UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Learning Causal Structure

Permalink
https://escholarship.org/uc/item/6kf9h3c7

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 24(24)

ISSN
1069-7977

Authors
Lagnado, David A
Sloman, Steven

Publication Date
2002

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Learning Causal Structure

David A. Lagnado (David_Lagnado@Brown.Edu)
Department of Cognitive and Linguistic Sciences, P.O.Box 1978
Providence, RI 02912 USA

Steven Sloman (Steven_Sloman@Brown.Edu)
Department of Cognitive and Linguistic Sciences, P.O.Box 1978
Providence, RI 02912 USA

Abstract

The central aims of this experiment were to compare observational and interventional learning of a simple causal chain, and to ascertain whether people represent their interventions in accordance with the normative model proposed by Pearl (2000). In the observation condition people treated putative causes as independent, and systematically selected the wrong model. In the intervention condition performance improved, in particular greater sensitivity was shown to the relevant conditional independencies. However, participants' likelihood judgments approximated the observed frequencies rather than reflecting the appropriate causal model.

Introduction

Our causal knowledge of the world is closely tied to our ability to control or manipulate certain aspects of it. On the one hand, we often learn about cause-effect relations by observing the effects of our own interventions (e.g., running controlled experiments). On the other, we can exploit such knowledge by manipulating the causes appropriate to our desired ends. Further, our causal knowledge allows us to predict or imagine the consequences of our actions, and is thus a prerequisite for deliberative decision-making.

Given the central role that intervention plays in causal reasoning, it has received scant attention in most accounts of human causal learning. In part this is due to the lack of a formal analysis of intervention, and the failure of standard probability theory to distinguish *action* from *observation* (Pearl, 2000). These lacunas appear to be addressed by a body of recent research in AI, which provides a normative analysis of causal inference and a formal means of representing the difference between observation and intervention (e.g., Glymour, 2001; Pearl, 2000).

The formulation of a normative model is at best only a first step towards an understanding of how people acquire and employ causal knowledge. The current experiment aims to gather some preliminary evidence about the difference between observational and interventional learning, and whether people represent their interventions in the manner suggested by this normative account.

Causal Models

The causal model framework offers a method for representing causal knowledge and formal rules for updating this knowledge in the light of either observation or intervention. Central to this formalism is the use of directed graphs to represent the mechanisms that underpin our causal knowledge of a domain, and the use of probability theory to reflect the uncertain and defeasible nature of this knowledge.

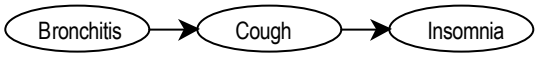

Figure 1: A causal graph

A causal model is made up of a set of nodes, a set of directed links between nodes, and a conditional probability distribution for each node. The nodes correspond to variables relevant to the domain (the pre-selection of which may be non-trivial); these may be binary, or take on a range of values. The directed links between variables correspond to the autonomous mechanisms that are supposed to mediate between these variables, and hence reflect the dependencies between them.

A simple causal graph is depicted in Fig. 1. In this example the model is restricted to three binary variables: *Bronchitis*, *Cough*, and *Insomnia*. There is presumed to be one mechanism that leads from *Bronchitis* to *Cough*, and another that leads from *Cough* to *Insomnia*.

Typically the dependencies between variables are probabilistic – reflecting either the incompleteness of the causal model or genuine noise. This uncertainty is represented by conditional probability distributions for each node (referred to as the *parameterization* of the

graph). Thus in our simple example the strength of dependency between *Bronchitis* and *Cough* is represented by two conditional probabilities – the probability of *Cough* given *Bronchitis*, and the probability of *Cough* given no *Bronchitis*. A high probability for the former would correspond to the belief that *Bronchitis* is very likely to cause *Cough*; a high probability for the latter would correspond to the belief that *Cough* is also very likely to be caused by other variables not represented in our simple model.

