Learning Causal Direction from Repeated Observations over Time

Benjamin M. Rottman (benjamin.rottman@yale.edu)
Frank C. Keil (frank.keil@yale.edu)
Department of Psychology, Yale U., 2 Hillhouse Ave
New Haven, CT 06520

Abstract

Inferring the direction of causal relationships is notoriously difficult. We propose a new strategy for learning causal direction when observing states of variables over time. When a cause changes state, its effects will likely change, but if an effect changes state due to an exogenous factor, its observed cause will likely stay the same. In two experiments, we found that people use this strategy to infer whether X→Y vs. X←Y, and X→Y→Z vs. X←Y→Z produced a set of data. We explore a rational Bayesian and a heuristic model to explain these results and discuss implications for causal learning.

Keywords: causal reasoning; causal structures; time

Introduction

Learning the direction of a causal relationship from observation is notoriously difficult. Science students are taught that "correlation does not imply causation." Indeed, the standard rational strategy proposed for how people learn causal structures (e.g., Steyvers et al., 2003) suggests that it is impossible to distinguish "Markov equivalent" structures such as [X→Y vs. X←Y] or [X→Y→Z vs. X←Y←Z vs. X←Y→Z]. In the three-variable structures above, all the three variables would be correlated. Additionally, X and Z would be conditionally independent given the state of Y. Because these structures have the same (in)dependence relations among variables, standard theories argue that it is impossible to learn which produced a set of data.

One way that causal direction can be learned from observational data is if there is a temporal delay between the cause and effect (e.g., Lagnado & Sloman, 2006). If Linda gets sick on Monday and Sarah gets sick on Wednesday, it is likely that Linda gave Sarah the cold and not the reverse. However, temporal delay is not always available as a cue to causal direction. Sometimes temporal delays are too short to be perceptible. Or, the learner may only have access to periodic snapshots rather than a continuous stream of data.

Here, we propose another way to learn the direction of causal relationships from observations over time. People may assume that when a cause changes its effects also change, but an effect may change due to an exogenous factor without its observed cause changing. This strategy is illustrated in the following example: Suppose you have two friends, Bill and Tim, and you are trying to figure out whether Bill's mood influences Tim's mood, or the reverse. For eight days, you observe whether Bill and Tim are in positive or negative moods. In Order 1 (Table 1), there are times when both Bill and Tim change between positive and negative moods simultaneously (e.g., Days 1-3), suggesting that there is some causal relationship between the two.

Table 1: Example of How Order Can Influence Inferred Causal Direction Note. 1 stands for a positive mood and 0 stands for a negative mood.

Day	Order 1: Bill→Tim		Order 2: Bill←Tim	
	Bill	Tim	Bill	Tim
1	1	1	1	1
2	0	0	0	0
3	1	1	1	1
4	1	0	0	1
5	1	1	1	1
6	0	0	0	0
7	0	1	1	0
8	0	0	0	0

Additionally, there are transitions in which Tim's mood changes (e.g., Days 3-4) while Bill's mood remains constant. These transitions suggest that some external event occurred to Tim (perhaps he did poorly on an exam), but the fact that Bill did not get into a bad mood on the same day suggests that Tim's mood does not influence Bill's mood.[1] Instead one might conclude that Bill's mood influences Tim's mood, which could explain why sometimes both of their moods change simultaneously.

Critically, according to this account, the *transitions* from day to day are used to infer the causal direction. Consider Order 2, which has the exact same eight days as Order 1, except that Days 4 and 7 are switched. As in Order 1, there are times when both Bill and Tim's moods change simultaneously (e.g., Days 1-3). However, Order 2 has transitions when Bill's mood changes but Tim's mood stays the same (e.g., Days 6-7; perhaps Bill got a job interview). This transition suggests that Bill's mood does not influence Tim's mood, so one might conclude that Tim's mood influences Bill's mood.

All our experiments use manipulations like that in Table 1, in which one set of trials is rearranged in two different orders. If one ignores the temporal sequence of events it would be impossible to determine the causal direction; collapsing across the eight days, there is a correlation of .5, but correlation does not imply a causal direction.

[1] There are also some transitions when Tim's mood changed and Bill's mood stayed constant, but which might not suggest that Tim's mood does not influence Bill's mood. From Trials 4-5, Tim goes from a negative to positive mood, but Bill was already in a positive mood. Such transitions do not necessarily suggest that Tim's mood does not influence Bill's mood because Bill's mood was already at "ceiling" or "floor."

