Teaching the Normative Theory of Causal Reasoning*

Richard Scheines,[1] Matt Easterday,[2] and David Danks[3]

Abstract

There is now substantial agreement about the representational component of a normative theory of causal reasoning: Causal Bayes Nets. There is less agreement about a normative theory of causal discovery from data, either computationally or cognitively, and almost no work investigating how teaching the Causal Bayes Nets representational apparatus might help individuals faced with a causal learning task. Psychologists working to describe how naïve participants represent and learn causal structure from data have focused primarily on learning from single trials under a variety of conditions. In contrast, one component of the normative theory focuses on learning from a sample drawn from a population under some experimental or observational study regime. Through a virtual Causality Lab that embodies the normative theory of causal reasoning and which allows us to record student behavior, we have begun to systematically explore how best to teach the normative theory. In this paper we explain the overall project and report on pilot studies which suggest that students can quickly be taught to (appear to) be quite rational.

Acknowledgements

We thank Adrian Tang and Greg Price for invaluable programming help with the Causality Lab, Clark Glymour for forcing us to get to the point, and Dave Sobel and Steve Sloman for several helpful discussions.

* This research was supported by the James S. McDonnell Foundation, the Institute for Education Science, the William and Flora Hewlett Foundation, the National Aeronautics and Space Administration, and the Office of Naval Research (grant to the Institute for Human and Machine Cognition: Human Systems Technology to Address Critical Navy Need of the Present and Future 2004).

[1] Dept. of Philosophy and Human-Computer Interaction Institute at Carnegie Mellon University.
[2] Human-Computer Interaction Institute at Carnegie Mellon.
[3] Department of Philosophy, Carnegie Mellon, and Institute for Human and Machine Cognition, University of West Florida.

1. Introduction

By the early to mid 1990s, a normative theory of causation with qualitative as well as quantitative substance, called "Causal Bayes Nets" (CBNs),[4] achieved fairly widespread acceptance among key proponents in Computer Science (Artificial Intelligence), Philosophy, Epidemiology, and Statistics. Although the representational component of the normative theory is at some level fairly stable and commonly accepted, how an ideal computational agent should learn about causal structure from data is much less settled, and is, in 2005, still a hot area of research.[5] To be clear, the Causal Bayes Net framework arose in a community that had no interest in modeling human learning or representation. They were interested in how a robot, or an ideal computational agent, with obviously far different processing and memory capacities than a human, could best store and reason about the causal structure of the world. Much of the early research in this community focussed on efficient algorithms for updating beliefs about a CBN from evidence (Spiegelhalter and Lauritzen, 1990; Pearl, 1988) , or on efficiently learning the qualitative structure of a CBN from data (Pearl, 1988, Spirtes, Glymour, and Scheines, 2000).

In contrast, the psychological community, interested in how humans learn, not in how they *should* learn if they had practically unbounded computational resources, studied associative and causal learning for decades. The Rescorla-Wagner theory (1972) was offered, for example, as models of how humans (and animals, in some cases), learned associations and causal hypotheses from data. Only later, in the early 1990s, did Causal Bayes Nets make their way into the pscychological community, and only then as a model that might describe everyday human reasoning. At the least, a broad range of psychological theories of human causal learning can be substantially unified when cast as different versions of parameter learning within the CBN framework (Danks, 2005), but it is still a matter of vibrant debate whether and to what degree humans represent and learn about causal claims as per the normative theory of CBNs (e.g., Danks, Griffiths, & Tenenbaum, 2003; Glymour, 1998, 2000; Gopnik, et al., 2001; Gopnik, et al., 2004; Griffiths, Baraff, & Tenenbaum, 2004; Lagnado & Sloman, 2002, 2004; Sloman & Lagnado, 2002; Steyvers, et al., 2003; Tenenbaum & Griffiths, 2001, 2003; Tenenbaum & Niyogi, 2003; Waldmann & Hagmayer, in press; Waldmann & Martignon, 1998).