Given certain assumptions,[1] the structure of a causal graph will fully capture the probabilistic dependencies amongst all of the represented variables. A fundamental relation here is that of 'screening off' or conditional independence. For any three variables *A*, *B*, *C*: *A* and *B* are conditionally independent given *C* if $P(A|B\&C) = P(A|C)$; once you know the value of C, learning the value of *B* does not provide additional information about *A*. One causal graph representation that implies screening off is when *C* intercepts all directed paths between *A* and *B*. Thus in the causal graph in Fig.1, the fact that the *Cough* node is in between the nodes for *Bronchitis* and *Insomnia* implies that *Bronchitis* and *Insomnia* are conditionally independent given *Cough*. Once you know the value of *Cough*, learning the value of *Bronchitis* tells you nothing more about the value of *Insomnia*.

By representing conditional independencies in this way, causal graphs provide a powerful tool for organizing knowledge, and for inferring the effects of new observations. As the graphs increase in size, these independence relations can greatly simplify such computations. For example, one could supplement the simple model in Fig.1 with a complex network of nodes and links between *Bronchitis* and *Cough*, but so long as the variable *Cough* still intercepts all links from *Bronchitis* to *Insomnia*, knowledge of *Cough* is all one needs to make inferences about *Insomnia*.

Making inferences given new information

The structure of a causal graph, in combination with the parameterization of its nodes, determines what inferences we can make on the basis of new information. When this information takes the form of an observation, then Bayesian updating tells us how we ought to modify our probabilities. For example, given the causal model in Fig. 1, if we find out that Jim has a cough, we should increase (to some degree, depending on the parameters) both the probability that Jim has *Bronchitis*, and the probability that he has *Insomnia*. However, what if we changed the value of *Cough* by giving him a cough suppressant? Such an action warrants a change in our belief that he has *Insomnia*, but does not warrant any change in the probability we assign to him having *Bronchitis*.

More generally, the probabilistic inferences we are licensed to draw after observing the value of a variable may not be the same as those after intervening to set that variable to the same value. Bayesian updating, indeed any formal probability model, fails to recognize this. It does not differentiate between *observing* and *acting*. That is, the same conditional probability P(X|Y) is used to represent the probability of X given that Y *is observed*, and the probability of X given that *we do* Y. But these can be quite different, as our example illustrates – the probability of *Bronchitis* given the absence of a cough is distinct from the probability of *Bronchitis* given that we remove the cough.

The Representation of Intervention

One of the innovative features of the causal model framework is that it proposes a normative account for the representation of interventions, and for the inferences that they license. In so doing, it formalizes the difference between observation and intervention.

Pearl (2000) achieves this through the introduction of the '*do(•)*' operator. In short, this amounts to representing an intervention in terms of a minimal modification of the causal graph. Thus a simple intervention to set a variable to a particular value is represented by the removal of all arrows into that variable, without altering the other directed links in the graph. The effects of the intervention are then computable through Bayesian updating on this 'mutilated' graph.

To illustrate using the graph in Fig.1, consider an intervention (e.g., use of a cough suppressant) that sets *Cough* to the value low. This leads to the modified graph in Fig. 2: The directed link from *Bronchitis* to *Cough* is deleted whilst the link from *Cough* to *Insomnia* is left unchanged. In effect the intervention amounts to placing the variable *Cough* under the influence of a new mechanism that sets its value to low.

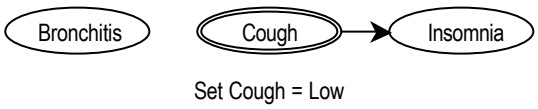

Figure 2: Causal graph after intervention

This account provides a normative model for the representation of both actual and imagined interventions, and tells us how these interventions will (or would) affect the values of the other variables in the system. In particular it dictates which probabilistic inferences we are entitled to make. Thus the modified graph in Fig. 2 permits us to infer a lower probability of *Insomnia*, but no change in the probability of

[1] For example, the explicit representation of any variable that affects two or more other variables in the model.

Bronchitis. The latter prohibition is reflected in Pearl's terminology by the difference between P(*Bronchitis*| ~*Cough*) and P(*Bronchitis*| *do*(~*Cough*)), and captures the basic asymmetry of the cause-effect relation: manipulating a cause can change an effect but not vice-versa.