We propose that in certain circumstances, believing that effects are more likely to change by themselves than causes is rational. Specifically, these inferences are rational if the variables are stable across time (i.e. temporally dependent), which is true for many variables such as people's moods. Consider X→Y. When an exogenous event changes the state of X, the change will transfer to Y (depending on how strong the causal relationship is). However, if an exogenous event changes the state of Y, X will stay *stable*; it would be coincidental for X to change at the exact same moment. The following two experiments examine whether people use this temporal strategy to infer causal structures.

Experiment 1

Experiment 1 tested whether people infer causal direction for two-variable causal structures.

Methods

Participants 36 participants from the Yale community were recruited for a 5 minute psychology experiment that paid $2. Participants were recruited at a main pedestrian crossroad on campus, and were tested on a laptop at the same location.

Stimuli The datasets used for the two conditions are presented in Table 2. The same sixteen trials were used in the two conditions and the overall contingency was 0.5.

The sixteen trials were presented in different orders in the two conditions. In the directional condition, Y sometimes changed on its own (e.g., Trials 1-5), which could occur from an external influence on Y. These transitions suggest that Y does not influence X. At other times both X and Y changed simultaneously (e.g., Trials 6-8), which can be explained by an external influence on X that transferred to Y. The temporal strategy suggests X→Y.

In the non-directional condition, X or Y changed by themselves about equally often (e.g., Trials 3-5, or Trials 5-7), which does not identify a causal direction. Additionally, sometimes X and Y stayed the same (e.g., Trials 1-2), and sometimes both X and Y changed simultaneously (e.g., Trials 2-3), suggesting some causal relation.

Procedures Participants read the following cover story:

> "…Please pretend that you are a psychologist studying the moods of roommates. You are trying to figure out if one person's mood influences the other. You might discover that Bill's mood influences Tim's mood, or Tim's mood influences Bill's mood, or both, or that neither influences the other. In the following scenario, you will observe Bill and Tim's moods over a period of 16 consecutive days. Please remember that moods influence one another on the same day. For example, if Bill gets into a negative mood on Monday, and if Bill's mood influences Tim's mood, then Tim will also be in a negative mood on Monday."

These instructions were intended to accomplish a number of goals. First, we intended for the mood transfer scenario to be plausible and an easy way to reason about causal structures. Second, we thought it would be intuitive for moods to be stable across a period of days (temporally dependent variables). Third, we hoped that participants would be able to easily consider possible external influences on peoples' moods (e.g., performing well on a test). The no delay instruction was included because we wanted to discourage this strategy so it did not interfere (Lagnado & Sloman, 2006). Indeed, the stimuli did not have any delay.

Table 2: Summary of Stimuli. Note. 1 represents a "positive mood" and 0 represents a "negative mood". A bolded variable represents a *hypothetical* external influence on this variable that participants may *infer* to explain the changes in the states of the variables. Bold numbers were not denoted in any way for participants.

	Experiment 1				Experiment 2					
	X→Y directional		not directional		X→Y→Z Chain			X←Y→Z Com. C.		
Trial	**X**	**Y**	**X**	**Y**	**X**	**Y**	**Z**	**X**	**Y**	**Z**
1	1	**1**	1	1	0	0	0	0	0	0
2	1	**0**	1	1	0	0	**1**	0	0	**1**
3	1	**1**	0	0	0	0	**0**	0	0	**0**
4	1	**0**	**1**	0	0	**1**	1	**1**	0	0
5	1	**1**	**0**	0	0	**0**	0	**0**	0	0
6	**0**	0	0	**1**	**1**	1	1	1	**1**	1
7	**1**	1	0	**0**	**0**	0	0	1	1	**0**
8	**0**	0	1	1	**1**	1	1	1	1	**1**
9	0	**1**	1	1	1	1	**0**	**0**	1	1
10	0	**0**	0	0	1	1	**1**	**1**	1	1
11	0	**1**	0	**1**	1	**0**	0	0	**0**	0
12	0	**0**	0	0	1	**1**	1	1	**1**	1
13	**1**	1	**1**	0	1	1	**0**	**0**	1	1
14	**0**	0	**0**	0	1	1	**1**	0	1	**0**
15	**1**	1	1	1	**0**	0	0	**1**	1	0
16	**0**	0	1	1	0	0	**1**	1	1	**1**
17					0	0	**0**	0	**0**	0
18					0	**1**	1	**1**	0	0
19					0	1	**0**	1	0	**1**
20					0	**1**	1	**0**	0	1
21					0	**0**	0	0	0	**0**
22					0	0	**1**	1	**1**	1
23					0	0	**0**	**0**	1	1
24					**1**	1	1	**1**	1	1
25					1	**0**	0	0	**0**	0
26					1	0	**1**	**1**	0	0
27					1	0	**0**	**0**	0	0
28					1	**1**	1	0	0	**1**
29					1	1	**0**	0	0	**0**
30					1	1	**1**	1	**1**	1
31					**0**	0	0	1	1	**0**
32					**1**	1	1	1	1	**1**

