_3) = \int_0^1\int_0^1\int_0^1\int_0^1\int_0^1 p(D|\boldsymbol{\theta}, G_3)\, p(\boldsymbol{\theta}|G_3)\mathrm{d}b_1\mathrm{d}b_2\mathrm{d}w_0\mathrm{d}w_1\mathrm{d}w_2 \quad (6)$$

and $p(D|G_3)$ is computed using the multinomical distribution M,

$$p(D|\boldsymbol{\theta}, G_3) = M(N(e,\boldsymbol{c});\, p(e,\boldsymbol{c}|\boldsymbol{\theta}, G_3)) \quad (7)$$

where $N(e,\boldsymbol{c})$ is again the vector of counts associated with each unique combination of C_1, C_2 and E and $p(e=1,\boldsymbol{c}|\boldsymbol{\theta}, G_3))$ is the vector of corresponding probabilities, where

$$p(e,\boldsymbol{c}|\boldsymbol{\theta}, G_3)) = p(e|\boldsymbol{c}, \boldsymbol{\theta}, G_3)p(\boldsymbol{c}) = p(e|\boldsymbol{c})b_1b_2 \quad (8)$$

and $p(e|\boldsymbol{c})$ is given by Eq. 5.

[1] Use of a binomial distribution introduces a normalizing constant in the computation of the likelihood, one that cancels out when the relative likelihood of two hypotheses is being computed (such as in Griffiths & Tenenbaum's support model). In contrast, in the current experiments subjects estimate absolute rather than relative likelihoods. Use of a binomial distribution may also be less appropriate in traditional learning experiments in which data is presented sequentially (where a non-representative sample may be less salient) rather than simultaneously.