Learning causal structure

The appropriate representation of intervention is not just critical to predicting the effects of our actions; it is also important for the discovery or learning of causal structure. Causal models can be learned from explicit instruction about how the world works, but we can also learn about causal structure through observation or through intervention. These are not exclusive, but it is useful to distinguish cases in which one is restricted to observational data alone from those in which one also has the opportunity to intervene.

Observational learning

The causal model literature in AI has developed various algorithms for inferring causal structure from observational data, many of which exploit the conditional dependencies encoded in the structure of a causal graph. So far none of these have been proposed as models of actual human discovery, although they do suggest some general principles that are relevant to such enquiries. For example, the establishment of conditional dependencies is a crucial starting point for the construction of a causal graph, so it is important that people are able to make judgments of conditional dependence versus independence. In contrast, the precise parameterization of those dependencies is not always required to discover correct causal structure

Moreover, the graphical approach clarifies which causal structures can be differentiated on the basis of observational data alone. It establishes equivalence classes of structures ("*Markov equivalence*") that share conditional dependencies and are thus indistinguishable on the basis of observation alone.[2] For example, in a model made up of just two nodes, *A* and *B*, ascertaining their probabilistic dependence does not tell us whether *A* causes *B*, or *B* causes *A*.

Even if causal structures are from different Markov equivalence classes, it might be difficult for people to distinguish them on the basis of observational data. Indeed, selecting between certain structures requires careful tracking of observed frequencies and subtle inferences based on what one would expect to see. For example, consider the two possible causal structures depicted in Fig. 3. In order to distinguish these on the basis of observation alone, one must determine whether or not blurred vision and headache are independent given high wine consumption (conditional independence would only hold if the data were produced by the model on the right). This may require many observations and careful tracking of the relevant relative frequencies.

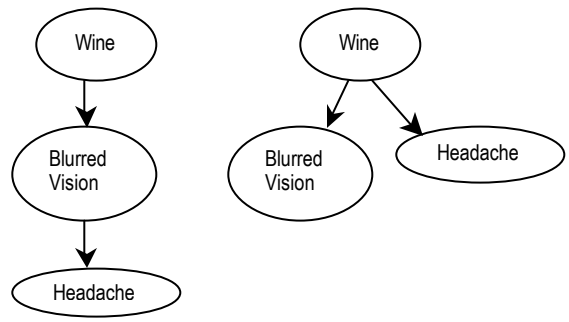

Figure 3: Two possible causal models

Interventional learning

Another way to learn about causal structure is to actively interact with the system under study and to observe the consequences. This seems to apply to the infant playing with a new toy as much as to the scientist running controlled experiments. Whilst this is often recognized as an important source of causal knowledge, it has received less attention in the human causal learning literature.[3]

Intuitively, the ability to intervene on a system should facilitate our learning about its causal structure. To take the simplest example, consider two variables that are known to be probabilistically dependent. Assuming no other relevant variables, the direction of this link can be determined by manipulating one of the variables and observing whether or not the other also changes. In a noisy system such learning may still require multiple trials and sensitivity to the observed frequencies. But interventional learning has several advantages over passive observation. Not only can it help to determine the direction of the causal links, it also allows selection of the kind of data to see, and thus to test out critical relations between variables. For example, let us return to the task of distinguishing between the two possible causal models in Fig. 3. One possible intervention is simply to drink a large amount of wine and then keep your eyes closed. If you don't get a headache, you can be reasonably sure that the chain model is the correct one. If the system is rather noisy you may have to repeat this experiment several times,

[2] One important qualification here is in the case of graphs in which causal links are necessary but not sufficient; that is, for a directed link from *A* to *B*: $1 > P(B|A) > 0$ and $P(B|\sim A) = 0$. Networks built from such links may be distinguishable even though they are Markov equivalent.

[3] The dominant approaches to human causal learning (e.g. Cheng, 1997; Dickinson, 2001; Shanks, 1995) concentrate on observational learning.

but it will still lead to greater confidence than making the distinction on the basis of observation alone.