Next, participants worked with both the conditions in a counterbalanced order. In each scenario, participants were shown a sequence of 16 screens representing 16 consecutive days. On each screen, a plus or minus sign appeared below each person's name signifying their mood as positive or negative. After each screen appeared, there was a delay of 2 seconds, and then participants were prompted to "Press the spacebar to see the next day," at which point the peoples'

moods on the next day appeared. Participants were never told about any external influences; they simply observed the moods of X and Y on 16 consecutive days.

At the end of the 16 days, participants chose one of the following four options: "'No Relationship' means that neither person's mood influences the other; 1→2 means that Person 1's mood influences Person 2's mood; 1←2 means that Person 2's mood influences Person 1's mood; and 1↔2 means that both people's moods influence each other."

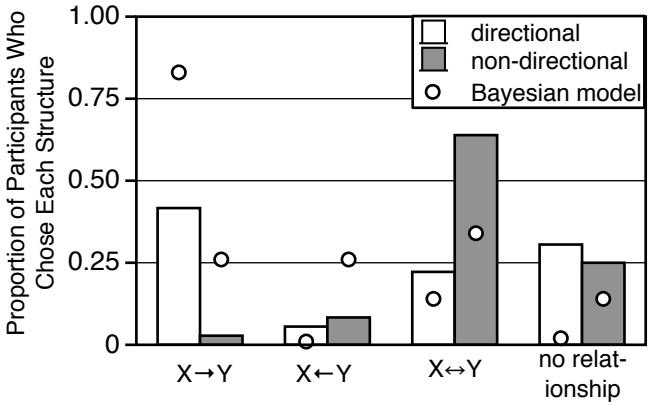


Figure 1: Results of Experiment 1

Results

Figure 1 presents the proportion of participants who chose each of the four response options in the directional and non-directional conditions. We were specifically interested in the hypothesis that participants would infer that X→Y more in the directional than non-directional condition. To test this hypothesis, we collapsed across the other three options. A McNemar test revealed that participants inferred X→Y more in the directional than non-directional condition, $p<.01$. As can be seen in Figure 1, most of the participants who inferred X→Y in the directional condition inferred X↔Y in the non-directional condition. This makes sense because in the non-directional condition there were times when both X and Y simultaneously changed (e.g., Trials 2-3 in Table 2). Such transitions would suggest that X and Y are correlated, but would not suggest a specific direction.

In sum, people readily learned the direction of causal relationships from observing changes in states over time. We believe that this is the first demonstration of such an ability for stimuli without a temporal delay.

Experiment 2

Experiment 2 tested whether people can also use changes in states over time to learn directional causal structures among three variables. We tested whether people could differentiate a chain structure (X→Y→Z) from a common cause structure (X←Y→Z). Non-temporal theories cannot distinguish these structures (see the introduction). However, if people use the temporal strategy they may be able to distinguish these two structures. For X→Y→Z, when X changes, Y and also Z will likely change. But for X←Y→Z, when X changes, Y and Z will likely stay constant.

Methods

Participants There were 28 participants from the same population as Experiment 1.

Design and Stimuli All participants worked with the chain and common cause conditions in a counterbalanced order. Both conditions had the same set of 32 trials with different orders (Table 2). The 32 trials were determined based on causal structures with the following parameters. Exogenous variables (X for the chain and Y for the common cause) had a base rate of .5. If a cause was present, its effects would be present with a probability of .75. If a cause was absent, its effect would be present with a probability of .25. Thus, Delta P, a measure of contingency, was .5. Because the two graphs are Markov equivalent, they have the same 32 trials.

