The Acquisition and Use of Causal Structure Knowledge

Benjamin Margolin Rottman

Learning Research and Development Center
University of Pittsburgh
3939 O'Hara Street
Pittsburgh PA 15260

In M.R. Waldmann (Ed.), *Oxford Handbook of Causal Reasoning.* Oxford: Oxford U.P.

Author Note: This research was supported by NSF BCS-1430439. I thank Michael Waldmann and Bob Rehder for providing helpful comments and suggestions.

Abstract

This chapter provides an introduction to how humans learn and reason about multiple causal relations connected together in a causal structure. The first half of the chapter focuses on how people learn causal structures. The main topics involve learning from observations vs. interventions, learn temporal vs. atemporal causal structures, and learning the parameters of a causal structure including individual cause-effect strengths and how multiple causes combine to produce an effect. The second half of the chapter focuses on how individuals reason about the causal structure, such as making predictions about one variable given knowledge about other variables, once the structure has been learned. Some of the most important topics involve reasoning about observations vs. interventions, how well people reason compared to normative models, and whether causal structure beliefs bias reasoning. In both sections I highlight open empirical and theoretical questions.

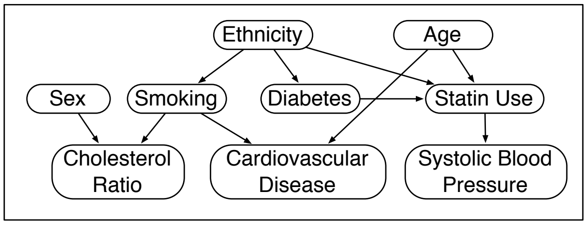
Keywords: Causal Structure Learning, Causal Reasoning, Causal Bayesian Networks

1 Introduction

In the past two decades, psychological research on casual learning has been strongly influenced by a normative framework developed by statisticians, computer scientists, and philosophers called Causal Bayesian Networks (CBN) or probabilistic directed acyclic graphical models. The psychological adoption of this computational approach is often called the CBN framework or causal models. The CBN framework provides a principled way to learn and reason about complex causal relations among multiple variables.

For example, Thrornley (2013) used causal learning algorithms to extract the causal structure in Figure 1 from medical records. Having the causal structure is useful for experts such as epidemiologists and biologists to understand the disease and make predictions for groups of patients (e.g., the likelihood of having cardiovascular disease among 70 year old smokers). It is also useful for scientists when planning future research; when researching cardiovascular disease as the primary outcome it is critical to measure and account for smoking status and age; it is not important to measure or statistically control for systolic blood pressure.

Figure 1: Causal Structure of Cardiovascular Disease (adapted from Thornley, 2013)



Though causal structures are surely useful for scientists, the causal models approach to causal reasoning hypothesizes that lay people also have an intuitive understanding of causal structures and use a framework similar to CBNs to learn and reason about causal relations. Adopting a “man as intuitive statistician” or “intuitive scientist” approach (Peterson & Beach, 1967), we can also contemplate how a doctor might develop a set of causal beliefs about cardiovascular disease somewhat akin to Figure 1. Of course the doctor likely has some knowledge of specific causal links from medical school and research articles. But these links may be reinforced or contradicted by personal experience such as noticing which patients have which symptoms and diseases, and tracking how patients' symptoms change after starting a treatment. Developing a set of causal beliefs such as in Figure 1 would allow a physician to make prognoses and treatment plans tailored to individual patients.

The CBN framework supports all of these different functions: learning, prediction, explanation, and intervention. The rest of this chapter will explain what the CBN framework entails, the evidence pertaining to how people learn and reason about causal networks, and how closely humans appear to mimic the normative CBN framework.

The outline of this chapter is as follows. I first explain what CBNs are, both normatively and as a model of human learning and reasoning. The bulk of the first half of this chapter is devoted to evidence about how people learn about causal networks including the structure, strength, and integration function. I then discuss evidence suggesting that instead of using the basic CBN framework, people may be using something akin to a generalized version of the CBN framework that allows for reasoning about time. The second half of the chapter is devoted to evidence on how people reason about their causal beliefs. At the end of the chapter I raise some questions for future research.

1.1 What Are Causal Bayesian Networks?

A Causal Bayes Network is a compact visual way to represent the causal relations between variables. Each variable is represented as a node, and arrows represent causal relations from causes to effects. The absence of an arrow implies the absence of a causal relation.

Though Causal Bayesian Networks capture the causal relations among variables, they also summarize the statistical relations between the variables. The CBN framework explains how causal relations should be learned from statistical relations. For example, given a dataset with a number of variables, the CBN framework has rules for figuring out the causal structure(s) that are most likely to have produced the data. Conversely, a causal structure can be read such that if the causal structure is believed to be true, it implies certain statistical relations between the variables; that some sets of variables will be correlated and others will not be correlated. In order to understand how to "read" a CBN, it is important to understand these relations between the causal arrows and the statistical properties they imply.

First, it is critical to understand some basic statistical terminology. "Unconditional dependence" is whether two variables are statistically related to (or "dependent on") each other (e.g., correlated) without controlling for any other variables. If they are correlated, they are said to be dependent, and if they are not correlated, they are said to be independent. Conditional dependence is whether two variables are statistically related to each other after controlling for one or more variables (e.g., whether there is a significant relation between two variables after controlling for a third variable in a multiple regression). Conditional and unconditional dependence and independence are critical for understanding a CBN, so it is important to be fluent with these terms before moving on.

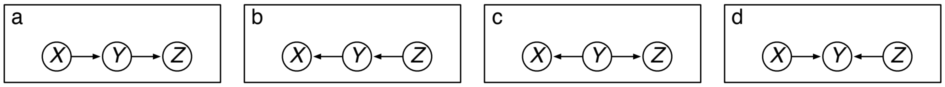
There are two properties, the Markov property, and the faithfulness assumption, that explain the relations between the causal arrows in a CBN and the statistical dependencies between the variables in a dataset that is summarized by the CBN (also see Rehder, this volume a and b). The Markov property states that once all the direct causes of a variable *X* are controlled for or held constant, *X* is statistically independent of every variable in the causal network that is not a direct or indirect effect of *X*.

For example, consider Figure 2b. The Markov assumption states that *X* will be independent of all variables (e.g., *Z*) that are not direct or indirect effects of *X* (*X* has no effects in Figure 2b) once controlling for all direct causes of *X* (*Y* is the only direct cause of *X*). Similar analyses can be used to see that *X* and *Z* are also independent conditional on *Y* in Figures 2a and 2c.

In regards to Figure 2d, the Markov Assumption implies that *X* and *Z* are un conditionally independent (not correlated). Neither *X* nor *Z* have any direct causes, so *X* will be independent of all variables (such as *Z*), that are not a direct or indirect effect of *X* (e.g., *Y*). The Markov property is symmetric; if *X* is independent of *Z*, *Z* is independent of *X* - they are uncorrelated.

The faithfulness Assumption states that the only independencies between variables in a causal structure must be those implied by the Markov assumption (Glymour, 2001; Spirtes, Glymour, & Scheines, 1993). Stated another way, all variables in the structure will be dependent (correlated), except when the Markov property states that they would not be. This means that if we collect a very large amount of data from the structures in Figures 2a, 2b, or 2c, *X* and *Y*, and *Y* and *Z*, and *X* and *Z* would all be unconditionally dependent; the only independencies between the variables arise due to the Markov assumption, that *X* and *Z* are conditionally independent given *Y*. If we collected a large amount of data and noticed that *X* and *Z* were unconditionally independent, this independency in the data would not be "faithful" to Figures 2a, 2b, or 2c, implying that the data do not come from one of these structures. For Figure 2d, the only independency implied by the Markov property is that *X* and *Z* are unconditionally independent. If a large amount of data were collected from structure 2d, *X* and *Y*, and *Z* and *Y* would be dependent (according to the faithfulness assumption).

Figure 2: Four CBNs



In sum, causal models provide a concise, intuitive, visual language for reasoning about complex webs of causal relations. The causal network diagram intuitively captures how the variables are causally and statistically related to each other. But causal networks can do much more than just describe the qualitative causal and statistical relations; they can precisely capture the quantitative relations between the variables.

To capture the quantitative relations among variables, causal networks need to be specified with a conditional probability distribution for each variable in the network given its direct causes. A conditional probability distribution establishes the likelihood that a variable such as *Y* will have a particular value given that another variable *X, Y*s cause, has a particular value. Additionally, exogenous variables, variables that have no known causes in the structure, are specified by a probability distribution representing the likelihood that the exogenous variable assumes a particular state.

For example, the CBN in Figure 2a would be specified by a probability distribution for *X*, a conditional distribution of *Y* given *X*, and a conditional distribution of *Z* given *Y*. If *X*, *Y*, and *Z* are binary (0 or 1) variables, the distribution for *X* would simply be the probability that *x*=1, *P*(*x*=1). The conditional probability of *Y* given *X* would be the probability that *y*=1 given that *x*=1, *P*(*y*=1|*x*=1), and the probability that *y*=1 given that *x*=0, *P*(*y*=1|*x*=0), and likewise for the conditional probability of *Z* given *Y*. (There is also another way to specify these conditional distributions with "causal strength" parameters, which will be discussed in later sections, and summarized in Section 3.6. See also Cheng & Lu, this volume; Griffiths, this volume; Rehder, this volume, a and b, for more details about parameterizing a structure with causal strengths.)

If the variables are normally-distributed continuous variables, the distribution for *X* would be captured by the mean and standard deviation of *X*. Then, the conditional distribution of *Y* would be captured by a regression coefficient of *Y* given *X* (e.g., the probability that *y*=2.3 given that *x*=1.7), as well as a parameter to capture the amount of error variance.

The CBN in Figure 2c would be specified by a distribution for *Y*, and conditional probability distributions for *X* given *Y* and *Z* given *Y*. The CBN in Figure 2d would be specified by probability distributions for *X* and *Z*, and a conditional probability distribution for *Z* given both *X* and *Y*. In this way, a large causal structure is broken down into small units.

Once all the individual probability distributions are specified, Bayesian inference can be used to make inferences about any variable in the network given any set of other variables, for example, the probability that *y*=3.5 given that *x*=-0.7 and *z*=1.1. CBNs also support inferring what would happen if one could intervene and set a node to a particular value. Being able to predict the result of an intervention allows an agent to choose the action that produces the most desired outcome.

CBNs have been tremendously influential across a wide range of fields including computer science, statistics, engineering, epidemiology, management sciences, and philosophy (Pearl, 2000; Spirtes, Glymour, & Scheines, 2000). The CBN framework is extremely flexible and supports many different sorts of tasks. They can be used to make precise predictions (including confidence intervals), they can incorporate background knowledge or uncertainty (e.g., uncertainty in the structure, or uncertainty in the strengths of the causal relations) for sensitivity analysis. They can be extended to handle processes that occur over time. And since CBNs are an extension of probability theory, they can incorporate any probability distribution (logistic, multinomial, Gaussian, exponential). In sum, the CBN framework is an extremely flexible way to represent and reason about probabilistic causal relations.

1.2 What is the Causal Bayesian Network Theory of Learning and Reasoning?

Most generally, the causal model theory of human learning and reasoning is that humans learn and reason about causal relations in ways that are similar to formal CBNs. This theory is part of a broader movement in psychology of using probabilistic Bayesian models as models of higher-level cognition.[1]

[1] Here "model" is serving two purposes. First probabilistic Bayesian models are intended to be objective models of how the world works (e.g., Figure 1 is an objective model of cardiovascular disease). The second sense of model, as used by the psychologist, is that the same probabilistic model could also serve as a model of human reasoning – treating Figure 1 as a representation of how a doctor thinks about cardiovascular disease.

The broader movement of using probabilistic models as models of higher-level cognition is typically viewed at Marr's computational level of analysis – identifying the problem to be solved. Indeed, articles appealing to causal networks have fulfilled the promise of a computational-level model – for example, they have reframed the problem of human causal reasoning by clarifying the distinction between causal strength vs. structure (Griffiths & Tenenbaum, 2005), and by identifying causal structure learning as a goal unto itself (Alison Gopnik et al., 2004; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003).

Though the flexibility of the CBN framework is obviously a tremendous advantage for its utility, the flexibility makes it challenging to specify a constrained descriptive theory of human learning and reasoning. The theoretical underpinning of the CBN framework (e.g., learning algorithms, inference algorithms) is an active area of research rather than a static theory. Additionally, there are many different instantiations of how to apply the framework in a specific instance (e.g., alternative learning algorithms, alternative parameterizations of a model).

Because of the flexibility and multifaceted nature of the CBN framework, it is not particularly useful to talk about the CBN framework as a whole. Instead, in the current chapter I focus on the fit between specific aspects of the framework and human reasoning, within a specific task.

2 Learning

2.1 Learning Causal Structure

2.1.1 Learning a Causal Structure from Observations

One of the most dramatic ways that the CBN framework has changed the field of human causal reasoning is by identifying causal structure learning as a primary goal for human reasoning (Steyvers et al., 2003). A fundamental principle of learning causal structure from observation is that it is often not possible to identify the exact causal structure; two or more structures may explain the data equally well. This is essentially a more sophisticated version of "correlation does not imply causation."

Consider the 9 observations in
Table 1, which summarizes the contingency between two variables, *X* and *Y.* The correlation between these two variables is .79. Just knowing that *X* and *Y* are correlated cannot tell us whether *X* causes *Y* [*X*→*Y*] or whether *Y* causes *X* [*X*←*Y*]. (Technically it is also possible that a third factor causes both *X* and *Y*, but I ignore this option for simplicity of explanation.) One simple way to see this is that under both of these causal structures [*X*→*Y*] and [*X*←*Y*], we would expect *X* and *Y* to be correlated, so the fact that they are correlated cannot tell us anything about the causal structure. A more sophisticated way to understand how it is impossible to determine the true causal structure just from observing the data in Table 1 is to see how parameters can be created for both causal structures that fit the data equally well. This means that the structures are equally likely to have produced the data.

The right side of Table 1 shows parameters for the respective causal structure that fit the data perfectly. For example, for [$X \rightarrow Y$], we need to find three parameters to specify the structure. The base rate of X, $P(x=1)$, can be obtained by calculating the percentage of times that $X=1$ regardless of Y, ((3+1)/9). $P(y=1|x=1)$ is simply the percent of times that $Y=1$ given that $X=1$, and $P(y=1|x=0)$ is the percent of times that $Y=1$ given that $X=0$. Parameters can be deduced for $X \leftarrow Y$ in a similar fashion. If we simulated a large number of observations that we would expect to see from each causal structure with the parameters specified in Table 1, we would find that both structures would produce data with proportions that looks similar to the data in Table 1. Specifically, we would observe trials in which both variables are 1 about 3 out of 9 times, trials in which both variables are 0 about 5/9 times, and trials in which $X=1$ and $Y=0$ about 1 in 9 times in the long run. Because we were able to find parameters for these structures that produce data very similar to the data we observed, these two structures are equally likely given the observed data.