Overview of Experiment

The central aims of this experiment were to compare the observational and interventional learning of a simple causal model, and to ascertain whether people represent their interventions in accordance with the normative model proposed by Pearl (2000). We used a typical observational learning paradigm (e.g. Shanks, 1995), but adapted it to include an interventional learning condition and a model selection task. The learning data were generated from a simple chain model (see Fig. 4).

Learning performance was assessed both through a model selection task and through the sensitivity of people's probability judgments to the appropriate conditional dependencies.

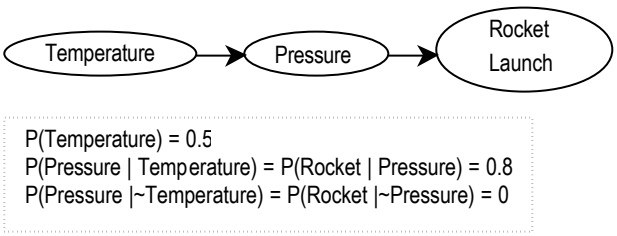

Figure 4: Causal graph used to generate stimuli for both observational and interventional tasks

Method

Participants. Thirty-three undergraduates from Brown University received course credit for their participation.

Materials and procedure. Initial instructions to the participants included an introduction to the notion of a causal model with examples of five candidate models. Each participant then completed both an observational and an interventional learning task. Two cover stories were used, one for each task (task order and scenario were counterbalanced across participants). Participants were asked to imagine that they were space engineers (chemists) running tests on a new rocket (perfume) in order to discover the underlying causal structure. They were told that previous tests had identified two variables as relevant to the success of the test. In the space engineer scenario the relevant variables were *Temperature* (either high or low) and *Pressure* (either high or low), and the outcome variable was whether or not the rocket launched. In the chemist scenario the variables were *Acid level* (either high or low) and *Ester level* (either high or low), and the outcome variable was whether or not the perfume was produced. In the *observation* task participants viewed the results of 50 test trials. On each trial they were shown the values of the two relevant variables, and then clicked on a button to view whether or not the outcome occurred. The learning set was constructed according to a chain model (see Fig. 4) and is shown in Table 1 (order of presentation was randomized for each participant).

Table 1: Frequency of presented instances in Observational Learning condition.

Temperature	Pressure	Rocket Launch	No	Prob
High	High	Yes	16	0.32
High	High	No	4	0.08
High	Low	Yes	0	0
High	Low	No	5	0.1
Low	High	Yes	0	0
Low	High	No	0	0
Low	Low	Yes	0	0
Low	Low	No	25	0.5

Participants then proceeded to a test phase, in which they made various conditional likelihood judgments (e.g., given that *Temperature* is high, and *Pressure* low, what is the likelihood that the rocket launches?) plus a model selection question. This question presented participants with five candidate causal models – two chains, two forks, and a collider (Fig.5 shows one model from each category) – and asked them to select the model that they believed was most likely to have produced the data.

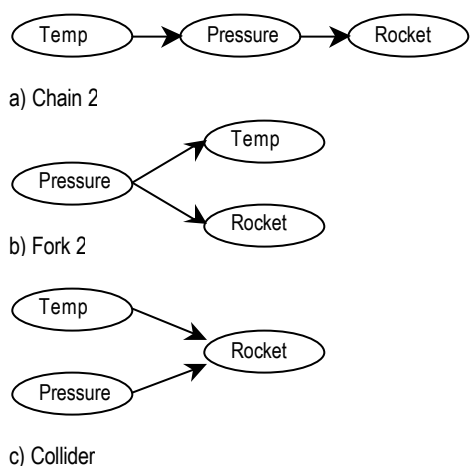

Figure 5: Three models from the selection task

In the learning phase of the intervention task, participants were able to set the value of one of the two relevant variables. They then viewed the resulting values of the outcome variable and the variable they had not intervened on. This learning set was generated from a pseudo-random table constructed in accordance with the same chain model. After running 50 tests they

proceeded to an identical test phase as in the observation task.