In the chain condition X→Y→Z, the trials were ordered such that sometimes Z changed by itself, sometimes Y and Z changed together and sometimes X, Y, and Z all changed together. If participants use the temporal strategy, they might attribute Z changing state by itself (e.g., Trials 1-3) to an unobserved influence on Z, and also infer that Z does not influence X or Y. They might interpret Y and Z changing together (e.g., Trials 4-5) as evidence of an unobserved influence on Y (after all Z does not appear to influence Y), which further influences Z. Finally, they might interpret all three variables changing state together (e.g., Trials 6-8) as evidence of an unobserved influence on X (the above transitions suggest that neither Y nor Z influences X), which influences Y and Z.

In the common cause structure X←Y→Z, sometimes X and Z changed state by themselves and sometimes all three variables change state together. If participants use the temporal strategy, they would interpret X and Z changing state by themselves (e.g., Trials 2-3, 4-5) as evidence of an unobserved influence on X or Z, and also infer that neither X nor Z influences any of the other two variables. They might interpret transitions when all three variables change state simultaneously (e.g., Trials 11-12) as evidence of an unobserved influence on Y, which in turn influences X and Z (X and Z seem not to influence Y).

Procedure Participants were first introduced to the emotion transmission cover story used in Experiment 1 with the following modifications. The instructions were changed to contain three friends instead of two. Participants were told at the beginning about chains ("one person's mood influences a second person's mood, which in turn influences a third person's mood") and common causes ("there is one main person whose mood influences both other people"). Participants were also told that their goal was to determine which graph best describes these three friends.[2]

[2] To encourage participants to think of moods as states that are stable over days, they were told: "People can also stay in good moods or bad moods for a period of days." Finally, to encourage them to think about unobserved events that may have manipulated the peoples' moods, which we believed would facilitate using the temporal strategy, participants were instructed: "On each day, please consider possible events that influenced peoples' moods. For

After reading the instructions, participants worked with both the common cause and chain scenarios in a counter-balanced order. Participants then observed the 32 days; after observing each day, they were prompted to press the spacebar to observe the next day. The chain and common cause graphs were shown during the entire scenario, and at the end, participants chose whether the scenario was best described by the chain (Person 1→Person 2→Person 3) or common cause (Person 1←Person 2→Person 3).

Results

Seventy-one percent (20 out of 28) of the participants chose the chain structure in the chain condition, which is significantly above chance, p=.01. Seventy-five percent (21 out of 28) of the participants chose the common cause structure in the common cause condition, p<.01. A McNemar test suggests that participants more often chose the chain structure in the chain condition than in the common cause condition, p<.01.

Models

Temporal Bayesian Model

We propose a Bayesian model to demonstrate how it is possible to *rationally* infer X→Y or X←Y by observing variables over time. We define a state as the current values of X and Y (e.g, [x=1, y=0] condensed as [10]). A transition is two consecutive states (e.g., [10 to 11]). With two variables, there are four states and 16 types of transitions.

The basic idea behind the model is that different causal structures (graphs) produce different types of transitions. It is possible to infer the probability of a graph g given the observed transitions t by using Bayes theorem to invert the probability of a graph producing the transitions.

$$P(g|t) \propto P(t|g)P(g) \qquad \text{(Eq. 1)}$$

We assume that any transition must be caused by an exogenous event that influenced one of observed variables. Given a particular state, there are three possible unobserved influences; an influence changing the state of X, an influence changing the state of Y, or simultaneous influences that change the states of both X and Y. We define i_x, i_y, and i_{xy}, as the probability of these three influences and they are mutually exclusive and exhaustive; $i_x+i_y+i_{xy}=1$. Because we have no reason to believe that X or Y is more likely to be influenced by external variables than the other, we assume that $i_x=i_y$, thus, $i_{xy}=1-2i_x$. We use "s" to refer to the causal strength, or the likelihood of an influence on a cause producing a change in the effect (assuming that the effect is not already at ceiling or floor).

Here we focus on the likelihood of the graph X→Y producing a particular transition. Figure 2a presents the transitions probabilities for X→Y in a Markov chain; the corners are the four states and the arrows represent the 16 transitions with associated probabilities (also see Figure 2b).