Table 1: Sample Data for Two Variables

X	Y	Number of Observations	Parameters that fit the data perfectly for each causal structure	
			$X \rightarrow Y$	$X \leftarrow Y$
1	1	3	$P(x=1)=4/9$ $P(y=1\mid x=1)=3/4$ $P(y=1\mid x=0)=0$	$P(y=1)=3/9$ $P(x=1\mid y=1)=1$ $P(x=1\mid y=0)=1/6$
1	0	1		
0	1	0		
0	0	5		

The same logic also applies with more variables. Consider the data in Table 2 with three variables. If you ran a correlation between each pair of variables you would find that X and Y are correlated (r=.25), Y and Z are correlated (r=.65), and X and Z are correlated (r=.17) but are independent (r=0) once Y is controlled for. According to the Markov and Faithfulness assumptions, this pattern of dependencies and conditional independencies is consistent with three and only three causal structures; $X \rightarrow \mathrm{Y} \rightarrow \mathrm{Z}$, $X \leftarrow Y \leftarrow Z$, and $X \leftarrow Y \rightarrow Z$.

Table 2 shows parameters for each of these causal structures that fit the data perfectly. If we sampled a large amount of data from any of the three structures in Table 2 with the associated parameters, the proportion of the types of 8 observations of X, Y, and Z would be very similar to the proportions of the number of observations in Table 2. These three causal structures are said to form a Markov class because they are all equally consistent with (or likely to produce) the set of conditional and unconditional dependencies in the observed data. Thus, it is impossible to know which of these three structures produced the set of data in Table 2.

Table 2: Sample Data for Three Variables

X	*Y*	*Z*	Number of Observations	Parameters that fit the data perfectly for each causal structure		
				$X\rightarrow \mathrm{Y}\rightarrow Z$	$X\leftarrow \mathrm{Y}\leftarrow Z$	$X\leftarrow Y\rightarrow Z$
1	1	1	6	$P(x=1)=1/2$ $P(y=1\|x=1)=3/4$ $P(y=1\|x=0)=1/2$ $P(z=1\|y=1)=2/3$ $P(z=1\|y=0)=0$	$P(z=1)=5/12$ $P(y=1\|z=1)=1$ $P(y=1\|z=0)=5/14$ $P(x=1\|y=1)=3/5$ $P(x=1\|y=0)=1/3$	$P(y=1)=5/8$ $P(x=1\|y=1)=3/5$ $P(x=1\|y=0)=1/3$ $P(z=1\|y=1)=2/3$ $P(z=1\|y=0)=0$
1	1	0	3			
1	0	1	0			
1	0	0	3			
0	1	1	4			
0	1	0	2			
0	0	1	0			
0	0	0	6			

Importantly, non-Markov equivalent causal structures can be distinguished from one another with observations. For example, common effect structures such as $X\rightarrow Y\leftarrow Z$ are in their own Markov equivalence class, so they can be uniquely identified. According to the Markov assumption for $[X\rightarrow Y\leftarrow Z]$, *X* and *Y* are dependent, and *Z* and *Y* are dependent, but *X* and *Z* are independent. There are no other three-variable structures with this particular set of conditional and unconditional dependencies. This means that even through $X\rightarrow \mathrm{Y}\rightarrow Z$, $X\leftarrow \mathrm{Y}\leftarrow Z$, and $X\leftarrow Y\rightarrow Z$ are all equally likely to produce the data in Table 2, $X\rightarrow Y\leftarrow Z$ is much less likely. Suppose we tried to find parameters for $X\rightarrow Y\leftarrow Z$ to fit the data in Table 2. It would be possible to choose parameters such that *Y* and *X* are correlated roughly around *r*=.25, and that *Y* and *Z* are correlated roughly around *r*=.65, matching the data in Table 2 fairly closely. But critically, we would find that no matter what parameters we chose, *X* and *Z* would always be uncorrelated, and thus, it would be very unlikely that the data from Table 2 would come from $X\rightarrow Y\leftarrow Z$.

In sum, by examining the dependencies between variables it is possible to identify which types of structures are more or less likely to have produced the observed data. Structures within the same Markov equivalence class always have the exact same likelihood of producing a particular set of data, which means that they cannot be distinguished, but structures from different Markov equivalence classes have different likelihoods of producing a particular set of data.

Steyvers et al., (2003) conducted a set of experiments to test whether people understand Markov equivalence classes and could learn the structure from purely observational data. First, they found that given a particular set of data, participants were above chance at detecting the correct Markov class. Furthermore, people seem to be fairly good at understanding that observations cannot distinguish $X\rightarrow \mathrm{Y}$ from $X\leftarrow \mathrm{Y}$. And people also seem to understand to some extent that common effect structures $X\rightarrow Y\leftarrow Z$ belong to their own Markov equivalence class.

However, Steyvers et al.'s participants were not good at distinguishing chain and common cause structures even when they were from different equivalence classes (e.g., $X\rightarrow Y\rightarrow Z$ vs. $X\rightarrow Z\rightarrow Y$). Distinguishing these structures was made more difficult in the experiment because the most common type of observation for all these structures was for the three variables to have the same state. Still, the participants did not appear to use the trials when two of the variables share a state different from a third to discriminate causal structures (e.g., the observation X=Y≠Z is more consistent with X→Y→Z than X→Z→Y).

Given that Markov equivalence class is so important for theories of causal structure learning from observation, it is surprising that there is not more work on how well lay people understand Markov equivalence. One important future direction would be give participants a set of learning data that unambiguously identifies a particular Markov-equivalent class, and test the percent of participants who 1) identify the correct class, 2) identify all the structures in the Markov class, and 3) include incorrect structures outside the class. Such an experiment would help clarify how good or bad people are at learning causal structure from observations. Additionally, Steyvers et al. used categorical variables with a large number of categories and nearly deterministic causal relations, which likely facilitated accurate learning because it was very unlikely for two variables to have the same value unless they are causally related. It would be informative to examine how well people understand Markov equivalence classes with binary or Gaussian variables, which will likely be harder. Another question raised by this article is to what extent heuristic strategies may be able to explain the psychological processes involved in this inference. In the studies by Steyvers et al. there are some simple rules that can distinguish the Markov equivalence classes fairly successfully. For example, upon observing a trial in which X=Y=Z, X←Y→Z is much more likely than X→Y←Z, but upon observing a trial in which X≠Y=Z, the likelihoods flip. But in other types of parameterizations such as noisy binary data or Gaussian data, this discrimination would not be so easy.

Even though Markov equivalence is a core feature of causal structure learning from observations, as far as I know this study by Steyvers et al. is the only study to test how well people learn causal structures purely from the correlations between the variables. There are a number of other studies that have investigated other observational cues to causality. For example, a number of studies have found that if X occurs followed by Y, people quickly and robustly use this temporal order or delay cue to infer that X causes Y (Lagnado & Sloman, 2006; Mccormack, Frosch, Patrick, & Lagnado, 2015). This inference occurs despite that fact that the temporal order may not necessarily represent the order in which these variables actually occurred, but instead it might reflect the order in which they become available for the subject to observe them. Another cue that people use to infer causal direction are beliefs about necessity and sufficiency. If a learner believes that all causes are sufficient to produce their effects, that whenever a cause is present, its effects will be present, then observing that X=1 and Y=0 implies that X is not a cause of Y, otherwise Y would be 1 (Mayrhofer & Waldmann, 2011). In sections 2.1.4 I will discuss one other way that people learn causal direction from observation. But the general point of all of these studies is that when these other cues to causality are pitted against pure correlations between the variables, people tend to use these other cues to causality (see Lagnado, Waldmann, Hagmayer, & Sloman, 2007 for a summary).

In sum, though there is some evidence that people do understand Markov equivalence class to some extent, this understanding appears limited. Furthermore, there are not many studies on how well people learn causal structures in a bottom-up fashion purely from correlational data when covariation is the only cue available. In contrast, what is clear is that when other strategies are available, people tend to use them instead of inferring causal structure purely from the dependencies and conditional independencies.

2.1.2 Learning Causal Structure from Interventions: Choosing Interventions

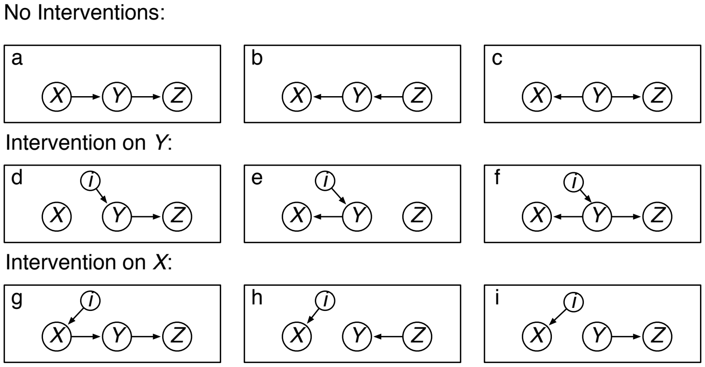
Another a core principle underlying causal structure learning is that interventions (manipulations) have the capability to discriminate causal structures that would not be distinguishable from observation; this is the same reason why experiments are more useful than observational studies. Going back to the example in Figure 1, if we were trying to figure out the causal relations between diabetes (D), statin use (S), and systolic blood pressure (BP), observational data would only be able to narrow the possibilities down to three structures: *D*→S→*BP, D*←S←*BP,* and *D*←*S*→*BP*. However, if we could do a randomized experiment such that half the patients take a statin and the other half do not, we could infer the causal structure. If *D*→*S*→*BP* is the true causal structure, then the patients who take a statin would have lower *BP* than those who do not, but there would not be any difference in *D* across the two groups. In contrast, if *D*←S←*BP* is the true causal structure, there would be a difference of *D* across the two groups but there would not be a difference in *BP* across the two groups. Finally, if *D*←*S*→*BP* is the true causal structure, there would be a difference in both *D* and *BP* across the two groups. The rest of this section will explain in more detail how interventions can be used to precisely identify a causal structure, and how humans use interventions.

The language of causal structure diagrams has a simple notation to represent interventions. When an intervention sets the state of a variable, all other variables that would otherwise be causes of the manipulated variable are no longer causes, so those links get removed. For example, when the patients in our example are randomly assigned to take a statin, even though normally having diabetes is a cause of taking a statin, now because of the random assignment diabetes is no longer a cause of taking a statin.

More generally, the reason why interventions can make causal structures that are in the same Markov equivalence class distinguishable is that the intervention changes the causal structure. For this reason, interventions are sometimes called "graph surgery" (Pearl, 2000). Figure 3a-c shows three causal structures that are not differentiable from observation. Figure 3d-f and g-i show the same three causal structures under either an intervention on *Y* or an intervention on *X*; the *i* nodes represent the intervention. (The intervention on *Y* is analogous to the previous example of the randomized experiment about taking a statin; it could be useful to compare these two examples for generality.)

Under the intervention on *Y* all three causal structures now have different dependence relations. In Graph D, *Z* and *Y* would still be correlated, but neither would be correlated with *X*. In Graph E, *X* and *Y* would be correlated, but neither would be correlated with *Z*. And in Graph F, all three variables would be correlated, but *X* and *Z* would become uncorrelated conditional on *Y*. In sum, interventions on *Y* change the causal structure such that the resulting structures no longer fall within the same Markov equivalence class, so they can be discriminated. In contrast, an intervention on *X* can discriminate Graph G from H, but cannot discriminate Graph H from I. This means that an intervention on *X* does not provide as much information for discriminating these three causal structures as does an intervention on *Y*.

Figure 3: Three Causal Structures with Different Types of Interventions.



Do people choose interventions that maximize "information gain," the ability to discriminate between multiple possible structures? Before getting to the evidence, it is useful to consider an alternative strategy for choosing interventions to learn about causal structure aside from maximizing information gain; selecting interventions that have the largest influence on other variables. For example, consider again the three structures in the No Interventions row in Figure 3. In Graph A, *X* influences two variables directly or indirectly, and in Graphs B and C *X* does not influence either other variable, for a total "centrality" rating of 2. Z has the same centrality rating - 2. *Y*, in contrast, influences *Z* in Graph A, *X* in Graph B, and *X* and *Y* in Graph C, for a total centrality rating of 4. In sum, looking across all three possible structures *Z* is more "central" or more of a "root cause." If a learner chooses to intervene to maximize the amount of changes in other variables they will tend to intervene on *Y* instead of *X* or *Z*.

Sometimes, as in the example in Figure 3 the Information Gain strategy and the Root Cause strategy produce the same interventions; *Y* is the most central variable and interventions on *Y* help discriminate the three structures the most. However, sometimes the two strategies lead to different interventions. For example, when trying to figure out whether [*X*→*Z*→*Y*] or [*X*→*Y*→*Z*] is the true structure, *X* has the highest centrality rating. However, intervening on *X* would not discriminate the structures well (low information gain) because for both structures it would tend to produce data in which *X*=*Y*=*Z*. Intervening on *Y* or *Z* would more effectively discriminate the structures. For example, intervening on *Y* would tend to produce data in which *X*=*Z*≠*Y* for [*X*→*Z*→*Y*] but would tend to produce data in which *X*≠*Y*=*Z* for [*X*→*Y*→*Z*], effectively discriminating the two structures. The Root Cause Strategy, intervening on *X*, can be viewed as a type of positive or confirmatory testing strategy in the sense that it confirms the hypothesis that *X* has some influence on *Y* and *Z*, but does not actually help discriminate between the remaining hypotheses.

Coenen et al. (2015) tested whether people use these two strategies and found that most people use a mixture of both, though some appear to use mainly one or the other. In another experiment Coenen tested whether people can shift towards primarily using the information gain strategy if they are first trained on scenarios for which the root cause positive testing strategy was very poor at discriminating the causal structures. Even without feedback, over time participants switched more towards using information gain. They also tended to use the root cause strategy more when answering faster. In sum, root cause positive testing is a heuristic that sometimes coincides with information gain, and it appears that people sometimes can overcome the heuristic when it is especially unhelpful.

Though Steyvers et al. (Steyvers et al., 2003) did not describe it in the same way as Coenen et al. (2015), they actually have evidence for a fairly similar phenomenon to the positive test strategy. They found that people tended to intervene on root causes more than would be expected by purely using information gain. In their study, participants first saw 10 observational learning trials, and then chose the causal structure that they thought was most plausible, for example [$X \rightarrow Y \rightarrow Z$]. Technically since the data was observational, they could not distinguish models within the same Markov equivalent class at this stage. Next they selected one intervention on either X, Y, or Z and would get 10 more trials of the same intervention repeatedly. Steyvers et al. found that their participants tended to select root cause variables to intervene upon. If they thought that the chain [$X \rightarrow Y \rightarrow Z$] structure was most plausible, they most frequently intervened on X, then Y, and then Z. This pattern fits better with the root cause heuristic than the information gain strategy, which suggests intervening on Y.