Results and discussion

Model Selection. The results for the model selection task are shown in Fig. 6, with the correct chain model designated as chain 2.[4] There were more correct model selections in the intervention than in the observation condition. However, whilst the correct model was the modal response in the intervention condition, overall responses were not significantly different from the uniform distribution ($\chi^2(4) = 2.91$, *ns.*). By contrast in the observational condition there was an overwhelming bias in favor of the collider ($\chi^2(4) = 40.79$, $p < 0.001$).

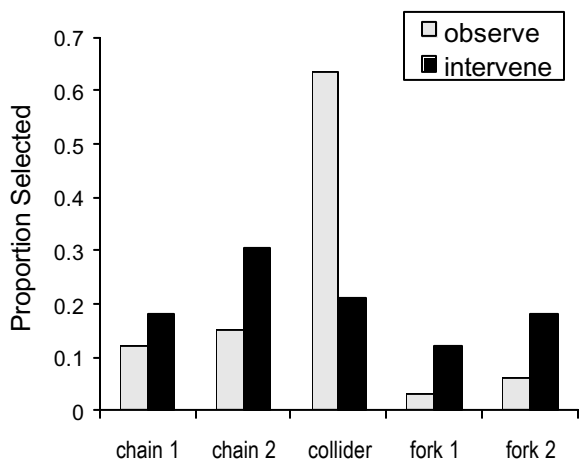

Figure 6: Model selection results in interventional and observational conditions (the correct model is chain 2).

Derived judgments of conditional independence. On the model used to generate the learning set (see Fig. 4), *Temperature* was independent of *Rocket launch* conditional on *Pressure*, that is: $P(R|T\&P) = P(R|P)$. Participants' mean ratings for these two likelihoods are shown in Fig. 7. No significant difference obtained between the two likelihoods in the intervention condition, suggesting that participants were sensitive to this conditional independence. This is reinforced by the fact that 19 out of 33 participants judged the two likelihoods equal. This contrasts with the observation condition, in which the mean likelihoods differed substantially, and only 8 out of 33 participants judged them equal.

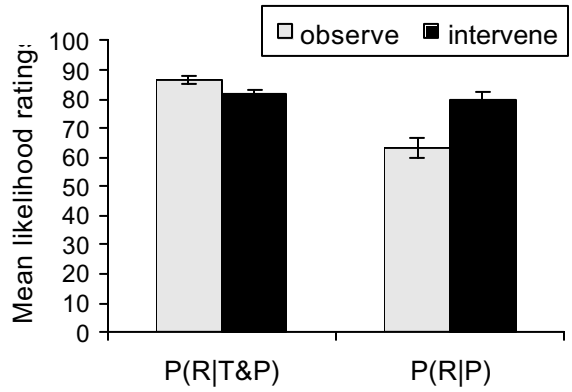

Figure 7: Mean conditional likelihood ratings for the outcome variable R (rocket launch).

Compatibility of judgments with the *do* operator. One way to assess the extent to which participants represent their interventions in line with the *do* operator is to look at their judgments of the likelihood that *Pressure* was high given that *Temperature* was low, $P(P|\sim T)$. Recall that the correct judgment for this likelihood is zero; *Pressure* is never high if *Temperature* is low. However, when participants intervene on the *Pressure* variable and set it to high they temporarily break the link between *Temperature* and *Pressure*. In such cases the value of *Temperature* is equally likely to be high or low (its base rate = 0.5). If participants fail to represent their interventions appropriately, by not 'mentally' removing the link from *Temperature* to *Pressure* when they intervene on *Pressure*, they may erroneously judge that $P(P|\sim T) > 0$. This is because 50% of the time when they set *Pressure* high they will observe *Temperature* as low. In other words, they might fail to mark the distinction between action and observation.

To test out this possibility we compared people's judgments for $P(P|\sim T)$ with the relative frequencies they actually observed; i.e., with the proportion of times they observed both low *Temperature* and high *Pressure* (regardless of whether they intervened on *Temperature* or *Pressure*). As shown in Fig. 8, participants' mean judgments for $P(P|\sim T)$ were very close to the frequencies they observed, and significantly different from the normative value of zero.