Suppose that the current state is [00]. The transition [00 to 11] could arise if X is turned on and succeeds in turning on Y ($i_x s$), or by simultaneous influences on X and Y (i_{xy}). (Hereon transition probabilities are in parentheses.) The transition [00 to 01] arises from an influence on Y (i_y). An influence that changes the state of Y cannot transfer to X because Y is the effect, so this transition probability does not include s. The transition [00 to 10] arises from of an influence on X *failing* to produce a change in Y ($i_x(1-s)$). As s becomes stronger, there would be fewer transitions when the cause fails to produce a change in the effect.

Consider the initial state [01]. The transition [01 to 11] arises from an influence on X (i_x). Because Y was already on (a "ceiling effect"), X cannot have any influence on Y, so the transition probability does not include s. The transition [01 to 00] arises from an influence on Y (i_y). The transition [01 to 10] arises from simultaneous influences on X and Y (i_{xy}). Since X and Y were both changed, the influence on X cannot have any influence on Y, so this probability does not include s.

Note that given a particular causal structure and an initial state, there are only three possible transitions. From [00], there can be a transition to [01], [10], or [11], and these respective transition probabilities sum to 1. We treat transitions to the same state as a single trial (e.g., [00 to 00] as just [00]).[3] Also, the transition probabilities with initial states [11] and [10] can be deduced from [00] and [01] discussed above; Transitions 1-8 in Figure 2b are the mirror image of Transitions 9-16. The transition probabilities for X←Y can be deduced from X→Y by switching X and Y.

How can the model infer whether X→Y or X←Y produced the observed data? Consider a transition when Y changes by itself [e.g., 00 to 01] (Fig. 2b row 15). Under the structure X→Y, the likelihood of this transition is just the probability of an external influence on Y (i_y). Under X←Y the likelihood is the probability of an influence on Y multiplied by the probability of Y *failing* to change X ($i_y(1-s)$). Effects are more likely to change by themselves than causes even if s is weak; $i_y > i_y(1-s)$.[4]

For the bidirectional graph X↔Y, X and Y change by themselves equally often. When the state of X is changed by an exogenous factor, it is no longer influenced by Y, so the transition probabilities are the same as X→Y. When the state of Y is changed by an external factor, the probabilities are the same as X←Y. For the graph in which X and Y are unrelated (no links), the likelihood of either X or Y changing by itself is the probability of an influence on that

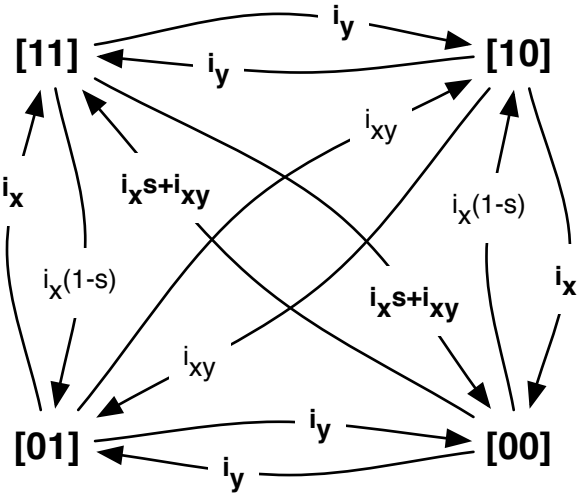
example, Allison may have performed well on a test, which put her into a good mood, and spread to Bill.…"

[3] The reason is that in such cases there is no implied unobserved influence and these transitions do not discriminate between causal structures. Additionally, any temporally extended state could be parsed into infinitely many periods. Note that transitions to the same state did sometimes occur in the stimuli.

[4] If an effect is at ceiling or floor [e.g., 01 to 11] the transition has the same probability under X→Y and X←Y (see Figure 2b).

variable (i_x or i_y). If both change, the probability is i_{xy}.