Because this finding cannot be explained by information gain alone, Steyvers et al. created two additional models; here I only discuss Rational Test Model 2. This model makes two additional assumptions. First, it assumes that participants had a distorted set of hypotheses about the possible causal structure. Normatively their hypothesis space for the possible causal structures should have been the Markov equivalent class; if they selected [$X \rightarrow Y \rightarrow Z$] as the most likely structure after the 10 observational trials, they should have also viewed [$X \leftarrow Y \leftarrow Z$] and [$X \leftarrow Y \rightarrow Z$] as equally plausible, and the goal should have been to try to discriminate among these three. However, this model assumes that instead of trying to discriminate between the three Markov equivalent structures, participants were trying to discriminate between [$X \rightarrow Y \rightarrow Z$], [$X \rightarrow Y$; Z], [X; $Y \rightarrow Z$], and [X; Y; Z]; the latter three are the subset of the chain structures that include the same causal links or fewer links. I call this assumption the "alternate hypothesis space assumption" in that the set of possible structures (hypothesis space) is being changed.

Under this new hypothesis space, when X is intervened upon, each of these four structures would produce different patterns of data, making it possible to determine which of these four structures is most likely; see Table 3. This means that X has high information gain. In contrast, when Y or Z is manipulated, some of the structures produce the same patterns of data, meaning that they provide less information gain.

Table 3: Most Common Pattern of Data Produced by An Intervention On A Particular Node for a Particular Structure

	Alternate Hypothesis Space Assumption			Alternate Hypothesis Space Assumption & Only Attending to Variables with Same State as Manipulated Variable		
	Intervention On:			Intervention On:		
Structure Hypothesis Space	X	Y	Z	X	Y	Z
$X \rightarrow Y \rightarrow Z$	$X=Y=Z$	$X \neq Y=Z$	$X=Y \neq Z$	$X=Y=Z$	$Y=Z$	Z
$X \rightarrow Y$; Z	$X=Y \neq Z$	$X \neq Y \neq Z$	$X=Y \neq Z$	$X=Y$	Y	Z
X ; $Y \rightarrow Z$	$X \neq Y=Z$	$X \neq Y=Z$	$X \neq Y \neq Z$	X	$Y=Z$	Z
X ; Y ; Z	$X \neq Y \neq Z$	$X \neq Y \neq Z$	$X \neq Y \neq Z$	X	Y	Z
How Informative This Intervention Is	High	Medium	Medium	High	Medium	Low

Steyvers et al. introduce another assumption as well; they assume that people only attend to variables that take on the same state as the intervened-upon variable and ignore any variables that have a different state from the manipulated variable. The combination of the two assumptions is detailed in the right column of Table 3. When *X* is intervened upon it would produce three different patterns of data for the four causal structures, which means that it has fairly high information gain; it can distinguish all but the bottom two structures in Table 3. The reason why an intervention on *X* can no longer distinguish between the bottom two structures is because of the assumption that the other two variables that do not equal *X*, *Y* and *Z,* are ignored.

When *Y* is intervened upon, it can narrow down the space of 4 structures down to 2, a medium amount of information gain. When *Z* is intervened upon all the structures produce the same pattern of data so an intervention on *Z* does not help at all to identify the true structure. In sum, the combination of these two hypotheses now makes it such that intervening on *X* is more informative than *Y*, which is more informative than inventing on *Z*. This pattern matches the frequency of participants' interventions, which were most frequently on *X*, then on *Y*, and lastly on *Z*.

There are two key points made by this analysis of the similarities between the findings of Coenen et al. and Steyvers et al. First, even though they approach the results from different perspectives and talk about the results in different ways, they both found that people tended to intervene on root causes. Second, even though Steyvers' model has rational elements to it, the resulting model is not very close to the ideal model, for which *Y* is the most informative intervention. Finally, by comparing different models with different assumptions it can be seen how the two assumptions made by Steyvers et al. effectively amount to the positive test strategy put forth by Coenen at al. Restated, the same behavioral pattern of intervening primarily on root causes could be explained in more than one way.

Bramley et al., (2015) also studied a number of important factors related to learning from interventions. Overall, they found that humans were highly effective causal learners, and able to select and make use of interventions for narrowing down the number of possible structures. One particular factor he introduced was the possibility of intervening on two variables simultaneously rather than just one. Double interventions are particularly helpful to distinguish between [*X*→*Y*→*Z* and *X*→*Z*] vs. [*X*→*Y*→*Z*]. With a single intervention, these two structures are likely to produce very similar outcomes. For example, an intervention on *X* is likely to produce data in which $X=Y=Z$, an intervention on *Y* is likely to produce data in which $X\neq Y=Z$, and an intervention on *Z* is likely to produce data in which $X=Y\neq Z$.

However, consider a double intervention setting $X=1$ and $Y=0$. Under the simple chain [*X*→*Y*→*Z*], *Z* is most likely to be 0, but in the more complex structure in which *Z* is influenced by both *X* and *Y*, *Z* has a higher chance of being 1 because $X=1$. Bramley et al. found that distinguishing between these two types of structures was the hardest discrimination in this study and produced the most errors. 61 of the subjects rarely used double interventions, whereas 49 were more likely to use them, suggesting that people may not use double interventions frequently enough for causal learning.

In sum, there are many open questions about how people choose interventions. There is some evidence that people do use information gain when selecting interventions, but there are also a variety of heuristics (only use single interventions, intervene on root causes) and or simplifying assumptions (focus on a limited and distorted hypothesis space, only attend to variables with the same value as the manipulated variable). Clearly there is more work to be done to have a fuller and more robust understanding of how humans choose interventions to learn about causal structures.

2.1.3 Learning Causal Structures from Interventions: Interpreting Interventions and Updating Beliefs about the Structure

The prior section discussed how people choose interventions to learn a causal structure. This section examines how people interpret the outcome after making an intervention. Four patterns have been proposed to explain how people interpret the outcomes of interventions. The first is that if a variable *X* is manipulated and another variable *Z* assumes the same state of *X*, people tend to infer a direct link from *X* to *Z*. Though this heuristic makes sense when learning the relations between two variables, it can lead to incorrect inferences in cases involving three or more variables linked together in a chain structure such as *X*→*Y*→*Z* because it can lead people to infer additional links not in the structure. If a learner intervenes on *X* such that it is 1, and subsequently *Y* and *Z* are both 1, this heuristic implies that *X*→*Y* and *X*→*Z*. Indeed, people often infer that there is an *X*→*Z* link above and beyond *X*→*Y*→*Z*, even in cases when there is not a direct link from *X* to *Z* (Bramley et al., 2015; Fernbach & Sloman, 2009; Lagnado & Sloman, 2004; Rottman & Keil, 2012). (In reality, the correct way to determine whether there is a *X*→*Z* above and beyond *X*→*Y*→*Z* is to see whether the probability of *Z* is correlated with the state of *X* within trials in which *Y* is 1 or within trials in which *Y* is 0, or to use double interventions as explained previously.)

This heuristic is problematic for two reasons. At a theoretical level, it suggests that people fail to pay attention to the fact that *X* and *Z* are independent conditional on *Y*. As already discussed, attending to statistical independencies is critical for understanding Markov equivalence classes, and this finding suggests that people do not fully understand the relations between statistical independence and causal Markov equivalence class. At a more applied level, adding this additional link *X*→*Z* could lead to incorrect inferences (see section 3.2). In particular, when inferring the likelihood that *Z* will be present given that *Y* is present, people tend to think that *X* has an influence on *Z* above and beyond *Y*. In reference to a subset of Figure 1, even though the true causal structure is *Ethnicity → Smoking → Cardiovascular Disease*, this heuristic could lead doctors to incorrectly predict that people of certain ethnicities are more likely to develop cardiovascular disease even after knowing their smoking status, even though ethnicity has no influence on cardiovascular disease above and beyond smoking (according to Thornley, 2013). Such a misperception could lead people of those ethnicities to feel unnecessarily worried that their ethnicity will cause them to have cardiovascular disease.

The second pattern of reasoning was already discussed in the previous section. Steyvers et al., (2003) proposed that when a person intervenes on a variable, that they only attend to other variables that assume the same state as the manipulated variable. The previous section explained how this tendency would bias reasoners to intervene on root causes (Table 3). But this tendency would also decrease the effectiveness of learning from interventions. If one intervenes on *Z* and the resulting observation is *X*=*Y*≠*Z*, the fact that *X* and *Y* have the same state should increase the likelihood that there is some causal relation between *X* and *Y*; however, this heuristic implies that people would not learn anything about *X* or *Y* because people only attend to variables with the same state as the intervened-upon variable (*Z*). In sum, this simplification means that people do not extract as much information from interventions as they could.

The third and fourth habits of updating causal beliefs after interventions come from the study by Bramley et al's study (2015). In this study participants made a series of interventions on *X*, *Y* or *Z*, and after each intervention they drew the causal structure that they believed to be the most plausible structure given the evidence up to that point. They discovered two interrelated habits. First, participants updated their drawings of the causal structure slowly. This can be explained as a conservative tendency; people need considerable evidence before adding or deleting a causal relation to their set of beliefs. The second pattern is that when drawing the causal structures participants were influenced by the most recent intervention and appeared to forget many of the outcomes of prior interventions. The combination of these two habits, conservatism and forgetting can be explained with an analogy to balancing a checkbook. After each transaction one updates the current balance by adding the most recent transaction to the prior balance, but one does not re-calculate the balance from all past experiences after each transaction. Keeping the running balance is a way to simplify the calculation. Likewise, storing a representation of the causal structure as a summary of the past experience allows the learner to get by without remembering all the past experiences; the learner just has to update the prior causal structure representation. In a related vein, Fernbach and Sloman (2009) found that people have a recency bias – they are most influenced by the most recent data, which is similar to forgetfulness. Understanding the interplay between all of these habits will provide insights into how people learn causal structures from interventions in ways that are cognitively tractable.

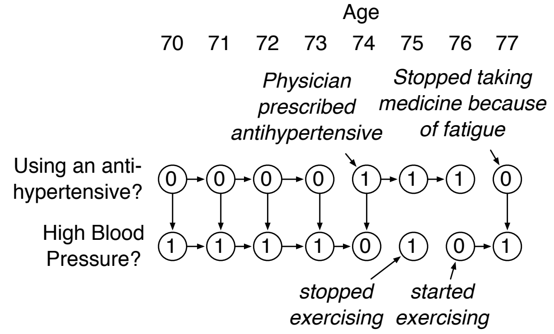
2.1.4 Learning Temporal Causal Structures

So far this chapter has focused on how people learn about atemporal causal networks in which each observation is assumed to be temporally independent. In the example at the beginning of the chapter about cardiovascular disease, each observation captured the age, sex, smoking status, diabetes status, and other variables of an individual patient. The causal link between smoking and cardiovascular disease, for example, implies that across patients, those who smoke are more likely to have cardiovascular disease.

However, often it is important to understand how variables change over time. For example, a physician treating patients with cardiovascular disease is probably less interested in population-level effects, and instead is more interested in understanding how a change in smoking would influence an individual patient's risk of developing cardiovascular disease.

Temporal versions of CBNs can be used to represent learning and reasoning about changes over time (Ghahramani, 1998; Murphy, 2002 also see Rehder, this volume b). Temporal CBNs are very similar to standard CBNs, except each variable is represented by a series of nodes for each time point *t*. The causal structure is often assumed to be the same across time, in which case the causal structure is repeated at each time point. Additionally, often variables are assumed to be influenced by their past state; positive autocorrelation means that if the variable was high at time *t*, it is likely to be high at *t*+1.

Figure 4: A Temporal CBN



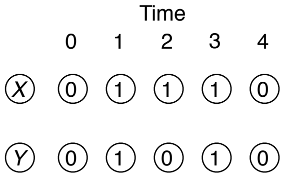
For example, Figure 4 shows a causal network representing the influence of using an antihypertensive on blood pressure. 1 represents using an antihypertensive or high blood pressure, whereas 0 represents not using an antihypertensive or having normal blood pressure. Instead of just having one node that represents using an antihypertensive and another for blood pressure, now the structure is repeated at each time point. Additionally, the autocorrelation can be seen with the horizontal arrows. All things being equal, if a patient's blood pressure is high, it will tend to stay high for periods of time. Likewise, if a patient starts using an antihypertensive, they might continue to use it for a while.

Like all CBNs, temporal CBNs follow the same rules and conventions. Here instead of using *i* nodes to represent interventions, I used text to explain the intervention (e.g., a physician prescribed an antihypertensive). The interventions are the reason that some of the vertical and horizontal arrows are removed in Figure 4 because an intervention modifies the causal structure. The Markov condition still holds in exactly the same way as in temporal CBNs. For example, a patient's blood pressure (BP) at age 73 is influenced by his BP at 72, but his BP at 71 does not have an influence on his BP at 73 above and beyond his BP at age 72.

Causal learning from interventions works in essentially the same way in temporal and atemporal causal systems. In Figure 4 it is easy to learn that using an antihypertensive influences blood pressure, not the reverse. When the drug is started, the patient's BP decreases, and when the drug is stopped, the patient's BP increases. But when another intervention (e.g., exercising) changes the patient's blood pressure, it does not have an effect on whether the patient uses a statin.

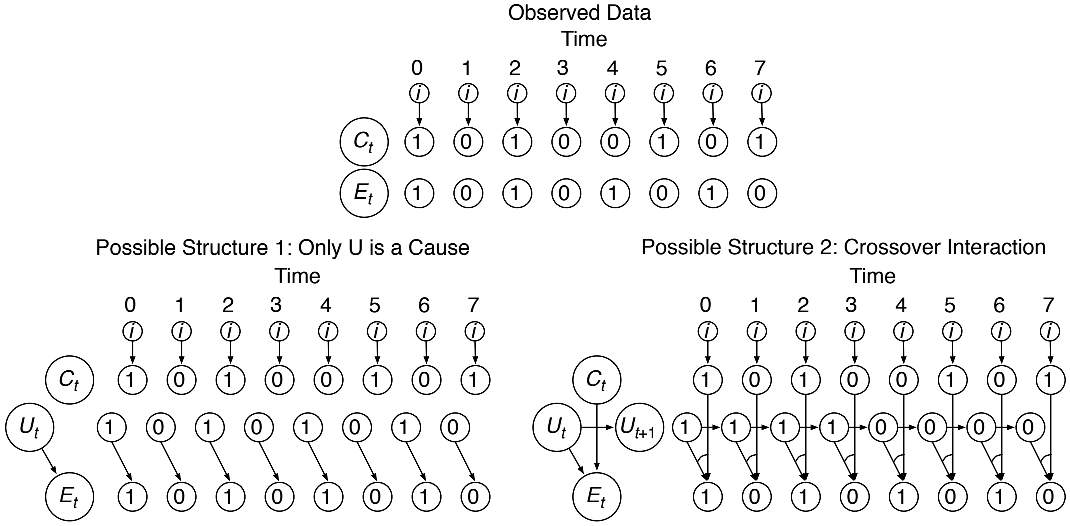
One interesting aspect about temporal causal systems, is that is possible to infer the direction of a causal relationship from observations, which is not possible with atemporal systems. Consider the data in Figure 5; the direction of the causal relation is not shown in the figure. There is an asymmetry in the data; sometimes *X* and *Y* change together, and sometimes *Y* changes without *X* changing, but *X* never changes without *Y* changing. Colleagues and I have found that both adults and children notice this asymmetry and use it to infer that *X* causes *Y* (Rottman & Keil, 2012; Rottman, Kominsky, & Keil, 2014; Soo & Rottman, 2014). The logic is that *Y* sometimes changes on its own, implying that whatever caused *Y* to change did not carry over to *X*; *Y* does not influence *X*. Furthermore, sometimes *X* and *Y* change together. Since we already believe that *Y* does not influence *X*, one way to explain the simultaneous change in X and Y is that a change in X caused the change in Y.[2] This is one way in which human causal learning seems more akin to learning temporal CBN rather than an atemporal CBN; the temporal aspect of this data is critical for inferring the causal direction.