---

[4] See (Spirtes, Glymour, and Scheines, 2000; Pearl, 2000; Glymour and Cooper, 1999),

[5] See, for example, recent proceedings of Uncertainty and Artificial Intelligence Conferences: http://www.sis.pitt.edu/~dsl/UAI/

Nearly all of the psychological research on human causal learning involves naïve participants, that is, individuals who have not been taught the normative theory in any way, shape, or form. Almost all of this research involves single-trial learning: observing how subjects form and update their causal beliefs from the outcome of a series of trials, each either an experiment on a single individual, or a single episode of a system's behavior. No work, as far as we are aware, attempts to train people normatively on this and related tasks, nor does any work we know of compare the performance of naïve participants and those taught the normative theory. The work we describe in this paper begins just such a project. We are specifically interested in seeing if formal education about normative causal reasoning helps students draw accurate causal inferences.

Although there has been, to our knowledge, no previous research on subjects trained in the normative theory, there has been research on whether naïve subjects approximate normative learning agents. Single trial learning, for example, can easily be described by the normative theory as a sequential Bayesian updating problem. Some psychologists have considered whether and how people update their beliefs in accord with the Bayesian norm (e.g., Danks, *et al*., 2003; Griffiths, *et al*., 2004; Steyvers, *et al.*, 2003; Tenenbaum & Griffiths, 2001, 2003; Tenenbaum & Niyogi, 2003), and have suggested that some people at least approximate a normative Bayesian learner on simple cases. This research does not extend to subjects who have already been taught the appropriate rules of Bayesian updating, either abstractly or concretely.

In the late 1990s, currcular material became available that taught the normative theory of CBNs.[6] Standard introductions to the normative theory in computer science, philosophy, and statistics do not *directly* address the sorts of tasks that psychologists have investigated, however. First, as opposed to single trial learning, the focus is on learning from samples drawn from some population. Second, little or no attention is paid to the severe computational (processing time) and representational (storage space) limitations of humans. Instead, abstractions and algorithms are taught that could not possibly be used by humans on any but the simplest of problems.

In the normative theory, learning about which among many possible causal structures might obtain is typically cast as iterative:

1) enumerate a space of plausible hypotheses,
2) design an experiment that will help distinguish among these hypotheses,
3) collect a sample of data from such an experiment,

[6] See, for example: www.phil.cmu.edu/projects/csr.

4) analyze these data with the help of sophisticated computing tools like R[7] or TETRAD[8] in order to update the space of hypotheses to those supported or consistent with these data, and

5) go back to step 2.

Designing an experiment, insofar as it involves choosing which variable or variables to manipulate, is a natural part of the normative theory and has just recently become a subject of study.[9] The same activity, that is, picking the best among many possible experiments to run, has been studied by Lagnado and Sloman, 2004, Sobel and Kushnir, 2004, Steyvers, *et al.*, 2003, and Waldmann & Hagmayer, in press.

Another point of contact is what a student thinks the data collected in an experiment tells them about the model that might be generating the data. Starting with a set of plausible models, some will be consistent with the data collected, or favored by it, and some will not. We would like to know whether students trained in the normative theory are better, and if so in what way, at determining what models are consistent with the data.

In a series of four pilot experiments, we examined the performance of subjects partially trained in the normative theory on causal learning tasks that involved choosing experiments and deciding on which models are consistent with the data. Although we did not use single-trial learning, we did use tasks similar to those studied recently by psychologists, especially Steyvers, *et al.*, 2003. Our students were trained for about a month in a college course on causation and social policy. The students were not trained in the precise skills tested by our experiments. Although our results are not directly comparable to those discussed in the psychological literature, they certainly suggest that students trained on the normative theory act quite differently than naïve participants.

Our paper is organized as follows. We first briefly describe what we take to be the normative theory of causal reasoning. We then describe the online corpus we have developed for teaching it. Finally, we describe four pilot studies we performed in the fall of 2004 with the Causality Lab, a major part of the online corpus.