[4] One complication is that the chain model used to generate the data is Markov equivalent to fork 2. However, although not inconsistent with the observational data, this model requires an idiosyncratic parameterization whereby one effect (temperature) occurs more often than its sole cause (pressure). Very few people chose this model in the observation condition.

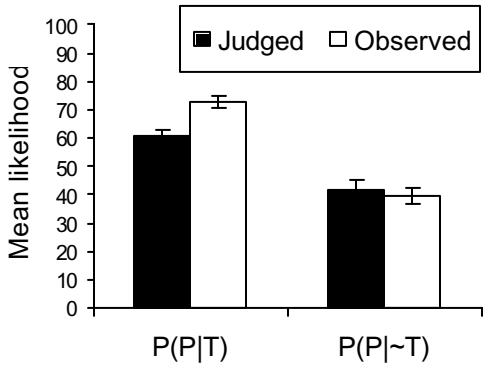

Figure 8: Mean likelihood ratings and observed relative frequencies in the intervention condition.

This result could indicate a failure by participants to implement the *do* operation when inferring the relation between *Pressure* and *Temperature*. However, there are alternative explanations for this finding. One possibility is that participants interpreted the likelihood question in terms of observational rather than interventional probabilities, and accurately reported the relative frequency with which low *Temperature* and high *Pressure* co-occurred, regardless of whether they believed that low *Temperature* would cause high *Pressure*. This fits with numerous studies showing that people encode the relative frequencies of events automatically, and often use these as a basis for their likelihood judgments (e.g., Hasher & Zacks, 1984).

Second, on Pearl's account the notion of an intervention is only well defined relative to a specific causal model. Thus if people uphold an incorrect model (as the majority of the participants did) they are unlikely to give appropriate estimates for the interventional probabilities. Moreover, even those participants that do select the correct model will have entertained various incorrect ones through the course of learning, and it may be very hard for them to retrospectively revise their prior observations.

Conclusions

This experiment demonstrated a contrast between observational and interventional learning, both with respect to people's model selection and their likelihood judgments. Under observational learning, participants exhibited a strong bias for the collider, despite the fact that the variables they judged to be independent were highly correlated in the data. This suggests that they were engaged in predictive learning of the outcome variable (e.g., *Rocket launch*) on the basis of two indicator cues (e.g., *Temperature* and *Pressure*), effectively treating them as independent causes of the outcome. This resonates with research on associative learning (e.g., Shanks, 1995), and multiple cue probability learning (e.g., Hammond, 1996), where models that assume the independence of causes fit the human data well. One factor likely to encourage this kind of learning was the manner in which the data were presented (e.g., indicator variables followed by outcome variable).

Interventional learning increased sensitivity to the appropriate conditional independencies and eliminated the bias for the collider, but the effect on model selection was not entirely beneficial. Although the correct chain was the modal choice, the majority of participants still chose the wrong model. Taken together with the observational results this implies that participants might have experienced too few trials to confidently discriminate between the models.

Whatever the precise reasons for sub-optimal performance in these tasks, the experiment shows that the automatic mechanisms that allow us to engage in the predictive learning and encoding of noisy information can sometimes override our discovery of the causal model that generates this information. Nevertheless, the difference we did find between observational and interventional learning encourages us that people are able to make use of the special kind of information afforded by intervention, and that future models of learning need to incorporate methods that represent the effect of action.

Acknowledgments

This work was funded by NASA grant NCC2-1217 to Steven Sloman. We thank Sean Stromsten and Dave Sobel for valuable comments.

References

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

Dickinson, A. (2001). Causal learning: an associative analysis. *Quarterly Journal of Experimental Psychology*, *49B*, 60-80.

Glymour, C. (2001). *The mind's arrows.* Cambridge, MA: MIT Press.

Hasher, L., & Zacks, R.T. (1984). The automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist, 39*, 1372-1388.

Hammond, K.R. (1996). *Human judgment and social policy*. New York: Oxford University Press.

Pearl, J. (2000*). Causality*. Cambridge: Cambridge University Press.

Shanks, D.R. (1995). *The psychology of associative learning.* Cambridge: Cambridge University Press.