Figure 5: Example of Learning Causal Direction from Temporal Data



A number of other phenomena fit well into the temporal CBN framework. Consider the data in Figure 6 Observed Data. In this situation, subjects know that C is a potential cause of E, not the reverse and the goal is to judge the extent that C influences E. Overall there is actually zero correlation between C and E. The faithfulness assumption states that the only independencies in the data arise through the Markov assumption. If C and E are unconditionally independent, it means that C cannot be a direct cause of E. Instead, another possibility (Possible Structure 1 in Figure 6) is that some unobserved third variable U is entirely responsible for E.

Figure 6: Learning about an Interaction with an Unobserved Factor



However, when faced with data like in Figure 6, people do not conclude that C is unrelated to E; instead they notice that there are periods of time in which C has a positive influence on E (Times 0-3), and other periods of time in which C has a negative influence on E (Times 4-7). They subsequently tend to infer that C does actually have a strong influence on E, but that there is some unobserved factor that is fairly stable over time, and C and the unobserved factor (U) interact to produce E (Rottman & Ahn, 2011). This explanation is

[2] Technically, the reason why it is possible to learn the direction of the causal relation is the autocorrelation, the belief that $Y_t \rightarrow Y_{t+1}$ and that $X_t \rightarrow X_{t+1}$. Thus, the learner is really discriminating between $[Y_t \rightarrow Y_{t+1} \leftarrow X_{t+1} \leftarrow X_t]$ and $[Y_t \rightarrow Y_{t+1} \rightarrow X_{t+1} \leftarrow X_t]$, which are in different Markov equivalence classes. I thank David Danks for pointing this out.

represented in Figure 6 Possible Structure 2. In this structure, both C and U influence E, and there is an ark between the two links, which represents an interaction; in this case the interaction is a perfect cross-over such that E is 1 if both C and U are 1 or both are 0. The reason people appear to make this inference about the crossover interaction with an unobserved cause rather than inferring that C is unrelated to E is because the data are grouped into distinct periods such that there are periods during which there is sometimes a positive relation and other times a negative relation between C and E. This allows the reasoner to infer that some unobserved factor U must account for the switch. If the same 8 trials were randomized then people tend to infer that only U is a cause of E, not C. This inference again suggests that people tend to represent causal systems as temporally extended (that variables such as U tend to be autocorrelated) rather than atemporal (see Rottman & Ahn, 2009, for another example).

Elsewhere, colleagues and I have argued (Rottman et al., 2014) that many of the causal learning phenomena that have been used as evidence that people learn about causal relations in ways akin to CBNs are even better explained by temporal CBNs. For example, one study found that children can learn about bidirectional causal relations in which two variables both cause each other (Schulz, Gopnik, & Glymour, 2007). Bidirectional causal structures can only be represented through temporal, not atemporal causal networks (Griffiths & Tenenbaum, 2009; Rottman et al., 2014).

In conclusion, there is growing evidence that, at least in certain situations, people appear to be learning something similar to a temporal causal network, and the temporal aspect of reasoning allows them to infer quite sophisticated causal relations that would otherwise be impossible to learn.

2.2 Learning about the Integration Function

Another aspect of a CBN that must be learned in addition to the structure is the integration function; the way that multiple causes combine to influence an effect (also see Griffiths, this volume, Rehder, this volume a and b). For example, in regression, the predictors are typically assumed to combine linearly. The CBN framework allows for the possibility that causes can potentially combine in any conceivable way, and humans are extremely flexible as well. For example, Waldmann (2007) demonstrated that people naturally reason about causes that are additive (e.g., the effect of taking two medicines is the sum of the two individual effects) and averages (e.g., the taste of two chemicals mixed together is the average of the two). Furthermore, people use background knowledge (e.g., about medicines and taste) to decide which type of integration function is more plausible in a given situation.

Most research on causal learning has focused on binary variables. The most prominent integration function for binary variables, called Noisy-OR, describes situations in which there are multiple generative causes (Cheng, 1997; Pearl, 1988). It stipulates that the probability of the effect being absent is equal to the probability that all the causes happen to simultaneously fail to produce the effect. If there are two causes, each of which produce the effect 50% of the time on their own (a causal strength of .5), then both would fail simultaneously 25% of the time; the effect should occur 75% of the time. If there are three causes, each of which produces the effect 50% of the time, then all three would simultaneously fail 5^3 = 12.5% of the time; the effect would be present 87.5% of the time. An analogous integration function called Noisy-And-Not can be used to describe inhibitory causes that combine in a similar fashion. It is not difficult to imagine other sorts of integration functions, and the following studies have examined how people learn about the integration function from data.

Beckers et al. (2005; see also the chapter in this volume by Boddez, De Houwer, & Beckers) studied how beliefs about the integration function influence learning. In one study participants first learned about two causes, *G* and *H*, both of which produce an outcome of 1 one their own. In the "additive" condition they saw that *G* and *H* together produce an outcome of 2, which is consistent with an integration function in which two causes add together. In another condition they saw that *G* and *H* together produce an outcome of 1. This is inconsistent with the notion that the two causes add together; instead it suggests some sort of "sub-additive" integration function in which the effect can never be higher than 1. Subsequently participants in both conditions experienced a blocking paradigm in which they learn that *A* by itself produces an outcome of 1, and *A* plus *X* produces an outcome of 1. In the subadditive condition participants still thought that *X* might be a cause because they believed that the effect could never go higher than 1. In contrast, in the additive condition they concluded that *X* was not a cause; if it was, then presumably the effect would have been 2.

Lucas and Griffith (Lucas & Griffiths, 2010) investigated a similar phenomenon, that initial training about how causes combine influences whether subjects interpret that a variable is a cause or not. They first presented people with data that suggested that the causes worked conjunctively (multiple causes were needed to be present for the effect to occur), or through the noisy-OR function (a single cause was sometimes sufficient to produce the effect). Afterwards, participants saw a cause *D* never produce the effect, and saw that two causes in combination, *D* and *F,* produced the effect. Participants in the conjunctive condition tended to conclude that both *D* and *F* were causes, whereas participants in the noisy-OR condition tended to infer that only *F* was a cause.

In sum, these results show that people quickly and flexibly learn about how causes combine to produce an effect and the integration rule that they learn dramatically influences subsequent reasoning about the causal system.

2.3 Learning Causal Strength

So far this chapter has focused on how people learn causal structure, and to a lesser extent integration functions. One other important component of causal relations is causal strength, our internal measurement of how important a cause is. For example, if a medicine works very well to reduce a symptom, it has high causal strength, but if it does not reduce the symptom at all it has zero causal strength.

Prior to the CBN framework, theories of causal strength learning were based on simple measures of the contingency between the cause and effect. For example, the ΔP model computes the strength of the influence of a cause (C) on an effect (E) by the extent of the difference in the probability of the effect when the cause is present vs. absent; $P(e{=}1|c{=}1)$-$P(e{=}1|c{=}0)$ (Cheng & Novick, 1992; Jenkins & Ward, 1965). This same contrast is calculated at asymptote by one of the most influential models conditioning as a way to capture how strongly a cue and outcome become associated by an animal (Danks, 2003; Rescorla & Wagner, 1972). This same model has also been proposed as a model of causal learning, the idea being that the stronger that a cue is associated with an outcome, the stronger that humans would infer that the cue causes the outcome (David R Shanks & Dickinson, 1987).

With the introduction of the CBN framework a number of theories of causal learning were proposed that incorporate different sorts of top-down causal beliefs into the learning process. A number of other chapters discuss causal strength learning including those by Griffiths, Cheng and Lu, and Perales, Catena, Maldonado and Cándido. Thus, I briefly discuss the connections between the CBN framework and theories of causal strength learning, while leaving the details to those other chapters.

2.3.1 Elemental Causal Induction: Learning Causal Strength Between Two Variables

One of the most important developments of models of causal strength learning is the Power-PC model (Cheng, 1997). This model builds off the ΔP model by incorporating causal beliefs and assumptions. This model assumes that one generative cause combines through the Noisy-OR integration function with another unobserved cause. For example, imagine that the effect E occurs 25% of the time without the observed cause C; $P(e{=}1|c{=}0){=}.25$. We can attribute this 25% to some background cause that has a strength of .25. Further, imagine that the observed cause has a strength of 2/3. When the observed cause is present, the effect should occur 75% of the time if C and the background cause combine through a noisy-OR function; $P(e{=}1|c{=}1){=}.75$. (The effect would fail with a probability of 1/3×3/4 = ¼).

Cheng used this sort of logic, in reverse, to deduce that if an observed cause combines with a background cause through a noisy-OR integration function, the correct way to calculate causal strength involves dividing ΔP by $P(e{=}0|c{=}0)$. Consider now the probabilities just presented, without knowing the causal strength: $P(e{=}1|c{=}1){=}.75$ and $P(e{=}1|c{=}0){=}.25$. According to ΔP, the causal strength is .5; the causes raises the probability of the effect by .5. According to Power-PC, the causal strength of C is (.75-.25)/(1-.25)=.67; the cause increases the effect by 2/3rds (from .25 to .75). In sum, by specifying a set of prior beliefs about the causal relation, Cheng specified how causal strength should be induced given those beliefs.

Another influential development to causal strength learning is the Causal Support model. Griffiths and Tenenbaum (2005) proposed that when people estimate causal strength, what they are actually doing is not judging the magnitude of the influence of the cause on the effect, similar to effect sizes in inferential statistics, but rather judging the extent to which there is evidence that there is any causal relation or not, similar to the function of a p-value in hypothesis testing. At a theoretical level, this model is calculated by determining the relative likelihood that the true causal structure is [$C \rightarrow E \leftarrow U$], that both C and an unobserved factor U influence E vs. that the true causal structure is [C; $E \leftarrow U$], that C does not influence E and E is determined by an unobserved factor U. Thus, causal support treats causal strength learning as discriminating between two possible causal structures, one in which C actually is a cause of E, and one in which C is not a cause of E.

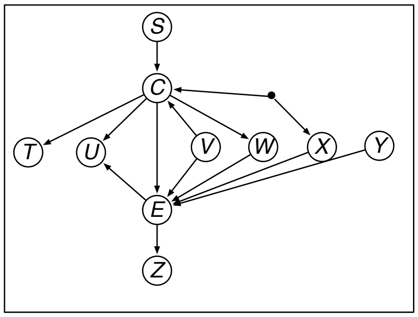
Causal Support has a number of behavioral implications, but the most obvious one and easiest to think about is sample size. Whereas ΔP and Power-PC are unaffected by sample size, Causal Support is influenced by sample size. Going back to the analogy of Causal Support as a p-value whereas ΔP and Power-PC are effect size measures, if there is a large enough sample size it is possible to have a very low p-value (confident that there is a causal relation) even if the effect size is small.

In sum, Power PC and Causal Support were both motivated by understanding causality through a CBN perspective, involving top-down beliefs about how an observed cause combines with other unobserved factors.

2.3.2 Inferring Causal Strength: Controlling for Other Causes

The previous section focused on how people infer causal strength given observations of just a single cause and effect, elemental causal induction. However, often there are more than two variables. When inferring the strength of one cause on an effect it is important to control for certain types of third variables (and not others), depending on the causal structure. Consider Figure 1. When studying the strength of the effect of a new drug on cardiovascular disease, it is important to control for age and smoking habits, either statistically or through the design of the study. One should not control for statin use because it is not a direct cause of cardiovascular disease.

Figure 7: Possible Third Variables when Learning the Causal Relation from C to E



More generally, consider trying to learn if there is a causal link from a potential cause *C* to a potential effect *E*, and if so, how strong the relation is. Figure 7 presents 8 different third variables (*S*-*Z*); the question is which of these variables should be controlled for. For readers familiar with multiple regression, you can think of *C* as one predictor in the regression that you are primarily interested in, and *E* is the outcome variable. The question about controlling for alternative variables is which of these variables should be included as predictors or covariates in the analysis? The following bullets systematically explain each of the third factors and whether it should be controlled for when inferring the strength of *C* on *E*:

- *V* and *X* are confounds and must be controlled for when inferring the relation of *C* on *E*. If they are not controlled for there would be a spurious correlation between *C* and *E* even if there is no causal relation between *C* and *E*. (*X* represents the case when some unobserved factor causes both *C* and *X*.)
- *W* represents an alternative mechanism from *C* to *E*. In order to test whether there is a direct influence of *C* on *E* above and beyond *W* it must be controlled for.
- *Y* is a noise variable. Accounting for it increases our power to detect a relation between *C* and *E*.
- *U* and *Z* should not be controlled for. The logic is a bit opaque (Eells, 1991, p. 203), but consider the simple case that *E* deterministically causes *Z* such that they are perfectly correlated. Controlling for *Z* explains all the variance in *E*, and there will be no left over variance for *C* to explain. Controlling for *Z* and *U* can distort the apparent relation between *C* and *E*.
- *S* and *T* never need to be controlled for. With large sample sizes it does not matter if *S* and *T* are controlled for or not when inferring the influence of *C* on *E*. The reason is that even though *S* and *T* are correlated with *C*, since *S* and *T* are screened off from *E*

(S and T are independent of E after controlling for C), they will not have any predictive power in a regression above and beyond C. However, with small sample sizes, most likely S and T will not be perfectly uncorrelated with E controlling for C, in which case they can change the estimated influence of C on E. Thus, they should not be controlled for.

In sum, the overall rule is that when inferring the strength of a relation of C on E, third variables that are believed to be potential direct causes of E should be controlled for; other variables should not be controlled (Cartwright, 1989; Eells, 1991; Pearl, 1996). This rule nicely dovetails with how causal structures are defined; each variable is modeled using a conditional probability distribution incorporating all of its direct causes.

Remarkably, a variety of research suggests that people have the ability to appropriately control for third variables when inferring causal strength. In fact, research on this topic was the first research on whether people intuitively use beliefs about causal structure when reasoning about causality (Waldmann & Holyoak, 1992; Waldmann, 1996, 2000). Michael Waldmann and colleagues called this theory the Causal Model theory; the idea was that when inferring causal strength, people use background knowledge about the causal structure ("model") to determine which variables to control for. In the first study on this topic, a scenario with three variables X, Y, and Z was set up. Based on the cover story the three variables were either causally related in a common effect structure [$X \rightarrow Y \leftarrow Z$] or in a common cause structure [$X \leftarrow Y \rightarrow Z$]. In the common effect condition [$X \rightarrow Y \leftarrow Z$], the goal for participants was to decide the extent to which X and Z were causes of Y; normatively people should control for alternative causes (e.g., control for X when determining whether Z is a cause of Y). In the common cause condition [$X \leftarrow Y \rightarrow Z$], the goal for participants was to decide the extent to which X and Z are effects of Y; normatively these two decisions should be made separately (e.g., one should ignore X when determining the influence of Y on Z).