a: Markov Chain with Transition Probabilities for X→Y



b: Transition Probabilities for the Four Structures with Two Variables

Trans. Type	Observed Data t x	t y	t+1 x	t+1 y	Graph X→Y	X←Y	No Link	X↔Y
1	1	1	1	1	0	0	0	0
2	1	1	1	0	i_y	$i_y(1-s)$	i_y	$i_y(1-s)$
3	1	1	0	1	$i_x(1-s)$	i_x	i_x	$i_x(1-s)$
4	1	1	0	0	$i_x s+i_{xy}$	$i_y s+i_{xy}$	i_{xy}	$i_x s+i_y s+i_{xy}$
5	1	0	1	1	i_y	i_y	i_y	i_y
6	1	0	1	0	0	0	0	0
7	1	0	0	1	i_{xy}	i_{xy}	i_{xy}	i_{xy}
8	1	0	0	0	i_x	i_x	i_x	i_x
9	0	1	1	1	i_x	i_x	i_x	i_x
10	0	1	1	0	i_{xy}	i_{xy}	i_{xy}	i_{xy}
11	0	1	0	1	0	0	0	0
12	0	1	0	0	i_y	i_y	i_y	i_y
13	0	0	1	1	$i_x s+i_{xy}$	$i_y s+i_{xy}$	i_{xy}	$i_x s+i_y s+i_{xy}$
14	0	0	1	0	$i_x(1-s)$	i_x	i_x	$i_x(1-s)$
15	0	0	0	1	i_y	$i_y(1-s)$	i_y	$i_y(1-s)$
16	0	0	0	0	0	0	0	0

Figure 2: Transition Probabilities. Note probabilities in bold highlight the transitions that are likely under a given graph.

$$P(g|t) \propto \iint_0^1 \prod_{m=1}^{16} P(t_m|g, i_x, s)^{N(t_m)} P(i_x, s|g)\, di_x\, ds \quad \text{(Eq. 2)}$$

The full model is represented in Eq. 2. The likelihood ($t|g$) of a graph producing the observed sequence of transitions ("transition path") is the product of the probabilities of each of the individual transitions. This can be simplified as the product of the probabilities of the 16 transition types $t_{m=1...16}$ each raised to the power of the number of transitions of that type ($N[t_m]$). There are only two parameters, s and i_x (note that $i_y=i_x$ and $i_{xy}=1-2i_x$), which are integrated over. We assume that the prior distribution $P(i_x,s|g)$ is uniform.

This model can be extended to three-variable causal structures using analogous reasoning. For brevity we do not explain all the transition probabilities, which can be obtained from the authors. For a brief example, consider a transition in which only X changes but Y and Z do not [000 to 100]. Under graph X←Y→Z, this transition would arise from an unobserved influence on X (i_x). But under graph X→Y→Z, this transition would require an influence on X that failed to transfer to Y ($i_x(1-s)$), which is less likely.

The model can also be extended to negative relations.[5]

Heuristic Model to Learn Causal Direction

It is possible that people use a heuristic approximation of the Bayesian model like the one presented here. For two variables X and Y, the model produces a score for the two links X→Y and X←Y; the higher the score, the more likely that the link exists. For each transition, the model runs the following function. If both variables changed (e.g., [00 to 11]), then the model adds 1 to the scores for both links because this is evidence of some causal relationship. If one variable changed (e.g., [00 to 01]) then the model subtracts 1 from the structure in which the variable that changed is the cause, in this case X←Y. The reason is that when Y changed to 1, X failed to change, which suggests that Y does not cause X. However, if one variable was already at ceiling or floor when the other changed [e.g., 01 to 00], then the scores are left unchanged. This model can be generalized for more variables by considering all links between all variables. In further work, we have demonstrated that the heuristic model approximates the Bayesian model in a wide variety of instances.[6]

Model Simulations of Experiments 1 and 2

Experiment 1 The heuristic model captured the essential difference between the two conditions. In the directional condition, X→Y had a higher score (7) than X←Y (3). In the non-directional condition, both graphs had a score of 2.

We used the Bayesian model to calculate the relative likelihoods of the four two-variable graphs producing the observed transitions (Figure 1). Both the model and our participants inferred that X→Y was more likely to have produced the directional data than X←Y. Additionally, both our participants and the model choose X↔Y more in the non-directional condition than directional condition. The

[5] Even for negative relationships, when a cause changes it would usually produce a change in its effect, but an effect can change by itself. Suppose that X→Y and X=1, and Y=0. If an exogenous factor sets X=0, Y would likely change to 1. However, if an exogenous factor sets Y=1, X would likely stay at 1.