2. The Normative Theory of Causal Reasoning

Although Galileo pioneered the use of fully controlled experiments almost 400 years ago, it wasn't until Sir Ronald Fisher's (1935) famous work on experimental design that

[7] www.r-project.org/
[8] www.phil.cmu.edu/projects/tetrad
[9] See Eberhardt, Glymour, and Scheines (2005), Murphy (2001), and Tong and Koller (2001).

real headway was made on the statistical problem of causal discovery. Fisher's work, like Galileo's, was confined to experimental settings in which treatment could be assigned. In Galileo's case, however, all the variables in a system could be perfectly controlled, and the treatment could thus be isolated and made to be the only quantity varying in a given experiment. In agricultural or biological experiments, however, it isn't possible to control all the quantities, e.g., the genetic and environmental history of each person. Fisher's technique of randomization not only solved this problem, but also produced a reference distribution against which experimental results could be compared statistically. His work is still the statistical foundation of most modern medical research.

Representing Causal Systems: Causal Bayes Nets

Sewall Wright pioneered representing causal systems as "path diagrams" in the 1920s and 1930s (Wright, 1934), but until about the middle of the 20th century the entire topic of how causal claims can or cannot be discovered from data collected in non-experimental studies was largely written off as hopeless. Herbert Simon (1954) and Hubert Blalock (1961) made major inroads, but gave no general theory. In the mid 1980s, however, artificial intelligence researchers, philosophers, statisticians and epidemiologists began to make real headway on a rigorous theory of causal discovery from non-experimental as well as experimental data.[10]

Like Fisher's statistical work on experiments, CBNs seek to model the relations among a set of *random variables*, such as an individual's level of education or annual income. Alternative approaches aim to model the causes of individual events, for example the cause(s) of the space shuttle *Challenger* disaster. We confine our attention to relations among variables. If we are instead concerned with a system in which certain types of events cause other types of events, we represent the occurrence or non-occurrence of the events by binary variables. For example, if a blue light bulb going on is followed by a red light bulb going on, we use the variables Red Light Bulb [lit, not lit] and Blue Light Bulb [lit, not lit].

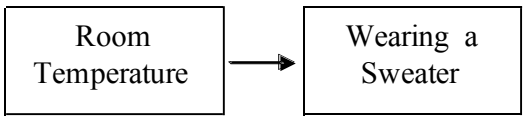
Any approach that models the statistical relations among a set of variables must first confront what we call the *ontological problem*: how do we get from a messy and complicated world to a coherent and meaningful set of variables that might plausibly be related either statistically or causally. For example, it is reasonable to examine the association between the number of years of education and the number of dollars in yearly

[10] See, for example, Spirtes, Glymour and Scheines (2000), Pearl (2000), Glymour and Cooper (1999).

income for a sample of middle aged men in Western Pennsylvania, but it makes no sense to examine the average level of education for the aggregate of people in a state like Pennsylvania and compare it to the level of income for individual residents of New York. Nor does it make sense to posit a "variable" whose range of values is not exclusive because it includes: has blond hair, has curly hair, etc. After teaching causal reasoning to hundreds of students over almost a decade, the ontological problem seems the most difficult to teach and the most difficult for students to learn. We need to study it much more thoroughly, but for the present investigation, we will simply assume it has been solved for a particular learning problem.

Assuming that we are given a set of coherent and meaningful variables, the normative theory involves representing the qualitative causal relations among a set of variables with a directed graph in which there is an edge from X to Y just in case X is a direct cause of Y relative to the system of variables under study. X is a direct cause of Y in such a system if and only if there is a pair of ideal interventions that hold the other variables in the system Z fixed and change only X, such that the probability distribution for Y also changes. We model the quantitative relations among the variables with a set of conditional probability distributions: one for each variable given each possible configuration of values of its direct causes (see Figure 1).