After the cover story manipulating the believed causal structure, participants first experienced a set of data in which X and Y were perfectly correlated; Z was not displayed. This training made it seem that there is a strong causal relation between X and Y. Then they experienced a set of data in which X, Y, and Z were all perfectly correlated; now Z is a redundant predictor of Y because X is entirely sufficient to predict Y. In sum, participants experienced the exact same data, and the only difference between the two conditions was their belief about the causal structure.

In the common effect condition [$X \rightarrow Y \leftarrow Z$], participants controlled for X when interpreting whether Z was a cause of Y, and consequently concluded that Z is not a cause of Y because X is entirely sufficient to predict whether Y was present or absent. In contrast, in the common cause condition [$X \leftarrow Y \rightarrow Z$], participants did not control for X, and concluded that Y was a cause of both X and Z.

Subsequently, a number of other studies have also shown that people control for alternative causes (V-Y in Figure 7) of the main effect and not alternative effects of the main cause (T in Figure 7) (Goodie, Williams, & Crooks, 2003; Spellman, Price, & Logan, 2001; Waldmann, 2000). There is even work suggesting that people do not control for variables like S and Z (Waldmann & Hagmayer, 2001); however, there has not been research on whether people control for variables like U.

In sum, when learning about a causal relation between C and E, people have some core intuitions to control for variables that they believe to be alternative causes of E, and not other roles, which is critical for correct causal learning (Glymour, 2001). This research is some of the most dramatic showing how top-down beliefs about causal structure influence learning, and consequently is some of the strongest evidence that human causal reasoning involves structured directional representations beyond just associations between variables (Waldmann, 1996).

3 Reasoning with the Causal Structure

So far this chapter has focused on how people learn about a causal network; the structure of the network, the parameters or causal strengths, and the functional form. The remainder of the chapter is how people use this knowledge (see also Oaksford and Chater, this volume). Going back to Figure 1, one might desire to explain whether a person's cardiovascular disease was caused by his age, or his smoking. One might desire to predict whether his cardiovascular disease will get worse as he ages. And one might desire to know which intervention, stopping smoking or starting to take a statin would have the largest influence on his cardiovascular disease in order to choose the action with the greatest rewards.

Though this second half of the chapter focuses on reasoning about the causal network rather than learning, it is impossible to completely divorce learning and reasoning. In the real world we learn about causal relations both from first-hand experience with data (e.g., did starting the statin lower my blood pressure) and also from communicated knowledge (e.g., from family members, teachers, doctors, newspaper articles). Research in psychology has used both personal experience and communicated knowledge, often in combination, to teach subjects about the causal structure before they reason about the structure. Typically words and pictures are used to convey the causal structure to participants, although the structural information is sometimes conveyed through or supplemented with experienced data. If the participants learn anything about the parameters (causal strengths) of the causal structure, it is usually conveyed through data-driven experience, though sometimes the parameters are conveyed textually. The integration function is often not mentioned at all, though sometimes it is mentioned.

One of the challenges with studying how well people reason about causal structures is that apparent flaws in reasoning can either be explained as reasoning biases, or as poor, biased, or insufficient learning about the causal structure. It is not clear how to cleanly differentiate the two because checking that the causal structure is learned appropriately involves questions that are typically viewed as reasoning about the causal structure. This sets up a difficult situation because any observed reasoning bias can potentially be explained away by claiming that the researcher failed to sufficiently convey the causal structure to the participants. Here I do not try to solve this problem, but instead just present the empirical findings of how closely reasoning appears to fit with the causal structures presented to subjects. These conclusions are based on a much more thorough analysis of the literature than can be presented here (Rottman & Hastie, 2014), though this chapter includes some newly published evidence.

3.1 Reasoning based on Observations vs. Interventions

In Section 2, I explained how the CBN framework treats observations and interventions very differently for learning a causal structure. Interventions change the causal structure by removing links from variables that were previously causes of the manipulated variable. For example, given the structure $X \rightarrow Y \rightarrow Z$, if Y is intervened upon, Y gets severed from X resulting in [X; $Y \rightarrow Z$]. Under an intervention on Y, X would be statistically independent or uncorrelated from *Y*, even though *Z* would still be dependent upon *Y*.

Practically, given the structure *X*→*Y*→*Z*, if a reasoner can *observe* the state of *Y*, they can make a prediction about both *X* and *Z*. In the types of situations typically studied in the lab with binary variables and positive causal relations, if *Y* is observed as 1, then *X* and *Z* are both likely to be 1 as well. However, if a reasoner *intervenes* on *Y* and sets its value to 1, then *Z* is likely to be 1, but this intervention would have no influence on *X*, so the best estimate of *X* is simply its base rate. In sum, interventions only influence variables down-stream from the manipulated variable, not up-stream (but see Hiddleston, 2005 for an alternative approach, and also see the chapter by Over on whether "if... then" conditionals are interpreted as interventions).

A number of researchers have found that people discriminate between observations and interventions when making inferences based on a causal structure. Sloman and Lagnado (2005) set up simple verbal descriptions in which one event (*X*) causes the other (*Y*), and found that when *Y* was *observed* to have a particular value, *X* would be inferred to have the same value, but when *Y* was *intervened upon* to have a particular value, *X* was inferred to have its normal default value. In sum, when it was made very clear whether there was an observation vs. an intervention, subjects' judgments largely followed the prescriptions of the CBN framework. In contrast, when more ambiguous language is used such that the value of a variable could be known either through an observation or an intervention, then the responses looked more muddy (see also Rips, 2010).

Another set of studies took this basic finding a step further by demonstrating that this difference between interventions vs. observations also holds in contexts in which participants are told the causal structure and then learn the parameters (e.g., the base rates and the causal strengths) from experience. Consider a set of studies that investigated reasoning on a diamond structure [*X*←*W*→*Y* and *X*→*Z*←*Y*] (Meder, Hagmayer, & Waldmann, 2008, 2009; Waldmann & Hagmayer, 2005). These studies are unique for involving more than three variables, and also for having two causal routes *W*→*X*→*Z* and *W*→*Y*→*Z*. Despite the complexities involved in these studies, the participants showed remarkable subtlety in reasoning about the causal structures, and distinguishing between interventions and observations differently.

Consider observing a low value of *X*, and trying to infer the value of *Z*. In the diamond structure there are two routes from *X* to *Z*: *X*←*W*→*Y*→*Z* and *X*→*Z*. Due to these two routes *X* and *Z* should be strongly correlated, and thus *Z* should be quite low when *X* is observed to be low. In contrast, if *X* is intervened upon and set to a low value, the route *X*←*W*→*Y*→*Z* is destroyed – the link from *W* to *X* is cut. The *X*→*Z* route is still open, so the predicted value of *Z* is still low, but it should not be as low as when *X* is observed. In fact, this is the exact pattern of reasoning that was observed; the inference of *Z* after an observation of *X* was lower than after an intervention on *X*. This finding further suggests that people reason about observations both down-stream and up-stream, but they reason about interventions only down-stream. This research also shows how people can reason about observations and interventions on more complex structures.

So far this section has focused on "perfect" interventions in which the intervention completely determines the state of the manipulated variables, and completely severs all other influences. However, often interventions are not perfect. For example, after prescribing a patient an antihypertensive to treat high blood pressure, the patient may not actually take it, or may not take it exactly as prescribed (e.g., as frequently as they should, at the right dose). Furthermore, even if the patient does take the medicine as prescribed, the medicine does not guarantee that all patients will have a 120/80 blood pressure. Patients who initially had very high blood pressures will probably still tend to have higher blood pressures than those who initially had moderately high blood pressures. Or, perhaps the medicine only succeeds to bring the blood pressure into a normal range for a certain percentage of patients, but not for others. In these ways, taking an antihypertensive is an "imperfect" intervention on blood pressure; a patient's blood pressure is not completely determined by the intervention. In such cases of imperfect interventions, reasoning up-stream is warranted to some extent, similar to observations. Unfortunately, there has been fairly little work examining how people reason about imperfect interventions (Meder, Gerstenberg, Hagmayer, & Waldmann, 2010; Meder & Hagmayer, 2009).

In sum, the existing research has found that people do distinguish between interventions and observations when reasoning about causal systems, in particular that interventions only influence variables down-stream from the intervened-upon variable. An important direction for future research is to examine how people reason about imperfect interventions. This seems especially important given that many of the actions or "interventions" humans perform are not perfect interventions.

3.2 Do People Adhere to the Markov Condition when Reasoning about Causal Structures?

Recall that the Markov condition states that once all the direct causes of a variable Z are controlled for or held constant, Z is statistically independent of every variable in the causal network that is not a direct or indirect effect of Z. For example, in the structure $X \rightarrow Y \rightarrow Z$, Z is conditionally independent of X once Y (the only direct cause of Z) is held constant. People have often been found to violate the Markov assumption; their inferences about the state of Z are influenced by the state of X even when they already know the state of Y (Mayrhofer & Waldmann, 2015; Park & Sloman, 2013; Rehder & Burnett, 2005; Rehder, 2014; Rehder, this volume b; Walsh & Sloman, 2008). Specifically, people tend to infer that $P(z=1|y=1,x=1) > P(z=1|y=1,x=0)$ even though they should be equivalent. Likewise, they use Z when inferring X even after knowing the state of Y. Going back to section 2.1.3, such a mistake could lead a doctor to incorrectly believe that ethnicity has an influence on cardiovascular disease above and beyond smoking even when the true causal structure is *Ethnicity → Smoking → Cardiovascular Disease.*

There are a variety of possible explanations for why inferences violate the Markov condition, and most of the explanations have attempted to find rationalizations for the violations, reasons that such judgments would make sense according to the CBN framework assuming some modification to the structure due to prior knowledge. For example, if subjects believe that there is some other causal link between X and Z (e.g., $\mathrm{X} \rightarrow Z$, $X \leftarrow Z$, or $X \leftarrow W \rightarrow Z$) in addition to the causal structure told to them by the experimenter ($X \leftarrow Y \rightarrow Z$), such additional information could justify their inferences. Three specific proposals are that people infer an unobserved factor that inhibits both X and Z, an unobserved factor that influences X, Y, and Z, or an intermediary mechanism M such that Z causes A, which in turn causes X and Y. Different articles in the list above have supported different accounts. For example, Burnett and Rehder (2005) argued for the account in which an unobserved factor influences X, Y, and Z. Park and Sloman (2013) found that people only make the Markov violation when the middle variable is present, not absent;

$P(X{=}1|Y{=}1,Z{=}1) > P(X{=}1|Y{=}1,Z{=}0)$ but that $P(X{=}1|Y{=}0,Z{=}1) = P(X{=}1|Y{=}0,Z{=}0)$. This finding is most consistent with the account that people infer an unobserved factor that inhibits X and Z. They also found that the size of the Markov violation was larger when participants believed that the two effects (X and Z) are both caused through the same mechanism (e.g., Y causes mechanism A, which in turn causes X and Z), than through separate mechanisms (e.g., $\mathrm{X} \leftarrow A \leftarrow Y \rightarrow B \rightarrow Z$, where A and B are the two mechanisms that explain how X and Z are each caused by Y). Mayrhofer and Waldmann (2015) have also found evidence that people infer an unobserved inhibitory factor that influences multiple effects of the same cause. And they further found that the size of the Markov violation was influenced by whether the causes and effects were described as agents vs. patients (e.g., cause "sending" information to effect vs. effect "reading" information from cause).

Rehder (2014) found some support for both the unobserved inhibitor and the one vs. two mechanism accounts, though more generally he found that none of these rationalizations provide a parsimonious and comprehensive explanation for all the reasoning errors. He argued that it is indeed highly likely that people embellish causal structures given in experiments with additional nodes and links based on their own prior knowledge. However, Rehder proposed that in addition to any embellishments due to background knowledge, some judgments followed an associative-style of reasoning that does not obey the Markov assumption. He proposed taking an individual-differences approach to understanding why certain people are more likely to use an associative style of reasoning.

One surprising aspect about the work on whether people uphold the Markov condition is that there have been very few studies in which people learn the parameters of the causal structure through trial-by-trial experience, and then make judgments.[3] Giving participants statistical experience with the correlations between the variables provides them with direct evidence that X and Z are statistically independent given Y. Park and Sloman (2013) conducted one experiment of this sort. Their participants inferred that $P(z{=}1|y{=}1,x1) > P(z{=}1|y{=}1,x{=}0)$, though $P(z{=}1|y{=}0,x{=}1) = P(z{=}1|y{=}0,x{=}0)$; a violation of the Markov condition only when $y{=}1$. As discussed above, this pattern actually fits the proposal that people infer an unobserved inhibitory cause of both X and Y. However, the modified structure with the unobserved inhibitory cause is still unfaithful to the data that they observed; in the learning data X and Z were independent when $y{=}1$. This raises a question for future research: if being told the structure and experiencing data faithful to the structure is not sufficient to stamp out violations of the Markov assumption, what is?

3.3 Qualitative and Quantitative Inferences when Reasoning about Causal Structures

Rottman and Hastie (2014) reviewed inferences on many different types of causal structures including one link [$X \rightarrow Y$], chains [$X \rightarrow Y \rightarrow Z$], common cause [$X \leftarrow Y \rightarrow Z$] common effect [$X \rightarrow Y \leftarrow Z$], and diamond [$X \leftarrow W \rightarrow Y$ and $X \rightarrow Z \leftarrow Y$] structures. For each of these structures we reviewed evidence about how well people make inferences on one variable given different observed combinations of the others (e.g., X given knowledge about Y, or Y given knowledge of X and Z, etc.).

[3] In the sections above on learning causal structures, when the true structure is $X \rightarrow Y \rightarrow Z$, people tend to also infer the link $X \rightarrow Z$, suggesting that they are not fully aware of the conditional independence. This section focuses on reasoning about the causal structure rather than learning, though of course they are related.

We concluded that for almost all the causal structures (see the section below on explaining away for an exception) the inferences tend to go in the right direction. For example, for the chain [$X\rightarrow Y\rightarrow Z$], if both causal relations between $X\rightarrow Y$ and $Y\rightarrow Z$ were positive or both were negative, people tended to infer a positive relation between X and Z. But if one of the links was positive and the other negative people infer a negative causal relation (Baetu & Baker, 2009).

The previously mentioned studies involving interventions and observations on a diamond structure [$X\leftarrow W\rightarrow Y$ and $X\rightarrow Z\leftarrow Y$] also reveal how sensitive people are to the parameters of the structure (Meder et al., 2008, 2009). These studies systematically manipulated the base rates of some of the variables, and also the strengths of some of the causal links. Even though the causal structures involved 4 variables, and the inference required reasoning with two routes from X to Z, all of these manipulations had influences on subjects' inferences in the predicted directions. In sum, reasoning habits often correspond to the qualitative predictions of the CBN framework.