[6] One difference can occur for X→Y→Z (see also, Fernbach & Sloman, 2009). The heuristic model would infer that X *directly* influences Z. The Bayesian model could infer whether there is a direct link from X to Z or not by testing these two alternative structures. X and Z changing simultaneously without Y is evidence in favor of the direct link. If X and Z do not change together without Y changing, this suggests that there is no direct link.

model does diverge from our participants in a number of ways. However, this is a *rational* model of the task, which need not coincide with human performance. Still, the model does predict the most critical difference between X→Y vs. X←Y. In further work we have examined how well the model fits individual participants' inferences.

Experiment 2 The heuristic model captured participants' inferences. For the chain X→Y→Z, the model gave a higher score for X→Y (7) than X←Y (3) and a higher score for Y→Z (15) than Y←Z (7). The model also gave a higher score for X→Z (7) than X←Z (-5). For the common cause condition X←Y→Z, the model gave higher scores for X←Y (7) than X→Y (1) and a higher score for Y→Z (7) than Y→Z (1). The model also had a somewhat higher score for X←Z (3) than X→Z (-1).

The Bayesian model also predicted the results. To simulate Experiment 2, we computed the relative likelihood of X←Y→Z vs. X→Y→Z. In the chain condition, the likelihood of the chain was 99%, and in the common cause condition, the likelihood of the common cause was 98%.

General Discussion

We proposed a theory for how people learn directional causal relationships by observing states of variables change over time. Experiment 1 demonstrated that people infer that X influences Y when Y changed more frequently by itself than X, and sometimes X and Y changed simultaneously. Experiment 2 demonstrated that given three variables, if X changes by itself, people tend to infer X←Y→Z, but if X usually changes with Y and Z, people tend to infer X→Y→Z. We also proposed two models. The heuristic model embodies the belief that effects are more likely to change by themselves than causes. The Bayesian model proposes a rational reason why people might adopt this belief. The Bayesian model assumes that the states of variables are stable across time and only change if an exogenous variable produces the change. Consider X→Y. When an exogenous factor changes X, the larger the causal strength, the more likely Y will also change. But when an exogenous factor changes Y, it would be coincidental for X to change simultaneously.

Previous research has demonstrated that people can learn causal structure from interventions, from atemporal observations (but cannot distinguish Markov-equivalent structures such as the structures used here), and from observations with delay (e.g., Lagnado & Sloman, 2006; Steyvers et al., 2003). There are likely many ways that people learn causal structures (Lagnado et al., 2007); the current strategy differs in two key ways from previous theories. First, most previous experiments have used punctate events that either happened or didn't happen on a given trial, but the variables in the current experiments had temporally extended states. Second, in most previous research, the trials were temporally independent, often randomized. In contrast, the states of the variables in the current experiments may stay stable for periods of time (e.g., in the X→Y condition in Experiment 1, X stayed stable from Trial 1-5). When a relatively stable variable changes state, it is possible to make inferences about the causes of the *change*. For example, if Bill and Tim were both in a good mood and then both get into a bad mood *at the same time*, this is strong evidence that there is some causal relation between the two. Alternatively, if Bill gets into a bad mood but Tim stays in a good mood, this transition suggests that Bill's mood does not influence Tim's mood. Thus, the current theory focuses on *transitions* between states rather than individual trials.

What implications should be drawn from the proposed models? Given that both models captured the basic asymmetry between causes and effects, and no other models can discriminate the difference, the current approaches appear promising. In other research, we have found that the models are highly correlated,[6] suggesting that the heuristic model approximates a rational strategy. However, given that there are some differences between the rational model and participants' responses, alternative heuristic models may provide insight to participants' reasoning strategies. These models also suggest some future areas of exploration. Do people infer causal strength along with structure? Can people learn negative causal relationships from observing variables over time?[5] And how can these models be generalized to multi-valued variables?

These experiments demonstrate that there is rich structure in how events unfold over time, and people readily identify the structure in these temporal patterns. In the real world, variables often are stable or temporally dependent. For example, after one observes a dent in one's car, the dent remains for a period of time until it gets fixed. Given the importance of causal reasoning and the fact that we experience the world temporally, the abilities demonstrated here may reflect common and vital learning processes.

Acknowledgments

This research was supported by an NSF Graduate Research Fellowship (Rottman) and NIH grant R37HD023922 (Keil).

References

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of experimental psychology: Learning, memory, and cognition*, *35*(3), 678-93.

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory and Cognition, 32*(3), 451-460.

Lagdano, 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. 154-72). Oxford: Oxford University Press.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27,* 453-489.