The asymmetry of causation is modeled by how the system responds to ideal intervention, both qualitatively and quantitatively. Consider, for example, a two variable system: Room Temperature (of a room an individual is in) [$<55^o$, $55\text{-}85^o$, $>85^o$], and Wearing a Sweater [yes, no], in which the following graph and set of conditional probability tables describe the system:



P(RT = <55) = .1
P(RT = 55 -85) = .8
P(RT = >85) = .1
P(Wearing a Sweater | RT < 55) = .98
P(Wearing a Sweater | RT = 55 -85) = .5
P(Wearing a Sweater | RT > 85) = .04

Figure 1: Causal Bayes Net

Ideal interventions are represented by adding an intervention variable that is a direct cause of only the variables it targets. Ideal interventions are assumed to have a simple property: if I is an intervention on variable X, then when I is active, it removes all the

other edges into X. That is, the "other" causes of X no longer influence X in the post-intervention, or **manipulated**, system. Figure 2 captures the change and non-change in the Figure 1 graph in response to interventions on Room Temperature (A) and on Wearing a Sweater (B), respectively.

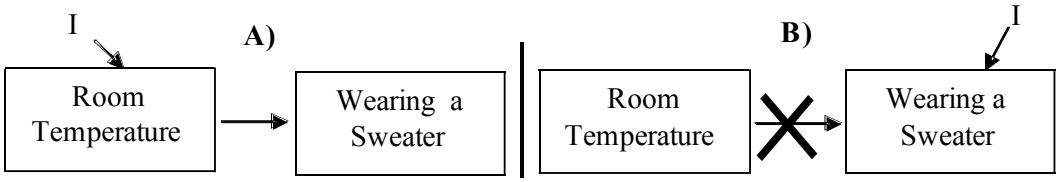


Figure 2: Manipulated graph

Modeling the system's quantitative response to interventions is almost as simple. Generally, we conceive of an ideal intervention as imposing not a value but rather a probability distribution on its target. We thus model the move from the original system to the manipulated system as leaving all conditional distributions intact save those over the manipulated variables, in which case we impose our own distribution. For example, if we assume that the interventions depicted in Figure 2 impose a uniform distribution on their targets when active, then **Figure 3** shows the two manipulated systems that would result from the original system shown in Figure 1.[11]

[11] Ideal interventions are only one type of manipulation of a causal system. We can straightforwardly use the CBN framework to model interventions that affect multiple variables (so-called "fat hand" interventions), as well as those that influence, but do not determine, the values of the target variables (i.e., that do not "break" all of the incoming edges). Of course, causal learning is significantly harder in those situations.