Yet, despite the qualitative correspondence between human inferences and the normative judgments based on the CBN framework, the quantitative correspondence is not so tight. For example, in one condition when inferring the probability of Z given X for the study above, the normative answer was 12.5%, yet subjects answered on average 37%. Given that 50 is the middle of the scale, 37% is actually considerably closer to a default of 50% than the normative answer. This pattern of conservative results, judgments too close to the center of the scale was very common across many studies reviewed in Rottman and Hastie (2014). For example, for both chain [$X\rightarrow Y\rightarrow Z$] and common cause [$X\leftarrow Y\rightarrow Z$] structures people do typically infer a correlation between A and C, however, often the correlation is considerably weaker than the correlation in the data that the subjects observed (Baetu & Baker, 2009; Bes, Sloman, Lucas, & Raufaste, 2012; Hagmayer & Waldmann, 2000; Park & Sloman, 2013). There are multiple possible interpretations of such effects such as response biases or memory errors (Costello & Watts, 2014; Hilbert, 2012) or potentially priors on the parameters (Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Yeung & Griffiths, 2011). More evidence is needed to understand why these effects occur, and also to understand the accuracy when reasoning with more than 3 or 4 variables.

3.4 Reasoning about Explaining Away Situations

The previous section already addressed quantitative inferences on causal networks, and the conclusion is that for the most part people are fairly good at making inferences, though there is a conservative bias. However, there is one type of inference called explaining away that stands out as particularly difficult. Explaining away inferences involve judgments of $P(x{=}1|y{=}1, z{=}1)$ and $P(x{=}1|y{=}1,z{=}0)$ on a common effect structure [$X\rightarrow Y\leftarrow Z$]. The reason that explaining away inferences are so challenging is that once the state of Y is known, X and Z actually become <u>negatively</u> dependent, so the normative pattern of inference is $P(x{=}1|y{=}1, z{=}1) < P(x{=}1|y{=}1,z{=}0)$. This is unlike any other type of inference. For example, on a chain structure [$X\rightarrow Y\rightarrow Z$], positive relations between X and Y and Y and Z mean that there is a positive relation between X and Z; $P(x{=}1|z{=}1) > P(x{=}1|z{=}0)$, and because of the Markov assumption $P(x{=}1|y{=}1, z{=}1) = P(x{=}1|y{=}1,z{=}0)$.

In terms of Figure 1 [*smoke* → *cardiovascular disease* ← *age*], explaining away could involve inferring the probability that someone smokes given their age and knowing that they have cardiovascular disease. Out of patients who have cardiovascular disease, knowing that a given patient is old means that it is less necessary to infer that they smoke in order to explain the cardiovascular disease; old age “explains away” the cardiovascular disease. If the patient is young it becomes more necessary to infer that they smoke - otherwise what explains the cardiovascular disease? In sum, when the two causes have a positive influence on the effect, the causes become negatively related controlling for the effect.

Prior evidence did not decisively identify how well people explain away (Morris & Larrick, 1995; Sussman & Oppenheimer, 2011). The newest and clearest evidence suggests that people have considerable difficulties when making explaining-away judgments (Rehder, 2014). Though sometimes people get the direction of the inference correct, $P(x=1|y=1, z=1) < P(x=1|y=1,z=0)$, they often are ambivalent about the direction of the inference, and sometimes think that Z would have a positive effect on X, $P(x=1|y=1, z=1) > P(x=1|y=1,z=0)$. Rehder proposed that this type of reasoning is more akin to an associative spreading-activation network than causal reasoning. Reid Hastie and I have also recently collected data on explaining away; unlike the previous research we gave participants learning data so that they could reason from experience rather than just from the causal structure, and so that they also have direct evidence that $P(x=1|y=1, z=1) < P(x=1|y=1,z=0)$. We sometimes found explaining away that was much weaker than normatively predicted by the CBN framework, and other times inference patterns in the opposite direction from explaining away.

The challenge people have with explaining away is somewhat mysterious. There are no other types of causal inference that give reasoners so much trouble, yet at the same time explaining away has also been touted as a fundamental strength of human reasoning (Jones, 1979; Kelley, 1972; Pearl, 1988, p. 49). There are also other results in which explaining away does occur. Oppenheimer et al. (2013) created stories to elicit explaining away. For example, participants were told about an animal with three features –feathers, lays eggs, and cannot fly – and asked to rate how likely this animal is to be an ostrich. Being an ostrich is a plausible explanation for why this bird cannot fly. Other participants were given the same three features with one additional feature, that it has a broken wing, which is an alternative cause for not being able to fly. These participants judged the likelihood of being an ostrich as lower than the participants who were not given this feature, suggesting explaining away (see also Oppenheimer & Monin, 2009). So sometimes people do get the direction of the inference correct.[4] (This study did not have normatively-correct quantitative answers to compare human inferences against, and it also tests a comparison of).

An additional complexity is that explaining away is related to another phenomenon. Explaining away involves inferring the probability of X given knowledge of Y and Z on the structure $[X \rightarrow Y \leftarrow Z]$. Another much studied topic is inferring the causal strength of X on Y. As already discussed, people know that they must control for Z when inferring the causal

[4] This study is different from the ones above in two ways. First, this study did not have a normatively correct quantitative answer to compare human inferences against. Second, this study tests the comparison P(ostrich | feathers, lays eggs, cannot fly, broken wing) vs. P(ostrich | feathers, lays eggs, cannot fly), not P(ostrich | feathers, lays eggs, cannot fly, no broken wing). This is analogous to $P(x=1|y=1, z=1)$ vs. $P(x=1|y=1)$ instead of $P(x=1|y=1,z=0)$, so it is a slightly different comparison.

strength of *X* on *Y*. However, when *Z* is a very strong cause of *Y*, it is not uncommon for people to infer that the strength of *X* is very weak, weaker than it actually is; sometimes this is called “discounting” (Goedert & Spellman, 2005). This discounting effect is related to explaining away in that both phenomena require understanding that two causes are competing to explain an effect.

In sum, there is conflicting evidence as to when, whether, and how much people explain away. Despite the fact that explaining away has been studied for 40 years, there is still important work to be done to reconcile these findings.

3.5 Do Causal Relations Bias Reasoning?

It is a fairly common view in psychology that it is it is easier for people to reason from causes to effects than from effects to causes (Pennington & Hastie, 1993; White, 2006), and this hypothesis is supported by evidence that cause to effect judgments are made faster than effect to cause judgments (Fernbach & Darlow, 2010). The question in this section is whether cognitive ease has an influence on the inferences themselves.

Tversky and Kahneman (1980) found that causal inferences are higher when reasoning from causes to effects. Similarly, Bes et al. (2012) found that when making inferences on the chain [$X \rightarrow Y \rightarrow Z$], inferences of $P(z=1|x=1)$ were higher than $P(x=1|z=1)$. Additionally, both of these inferences were higher than inferences $P(z=1|x=1)$ or $P(x=1|z=1)$ on a common cause [$X \leftarrow Y \rightarrow Z$] structure. These differences are especially instructive because their participants received trial-by-trial training, according to which all the inferences mentioned above should have been equivalent. They speculate that making inferences between *X* and *Z* on the common cause is harder because one must reason about causal relations going in two different directions, and this increased difficulty could lower the final judgment.

This study reaches a very different conclusion than most of the rest of the articles presented in this chapter. The conclusion is that strength of the inferences is determined by the ease of explaining how the two variables are connected, and that this cognitive ease overwhelms the probabilities participants experience. Even though the explanations for these findings appeal to causal structure and causal direction, they are inconsistent with the CBN framework; the CBN framework predicts that all the inferences mentioned above would be equal given the parameters used in the study.

Though the effects of causal direction were found consistently across three experiments, there are other results that do not entirely fit with the story that cause-to-effect judgments are higher than effect-to-cause judgments. First, Fernbach et al. (2011, p. 13) failed to replicate the study by Tversky and Kahneman (1980). More broadly, Fernbach et al. have found that inferences from causes to effects tend to be lower than the normative standard, but inferences from effects to causes tend to be roughly normative (Fernbach, Darlow, & Sloman, 2010; Fernbach et al., 2011; Fernbach & Rehder, 2013; see also Rehder, this volume b). The explanation is that when reasoning from causes to effects, people sometimes forget that alternative causes could produce the target effect aside from the main cause, though they do not forget about alternative causes when reasoning from the effect to a target cause.

There is some tension between these two sets of findings; Bes et al. found that effect-to-cause judgments are too low (lower than cause-to-effect judgments), whereas Fernbach et al. found that cause-to-effect judgments are too low. However, these results cannot be directly compared because they differ on a variety of dimensions.[5] Fernbach et al. used real world cover stories, asked participants their beliefs about the parameters of the causal structure, and then used those parameters to calculate the normative answers. Because of this approach, Fernbach et al. could not directly compare the cause-to-effect and effect-to-cause inferences and instead compared each inference to the normative standard for that inference. [6] In contrast, Bes et al. (Experiment 3) gave participants trial-by-trial learning data; because the learning data were symmetric the cause-to-effect and effect-to-cause inferences could be directly compared (although the cover story labels for the variables were not counterbalanced).

In sum, though it is intuitive that it is easier to reason from causes to effects rather than vice versa, it is still unclear weather or how cognitive fluency and neglect of alternative causes manifest in judgments; it is not clear exactly whether or when cause-to-effect judgments are higher than effect-to-cause judgments. It is especially important to come to consensus on these results, or explain why different patterns of reasoning are found in different situations, because both of the patterns of findings imply deviations from the CBN framework.

3.6 Alternative Representations for Causal Reasoning

So far this chapter has presented the CBN framework as a single method of learning causal structures and making inferences. However, like most sophisticated modeling tools, there are actually many choices that the modeler can make. Assuming that human cognitive representations of causality are somehow similar to the representation of a Causal Bayesian network (directed representations of causality, parameters to capture the strength of causal relations and base rates), these choices correspond to different cognitive representations of the task and background knowledge. An accurate description of causal reasoning requires clarifying the representations being used. In the next two sections I discuss some representational options, and whether they can be empirically distinguished.

Consider the case that you are told that X and Z both cause Y [$X \rightarrow Y \leftarrow Z$], you experience a set of learning trials that instantiate the statistical relations between these variables, and are subsequently asked to infer $P(x=1|y=1,z=1)$. Figure 8 details four possible processes for making the judgment.

The first route, the dashed line, involves making the inference directly from the experienced data. Whenever a learner experiences data that instantiates the causal structure it is possible to come to the correct inference by focusing on the experienced data

[5] I thank Michael Waldmann for highlighting these differences.

[6] Assuming a world in which causes and effects have the same base rates, on average, Fernbach et al.'s findings imply that cause-to-effect judgments would be lower than effect-to-cause judgments. However, Fernbach et al. actually assume a world in which effects have higher base rates than causes on average. Fernbach et al. (2011, p. 13) claim that a normative CBN analysis shows that inferences of $P(\text{effect=1}|\text{cause=1})$ should be higher than $P(\text{cause=1}|\text{effect=1})$ 65% of the time when integrating across the entire parameter space with uniform priors. The reason for this finding is due to the fact that they assumed that there are alternative factors that can generate effects but not inhibit effects. This same analysis shows that even though causes have a base rate .5 on average, effects have a base rate of .625. So their analysis is only appropriate in worlds in which there are no inhibitory factors.

and ignoring the causal structure. For example, in order to calculate $P(x=1 \lfloor y=1, z=1)$, a reasoner just needs to remember the total number of observations in which all three variables were 1, $P(x=1, y=1, z=1)$, and divide this by the total number of observations in which $y=1$ and $z=1$ ignoring X, $N(y=1, z=1)$; see Figure 8. This reasoning process can be thought of as similar to exemplar models of categorization; inference is performed by recalling specific exemplars.

The remaining three options all involve elaborating the causal structure with different kinds of parameters, and inference is performed through a computation on the parameters. Though in some ways the inference itself seems more complicated, the cognitive benefit is that the learner only needs to store the structure and the parameters, not all the individual instances. The difference between these three options is how they represent the conditional probability distribution of Y, the probability of Y given the causes X and Z. This conditional probability distribution is denoted as $P(Y=y|X=x, Z=z)$, which means the probability that Y is in a particular state (y = 0 or 1), given that X and Z are each in particular states, x and z.

Representation 1, involves calculating the conditional probability distribution $P(Y=y|X=x, Z=z)$ directly from the experienced data. For example, the probability that $y=1$ given that $x=1$ and $z=1$, is calculated directly from rows 1 and 3 from the experience table. Inference can then proceed through simple probability theory (Figure 8). Heckerman (1998) provides a tutorial on this approach, and provides citations to other exact and approximate inference algorithms.

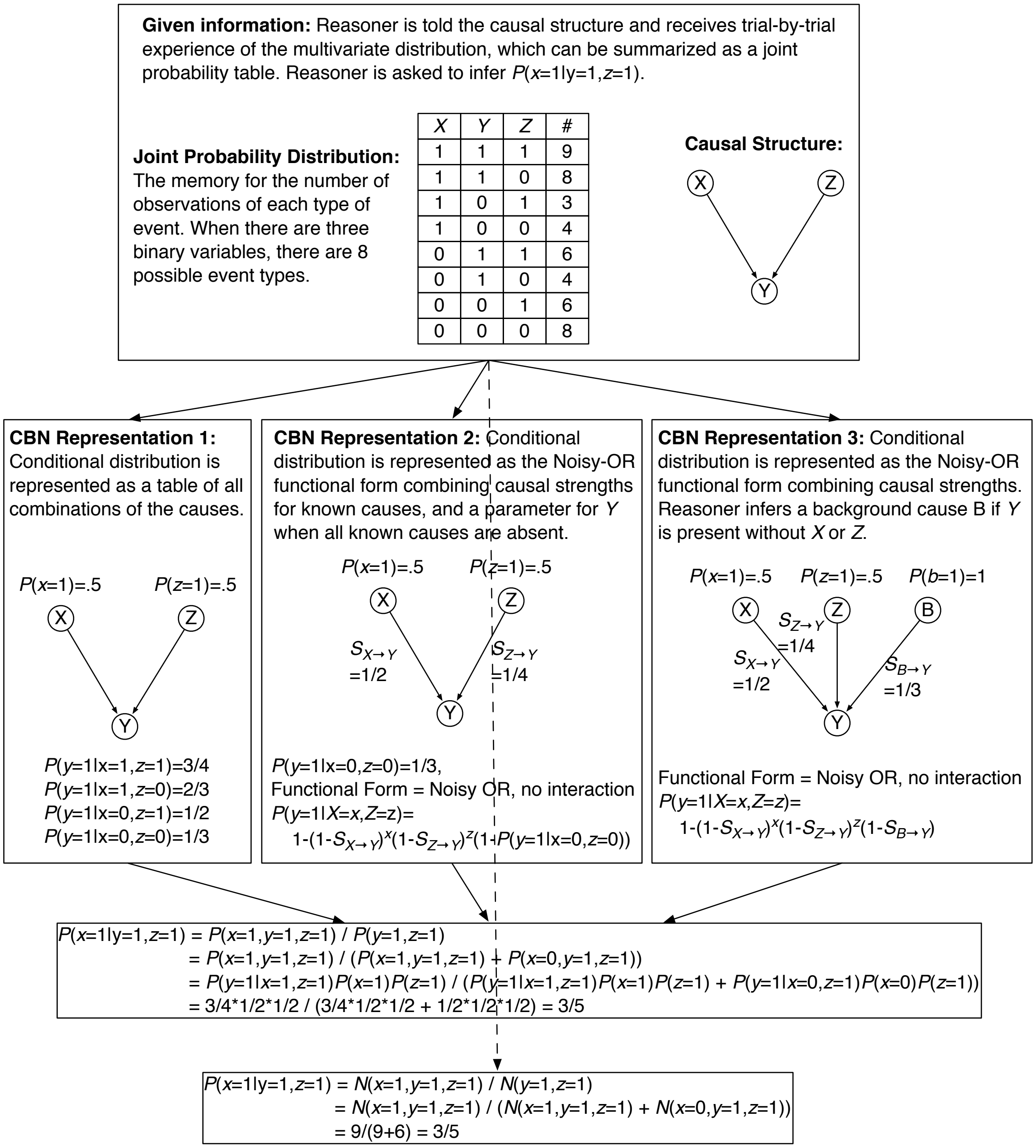
Representation 2 does not directly represent the conditional probability distribution $P(Y=y|X=x, Z=z)$, but instead assumes that people spontaneously infer causal strengths from the learning data. $S_{X\to Y}$ and $S_{Z\to Y}$ refer to the strength of X on Y and Z on Y, respectively. The most popular way to represent causal strengths in the normative psychological literature is using causal power theory, which assumes that causes combine through a Noisy-OR function (Cheng, 1997; Novick & Cheng, 2004 also see Sections 2.2 and 2.3). This approach also requires the learner to estimate the probability that the effect is present without any of its causes, $P(Y=1|x=0, z=0)$. The causal strengths and the functional form (Noisy-OR) subsequently allow a reasoner to deduce the conditional distribution $P(Y=y|X=x, Z=z)$, which would be used for making the inference $P(x=1 \lfloor y=1, z=1)$. The critical difference between Representation 1 vs. 2 is that Representation 2 embodies the assumption that X and Z combine through a Noisy-OR function and do not interact (Novick & Cheng, 2004); the Noisy-OR assumption is the reason why Representation 2 has only 5 parameters instead of the 6 parameters in Representation 1.

Representation 3 is very similar to Representation 2; however, instead of representing the parameter $P(Y=1|x=0, z=0)$, an additional background cause B is added that explains the cases when $Y=1$ but X and Z are 0. In Figure 8, B is assumed to always be present, and to have a strength of 1/3.

The question raised by these four options is whether some sort of representation of causal structure and strength mediates the process of making an inference based on experienced data, or whether the inference is made directly from the experienced data (dashed line). If indeed some sort of causal structure representation mediates the inference, which form of representation gets used? All four approaches make the exact same predictions, so they are difficult to distinguish empirically.

I do not know of any studies that address the first question, whether a causal structure representation mediates the process of making an inference based on experience data. However, there are some studies that have attempted to distinguish the nature of the CBN representation, specifically the difference between Representations 2 vs. 3.

Figure 8: Four Possible Processes for Making an Inference



Given information: Reasoner is told the causal structure and receives trial-by-trial experience of the multivariate distribution, which can be summarized as a joint probability table. Reasoner is asked to infer $P(x{=}1|y{=}1,z{=}1)$.

Joint Probability Distribution: The memory for the number of observations of each type of event. When there are three binary variables, there are 8 possible event types.

X	*Y*	*Z*	#
1	1	1	9
1	1	0	8
1	0	1	3
1	0	0	4
0	1	1	6
0	1	0	4
0	0	1	6
0	0	0	8

Causal Structure:

CBN Representation 1: Conditional distribution is represented as a table of all combinations of the causes.

$P(x{=}1){=}.5$ $P(z{=}1){=}.5$

$P(y{=}1|x{=}1,z{=}1){=}3/4$
$P(y{=}1|x{=}1,z{=}0){=}2/3$
$P(y{=}1|x{=}0,z{=}1){=}1/2$
$P(y{=}1|x{=}0,z{=}0){=}1/3$

CBN Representation 2: Conditional distribution is represented as the Noisy-OR functional form combining causal strengths for known causes, and a parameter for *Y* when all known causes are absent.

$P(x{=}1){=}.5$ $P(z{=}1){=}.5$

$P(y{=}1|x{=}0,z{=}0){=}1/3$,
Functional Form = Noisy OR, no interaction
$P(y{=}1|X{=}x,Z{=}z){=}$
$1-(1-S_{X\rightarrow Y})^x(1-S_{Z\rightarrow Y})^z(1-P(y{=}1|x{=}0,z{=}0))$

CBN Representation 3: Conditional distribution is represented as the Noisy-OR functional form combining causal strengths. Reasoner infers a background cause B if *Y* is present without *X* or *Z*.

$P(x{=}1){=}.5$ $P(z{=}1){=}.5$ $P(b{=}1){=}1$

X
Z
B
Y
$S_{Z\rightarrow Y}$ =1/4
$S_{X\rightarrow Y}$ =1/2
$S_{B\rightarrow Y}$ =1/3

Functional Form = Noisy OR, no interaction
$P(y{=}1|X{=}x,Z{=}z){=}$
$1-(1-S_{X\rightarrow Y})^x(1-S_{Z\rightarrow Y})^z(1-S_{B\rightarrow Y})$

$$\begin{aligned}P(x{=}1|y{=}1,z{=}1) &= P(x{=}1,y{=}1,z{=}1) / P(y{=}1,z{=}1)\\ &= P(x{=}1,y{=}1,z{=}1) / (P(x{=}1,y{=}1,z{=}1) + P(x{=}0,y{=}1,z{=}1))\\ &= P(y{=}1|x{=}1,z{=}1)P(x{=}1)P(z{=}1) / (P(y{=}1|x{=}1,z{=}1)P(x{=}1)P(z{=}1) + P(y{=}1|x{=}0,z{=}1)P(x{=}0)P(z{=}1))\\ &= 3/4*1/2*1/2 / (3/4*1/2*1/2 + 1/2*1/2*1/2) = 3/5\end{aligned}$$

$$\begin{aligned}P(x{=}1|y{=}1,z{=}1) &= N(x{=}1,y{=}1,z{=}1) / N(y{=}1,z{=}1)\\ &= N(x{=}1,y{=}1,z{=}1) / (N(x{=}1,y{=}1,z{=}1) + N(x{=}0,y{=}1,z{=}1))\\ &= 9/(9+6) = 3/5\end{aligned}$$

Note: *N* refers to the number of trials of observations of a particular type.

Krynski and Tenenbaum (2007) studied how well people make inferences on the famous mammogram problem. In this problem, participants are told that breast cancer (cause) almost always results in a positive mammogram test (effect), and they are told the base rate of breast cancer. They are also told that mammograms have false positives 6% of the time. Critically, this false positive rate is framed either as inherent randomness (Representation 2, which has a parameter to represent the probability of the effect when the known cause is absent), or due to a benign cyst (an explicit background cause like in Representation 3). Krynski and Tenenbaum found that participants' judgments about the probability of breast cancer given a positive mammogram were considerably more accurate when the false positive rate was framed as being caused by a benign cyst, suggesting that Representation 3 may be the most intuitive.

A number of recent studies help to clarify this finding by Krynski and Tenenbaum. First, though this facilitation of Bayesian responding by a causal framing has sometimes been found, the effect has not always been consistent (Hayes et al., 2015; Hayes, Newell, & Hawkins, 2013; McNair & Feeney, 2014, 2015). There appear to be two main reasons for the inconsistency. First, the causal framing has a bigger influence for participants who have higher mathematical abilities (McNair & Feeney, 2015). Second, the facilitation effect is often seen in a reduction in extreme overestimations (called base rate 'neglect'); however, the final judgments are often lower, closer to the normative response, but still not quite 'normative' (McNair & Feeney, 2014). A plausible explanation for this effect was put forth by Hayes, Hawkins, and Newell (2015; 2014), who found that the causal framing increases the perceived relevance of the false positive information. They concluded that the causal framing mainly has an influence on the attention paid to the false positive rate and possible the construction of a representation of the problem, but does not necessarily help participants to actually use the false positive rate in a normative way when calculating the posterior inference.

In sum, it seems like having explicit alternative causes (Representation 3) may facilitate accurate causal inference. That said, this finding raises a worrying prospect that causal reasoning is apparently fragile enough that it can be harmed by a small framing. If causal reasoning is robust why can't people translate between these representations by mentally generating an alternative cause to represent the false positive rate?

More broadly, the purpose of this analysis in Figure 8 was to show that the CBN framework can be instantiated in multiple possible ways. Different articles present different versions. Even though they all make similar if not identical predictions, these alternative versions present different cognitive processes involved in making the inference. In order to move from a computational-level theory to an algorithmic-level theory it will be necessary to further clarify the representations and inference process. It is especially critical to clarify whether a causal structure representation mediates causal inference when a reasoner has experienced learning data because in such instances it is possible to make inferences directly from the remembered experiences without thinking about the causal structure at all.

3.7 Even More Complicated Alternative Models for Causal Reasoning

The previous section discussed four possible implementations of the CBN framework. However, in reality there are many more possibilities. A fully Bayesian treatment of learning and inference allows for a way for prior knowledge to influence the learning and inference processes. In regards to a causal structure, there are three possible roles of prior information; prior beliefs about the network, about the integration function, and about the strengths or parameters.

First, whereas Representations 2 and 3 in Figure 8 both assume one particular functional form, the Noisy-OR, in reality learning is not this simple. Section 2.2 on functional forms already covered experiments on how people learn the specific way in which multiple causes combine to produce an effect, and how this belief shapes further learning and reasoning about the causal system. (Beckers & Miller, 2005; Lucas & Griffiths, 2010; Waldmann, 2007). Thus, a fully Bayesian version of Figure 8 would allow for multiple possible integration functions and priors on those functions.

Second, the parameters in Figure 8 were calculated by using point estimates. For example, the parameter $P(y=1|x=0, z=0)$ for Representations 1 and 2, and the $S_{B\rightarrow Y}$ parameter in Representation 3, are all given as exactly 1/3 in Figure 8, which was calculated by comparing rows 6 and 8 in the data table. If a point estimate of the parameters is used, then all four approaches produce exactly the same inferences. Alternatively, another option is that people represent uncertainty about all of the parameters based on the amount of data experienced. If this second approach is used, then Representation 1 will make somewhat weaker inferences than Representations 2 and 3, because Representation 1 requires inferring an additional parameter. Additionally, people may have prior beliefs about causal strengths that may bias the learning and inference process. For example, Lu et al. argued that people believe causes to be sparse and strong (Lu et al., 2008). Given the data in Figure 8, the sparse and strong priors pull the strengths downward; instead of a strength of .50, the sparse and strong priors would produce a strength estimate of .43 and with more data the estimate gets closer to .50. In contrast, Yeung and Griffiths (Yeung & Griffiths, 2015) found that people have priors such that they believe that most candidate causes are very strong. If people had such priors it would result in causal strength estimates above .50. Priors on strength would have a down-stream influence on inference; the stronger the causal strength beliefs, the stronger the inferences should be.

Third, people often have prior beliefs about the causal network. Lu et al.'s sparse and strong prior suggests that people believe that fewer causes are more likely than many causes (Lu et al., 2008). In a related vein, Meder et al. (2014) proposed that when performing an inference, even if told a causal structure, people may entertain the possibility that another causal structure could actually be the true structure, which can influence the judgment. In particular, Meder et al. told participants the structure [$X\rightarrow Y$], had them observe contingency data so that they could learn the statistical relation between X and Y, and then had them make an inference of $P(x=1|y=1)$. They found evidence that when the causal strength of X on Y is fairly weak, people may not believe the structure [$X\rightarrow Y$] and instead entertain the possibility that X and Y may be unrelated. This general approach, that people may entertain the possibility that the causal structure presented by the experimenter may not actually be the true causal structure has also been used to explain violations of the Markov assumption (see Section 3.2). One problem with this account, however, is that when there are more than two variables it is unclear what set of alternative structures is entertained, and considering multiple possibilities would quickly become cognitively unwieldy.

In sum, allowing for the possibility that people think about multiple possible strengths, functional forms, and causal structures makes the CBN framework very flexible, and on a case-by-case level it seems plausible that people may actually have priors for any of these aspects of the network. However, incorporating all of these priors makes the reasoning task much harder than any of the options in Figure 8, and it seems unlikely that people are always engaged in reasoning with all these priors simultaneously. Thus, it will be important to understand when people make use of the priors and how well they incorporate priors with observed data for making inferences.

4 Final Questions, Future Directions, and Conclusions

Throughout the chapter I have highlighted questions and future directions. In this section I repeat some of those questions and add some new ones. I believe that these questions are critical for having a thorough and accurate understanding of human causal learning and reasoning.

1) Though recently there have been more attempts to explore other functional forms, the vast majority of research on the CBN framework has investigated binary variables that combine through a Noisy-OR function. There has been very little theorizing about what causal strength means, for example, when causes and or effects are multilevel (Pacer & Griffiths, 2011; Rottman, 2016; White, 2001). For example, is the human interpretation of causal strength for multilevel (e.g., Gaussian) variables analogous to effect size measures for linear regression? What is the relation between function learning and causal strength learning? Do people face any challenges or use different heuristics when learning causal structures from multilevel rather than binary variables? In sum, causal reasoning is extremely diverse, and it will be critical to broaden our experimental paradigms to capture this diversity.
2) One of the goals of cognitive psychology is to understand the representations that people use for thought. As Figure 8 demonstrates, there are multiple possible representations for how people reason about causal structures, and many of these representations make exactly the same (or very similar) predictions. Clarifying which sorts of representations are used will help develop a more precise descriptive account of causal reasoning.
3) So far the CBN framework has been framed as a computational-level theory of human causal reasoning. However, the computations involved in inferring a causal structure from data, or making inferences on a network (e.g., Figure 8) are very complex. Thus, an important goal is to develop a process-level account of how people actually perform these inferences. A number of theorists have proposed various heuristics for causal learning, which often come close to the optimal solution, and often have equal or better fit to participants' inferences(Bramley et al., 2015; Coenen et al., 2015; Lagnado & Sloman, 2004; Rottman & Keil, 2012; Rottman et al., 2014; Steyvers et al., 2003). Yet so far this heuristics approach has been disconnected and has often taken the back seat to proof of concept demonstrations that the CBN framework can model human learning. More attention to how these inferences are actually made through a process-level account will help provide psychological insight into this fascinating and complex reasoning process.
4) Lastly, all of the studies on human causal reasoning give participants toy examples and sample data in short periods of time. It is unclear how well this research strategy captures actual causal reasoning in the real world, which involves long-term accumulation of data and many more variables. An ideal approach would be to find a real-world domain involving causes and effects that includes records of experiences. For example, a highly accurate electronic medical records system might

> in the future permit us to track a doctor's experiences with all the variables in Figure 1 to see if the doctor's judgments fit closely with his or her personal experiences.

The causal Bayesian network framework has entirely reshaped the landscape of research on causality to the point that it is now rare see articles that investigate causal learning without mentioning the CBN framework. Whereas research on causal reasoning used to be primarily about inferences between a single cause and effect, now the central questions are about larger causal structures. Thus, the new focus is on how people learn the structure and determine causal directionality, how people simplify complex structures into smaller units using the Markov assumption, and how various beliefs captured in the network such as the integration function influence learning and reasoning. Even older questions such as elemental causal learning have benefitted tremendously from the CBN framework by reinterpreting strength as a parameter in the causal network.

On the descriptive side, the most important fact about human causal reasoning is that humans are remarkably good causal reasoners; we adeptly incorporate many different beliefs when learning and reasoning (e.g., integration functions, autocorrelation, causal directionality), we can learn about quite complicated causal relations (e.g., unobserved causes that interact with observed causes), and we often do so with remarkably little data. The introduction of the CBN framework has revealed many of these capacities that were previously unknown and has also raised important questions such as how such as how to develop a process-level account of these sophisticated inferences, how closely do the representations of the CBN framework map on to the actual representations that we use for causal reasoning, how causal reasoning occurs with more diverse sorts of stimuli and in more naturalistic environments. Answering these questions will not only help us develop a more accurate and complete picture of human causal reasoning but may also identify ways to help people become even better causal reasoners.

References:

Baetu, I., & Baker, a G. (2009). Human judgments of positive and negative causal chains. *Journal of Experimental Psychology: Animal Behavior Processes*, *35*(2), 153–68. http://doi.org/10.1037/a0013764

Beckers, T., & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning, *31*(2), 238–249. http://doi.org/10.1037/0278-7393.31.2.238

Bes, B., Sloman, S. A., Lucas, C. G., & Raufaste, E. (2012). Non-bayesian inference: causal structure trumps correlation. *Cognitive Science*, *36*(7), 1178–203. http://doi.org/10.1111/j.1551-6709.2012.01262.x

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars - How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *41*(3), 708–731. http://doi.org/10.1037/xlm0000061

Cartwright, N. (1989). *Nature's capacities and their measurement.* Oxford, UK: Clarendon Press.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405. http://doi.org/10.1037//0033-295X.104.2.367

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*(2), 365–82. http://doi.org/10.1037/0033-295X.99.2.365

Coenen, A., Rehder, B., & Gureckis, T. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, *79*, 102–133. http://doi.org/10.1016/j.cogpsych.2015.02.004

Costello, F., & Watts, P. (2014). Surprisingly Rational : Probability Theory Plus Noise Explains Biases in Judgment, *121*(3), 463–480.

Danks, D. (2003). Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology*, *47*(2), 109–121. http://doi.org/10.1016/S0022-2496(02)00016-0

Eells, E. (1991). *Probabilistic causality*. Cambridge, UK: Cambridge University Press.

Fernbach, P. M., & Darlow, A. (2010). Causal Conditional Reasoning and Conditional Likelihood. In *Proceedings of the 32nd annual conference of the Cognitive Science Society.* (p. 305). Austin, TX: Cognitive Science Society. http://doi.org/10.1177/0272989X9101100408

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science*, *21*(3), 329–36. http://doi.org/10.1177/0956797610361430

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General*, *140*(2), 168–85. http://doi.org/10.1037/a0022100

Fernbach, P. M., & Rehder, B. (2013). Cognitive shortcuts in causal inference. *Argument & Computation*, *4*(1), 64–88. http://doi.org/10.1080/19462166.2012.682655

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 678–93. http://doi.org/10.1037/a0014928

Ghahramani, Z. (1998). Learning dynamic Bayesian networks. In *Adaptive processing of sequences and data structures* (Vol. 1387, pp. 168–197). Springer Berlin Heidelberg.

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

Goedert, K. M., & Spellman, B. A. (2005). Nonnormative discounting: There is more to cue interaction effects than controlling for alternative causes. *Animal Learning & Behavior*, *33*(2), 197–210. http://doi.org/10.3758/BF03196063

Goodie, A. S., Williams, C. C., & Crooks, C. L. (2003). Controlling for causally relevant third variables. *The Journal of General Psychology*, *130*(4), 415–30. http://doi.org/10.1080/00221300309601167

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*(1), 3–32. http://doi.org/10.1037/0033-295X.111.1.3

Griffiths, T. L., & Tenenbaum, J. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–84. http://doi.org/10.1016/j.cogpsych.2005.05.004

Griffiths, T. L., & Tenenbaum, J. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716. http://doi.org/10.1037/a0017201

Hagmayer, Y., & Waldmann, M. R. (2000). Simulating Causal Models : The Way to Structural Sensitivity. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 214–219). Austin, TX: Cognitive Science Society.

Hayes, B. K., Hawkins, G. E., Newell, B. R., Hayes, B. K., Hawkins, G. E., & Newell, B. R. (2015). Consider the Alternative: The Effects of Causal Knowledge on Representing and Using Alternative Hypotheses in Judgments Under Uncertainty. *Journal of Experimental Psychology : Learning , Memory , and Cognition*, *41*(6). http://doi.org/10.1037/xlm0000205

Hayes, B. K., Hawkins, G. E., Newell, B. R., Pasqualino, M., & Rehder, B. (2014). The role of causal models in multiple judgments under uncertainty. *Cognition*, *133*(3), 611–620. http://doi.org/10.1016/j.cognition.2014.08.011

Hayes, B. K., Newell, B. R., & Hawkins, G. E. (2013). Causal model and sampling approaches to reducing base rate neglect. *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, 567–572.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In M. I. Jordan (Ed.), *Learning in Graphical Models* (pp. 301–354). Springer.

Hiddleston, E. (2005). A causal theory of counterfactuals. *Nous*, *39*, 632–657. http://doi.org/10.1111/j.0029-4624.2005.00542.x

Hilbert, M. (2012). Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. *Psychological Bulletin*, *138*(2), 211–37. http://doi.org/10.1037/a0025940

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, *79*(1), 1–17.

Jones, E. (1979). The rocky road from acts to dispositions. *The American Psychologist*, *34*(2), 107–17.

Kelley, H. H. (1972). Causal schemata and the attribution process. In E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the Causes of Behavior* (pp. 151–174). Morristown, NJ: General Learning Press.

Krynski, T. R., & Tenenbaum, J. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology. General*, *136*(3), 430–50. http://doi.org/10.1037/0096-3445.136.3.430

Lagnado, D. A., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *30*(4), 856–76. http://doi.org/10.1037/0278-7393.30.4.856

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *32*(3), 451–60. http://doi.org/10.1037/0278-7393.32.3.451

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

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 955–84. http://doi.org/10.1037/a0013256

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical bayesian models. *Cognitive Science*, *34*(1), 113–47. http://doi.org/10.1111/j.1551-6709.2009.01058.x

Mayrhofer, R., & Waldmann, M. R. (2011). Heuristics in Covariation-based Induction of Causal Models: Sufficiency and Necessity Priors. In C. H. Carlson & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 3110–3115). Austin, TX.

Mayrhofer, R., & Waldmann, M. R. (2015). Agents and causes: Dispositional intuitions as a guide to causal structure. *Cognitive Science*, *39*(1), 65–95. http://doi.org/10.1111/cogs.12132

Mccormack, T., Frosch, C., Patrick, F., & Lagnado, D. A. (2015). Temporal and Statistical Information in Causal Structure Learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *41*(2), 395–416.

McNair, S., & Feeney, A. (2014). When does information about causal structure improve statistical reasoning? *Quarterly Journal of Experimental Psychology (2006)*, *67*(4), 625–45. http://doi.org/10.1080/17470218.2013.821709

McNair, S., & Feeney, A. (2015). Whose statistical reasoning is facilitated by a causal structure intervention? *Psychonomic Bulletin & Review*, *22*(1), 258–264. http://doi.org/10.3758/s13423-014-0645-y

Meder, B., Gerstenberg, T., Hagmayer, Y., & Waldmann, M. R. (2010). Observing and Intervening : Rational and Heuristic Models of Causal Decision Making. *The Open Psychology Journal*, (3), 119–135.

Meder, B., & Hagmayer, Y. (2009). Causal induction enables adaptive decision making. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society* (pp. 1651–1656). Austin, TX: Cognitive Science Society.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2008). Inferring interventional predictions from observational learning data. *Psychonomic Bulletin & Review*, *15*(1), 75–80. http://doi.org/10.3758/PBR.15.1.75

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2009). The role of learning data in causal reasoning about observations and interventions. *Memory & Cognition*, *37*(3), 249–64. http://doi.org/10.3758/MC.37.3.249

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure Induction in Diagnostic Causal Reasoning. *Psychological Review*, *121*(3), 277–301. http://doi.org/10.1037/a0035944

Morris, M. W., & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review*, *102*(2), 331–355. http://doi.org/10.1037/0033-295X.102.2.331

Murphy, K. P. (2002). *Dynamic bayesian networks: representation, inference and learning*. University of California, Berkeley.

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review*, *111*(2), 455–85. http://doi.org/10.1037/0033-295X.111.2.455

Oppenheimer, D. M., & Monin, B. (2009). Investigations in spontaneous discounting. *Memory & Cognition*, *37*(5), 608–614. http://doi.org/10.3758/MC.37.5.608

Oppenheimer, D. M., Tenenbaum, J., & Krynski, T. R. (2013). Categorization as Causal Explanation. Discounting and Augmenting in a Bayesian Framework. In *Psychology of Learning and Motivation - Advances in Research and Theory* (Vol. 58, pp. 203–231). Elsevier. http://doi.org/10.1016/B978-0-12-407237-4.00006-2

Pacer, M. D., & Griffiths, T. L. (2011). A rational model of causal induction with continuous causes. In *Advances in Neural Information Processing Systems* (pp. 2384–2392).

Park, J., & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. *Cognitive Psychology*, *67*(4), 186–216. http://doi.org/10.1016/j.cogpsych.2013.09.002

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers.

Pearl, J. (1996). Structural and Probabilistic Causality. In D. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *Psychology of learning and motivation: Causal Learning* (Vol. 34, pp. 393–435). San Diego: Academic Press.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press.

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

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*(1), 29–46.

Rehder, B. (2014). Independence and Dependence in Human Causal Reasoning. *Cognitive Psychology*, *72*. http://doi.org/10.1016/j.cogpsych.2014.02.002

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*(3), 264–314. http://doi.org/10.1016/j.cogpsych.2004.09.002

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory.* (pp. 64–99). New York: Appleton-Century-Crofts.

Rips, L. J. (2010). Two causal theories of counterfactual conditionals. *Cognitive Science*, *34*(2), 175–221. http://doi.org/10.1111/j.1551-6709.2009.01080.x

Rottman, B. M. (2016). Searching for the best cause: Roles of mechanism beleifs, autocorrelation, and exploitation. *Journal of Experimental Psychology: Learning, Memory and Cognition*. http://doi.org/http://dx.doi.org/10.1037/xlm0000244

Rottman, B. M., & Ahn, W. (2009). Causal learning about tolerance and sensitization. *Psychonomic Bulletin & Review*, *16*(6), 1043–9. http://doi.org/10.3758/PBR.16.6.1043

Rottman, B. M., & Ahn, W. (2011). Effect of grouping of evidence types on learning about interactions between observed and unobserved causes. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(6), 1432–48. http://doi.org/10.1037/a0024829

Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*, *140*(1), 109–39. http://doi.org/10.1037/a0031903

Rottman, B. M., & Keil, F. C. (2012). Causal structure learning over time: Observations and interventions. *Cognitive Psychology*, *64*(1-2), 93–125. http://doi.org/10.1016/j.cogpsych.2011.10.003

Rottman, B. M., Kominsky, J. F., & Keil, F. C. (2014). Children Use Temporal Cues to Learn Causal Directionality. *Cognitive Science*, *38*(3), 1–25. http://doi.org/10.1111/cogs.12070

Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science*, *10*(3), 322–32. http://doi.org/10.1111/j.1467-7687.2007.00587.x

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 21, pp. 229–261). San Diego: Academic Press.

Sloman, S. A., & Lagnado, D. A. (2005). Do We “do”? *Cognitive Science*, *29*, 5–39. http://doi.org/10.1207/s15516709cog2901_2

Soo, K., & Rottman, B. M. (2014). Learning Causal Direction from Transitions with Continuous and Noisy Variables. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Spellman, B. A., Price, C. M., & Logan, J. M. (2001). How two causes are different from one: The use of (un)conditional information in Simpson's paradox. *Memory & Cognition*, *29*(2), 193–208. http://doi.org/10.3758/BF03194913

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, Prediction, and Search*. N.Y.: Springer-Verlag.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search.* (2nd ed.). New York, N.Y.: MIT Press.

Steyvers, M., Tenenbaum, J., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489. http://doi.org/10.1016/S0364-0213(03)00010-7

Sussman, A., & Oppenheimer, D. (2011). A Causal Model Theory of Judgment. In C. Hölscher, Carlson & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1703–1708). Austin, TX: Cognitive Science Society.

Thornley, S. (2013). Using Directed Acyclic Graphs for Investigating Causal Paths for Cardiovascular Disease. *Journal of Biometrics & Biostatistics*, *04*(05). http://doi.org/10.4172/2155-6180.1000182

Tversky, A., & Kahneman, D. (1980). Causal Schemata in Judgments Under Uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (pp. 49–72). Hillsdale, New Jersey: Erlbaum.

Waldmann, M. R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. L. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation* (Vol. 34: Causal, pp. 47–88). San Diego.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 53–76. http://doi.org/10.1037//0278-7393.26.1.53

Waldmann, M. R. (2007). Combining versus analyzing multiple causes: how domain assumptions and task context affect integration rules. *Cognitive Science*, *31*(2), 233–56. http://doi.org/10.1080/15326900701221231

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: the role of structural knowledge and processing effort. *Cognition*, *82*(1), 27–58. http://doi.org/10.1016/S0010-0277(01)00141-X

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: two modes of accessing causal knowledge. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(2), 216–27. http://doi.org/10.1037/0278-7393.31.2.216

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*(2), 222–236. http://doi.org/10.1037/0096-3445.121.2.222

Walsh, C., & Sloman, S. A. (2008). Updating beliefs with causal models: Violations of screening off. *Memory and Mind: A Festschrift for Gordon H. Bower*, 345–358.

White, P. A. (2001). Causal judgments about relations between multilevel variables. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27*(2), 499–513.

White, P. A. (2006). The causal asymmetry. *Psychological Review*, *113*(1), 132–47. http://doi.org/10.1037/0033-295X.113.1.132

Yeung, S., & Griffiths, T. L. (2011). Estimating human priors on causal strength. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.* (pp. 1709–1714). Austin, TX: Cognitive Science Society.

Yeung, S., & Griffiths, T. L. (2015). Identifying expectations about the strength of causal relationships. *Cognitive Psychology*, *76*, 1–29. http://doi.org/10.1016/j.cogpsych.2014.11.001