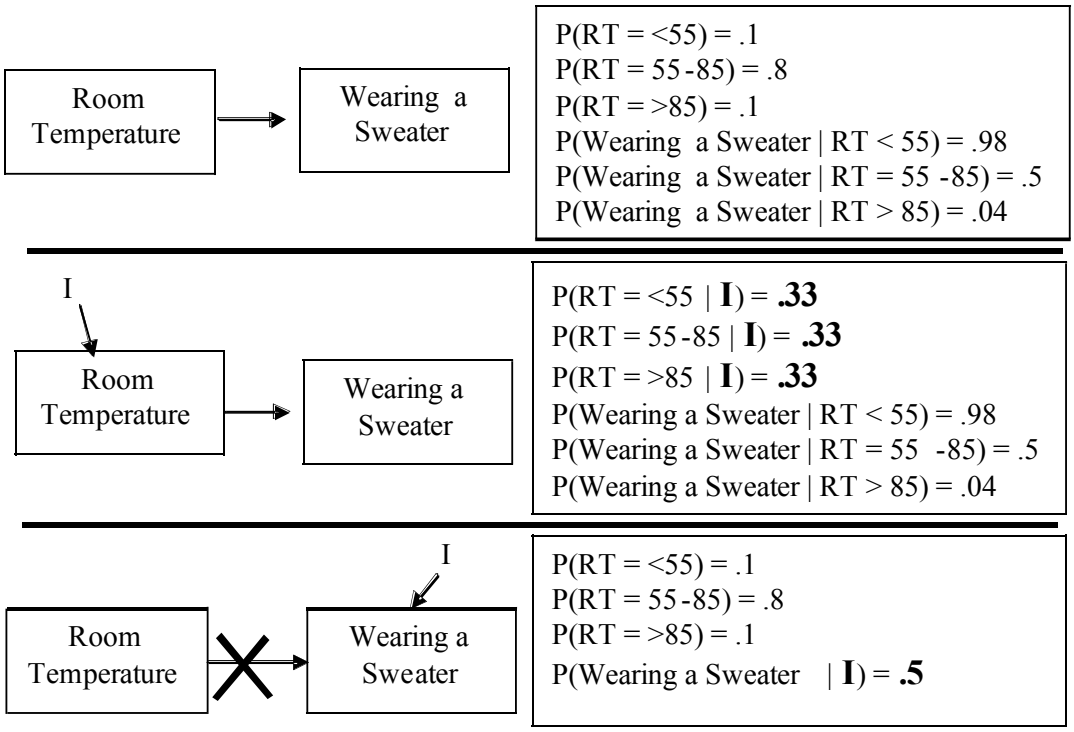


Figure 3: Original and Manipulated Systems

To simplify later discussions, we will include the "null" manipulation (i.e., we intervene on no variables) as one possible manipulation. A Causal Bayes Net *and* a manipulation define a joint probability distribution over the set of variables in the system. If we use "experimental setup" to refer to an exact quantitative specification of the manipulation, then when we collect data we are drawing a sample from the probability distribution defined by the original CBN and the experimental setup.

Learning Causal Bayes Nets

There are two distinct types of CBN learning given data: parameter estimation and structure learning. In parameter estimation, one fixes the qualitative (graphical) structure of the model and estimates the conditional probability tables by minimizing some loss function or maximizing the likelihood of the sample data given the model and its parameterization. In contrast, structure learning aims to recover the qualitative structure of graphical edges. The distinction between parameter estimation and structure learning is not perfectly clean, since "close-to-zero parameter" and "absence of the edge" are roughly equivalent. Danks (2005) shows how to understand most non-Bayes net psychological theories of causal learning (e.g., Cheng, 1997; Cheng & Novick, 1992;

Perales & Shanks, 2003; Rescorla & Wagner, 1972) as parameter estimation theories for particular graphical structures.

A fundamental challenge for CBN structure learning algorithms is the existence of *Markov equivalence classes*: sets of CBNs that make identical predictions about the way the world looks in the absence of experiments. For example, A → B and A ← B both predict that variables A and B will be associated. Any dataset that can be modeled by A → B can be equally well-modeled by A ← B, and so there is no reason—given only observed data—to prefer one structure over the other. This observation leads to the standard warning in science that "correlation does not equal causation." However, *patterns* of correlation can enable us to infer something about causal relationships (or more generally, graphical structure), though perhaps not a unique graph. Thus, structure learning algorithms will frequently not be able to learn the "true" graph from data, but will be able to learn a small set of graphs that are indistinguishable from the "truth."

For learning the structure of the causal graph, the normative theory splits into two approaches: constraint-based and scoring. The constraint-based approach (Spirtes, et. al, 2000) aims to determine the class of CBNs consistent with an inferred (statistical) pattern of independencies and associations, as well as background knowledge. Any particular CBN entails a set of statistical constraints in the population, such as independence and tetrad constraints. Constraint-based algorithms take as input the constraints inferred from a given sample, as well as background assumptions about the class of models to be considered, and output the set of indistinguishable causal structures. That is, the algorithms output the models which (i) entail all and only the inferred constraints, and (ii) are consistent with background knowledge. The inference task is thus split into two parts: 1) *statistical*: inference from the sample to the constraints that hold in the population, and 2) *causal*: inference from the constraints to the Causal Bayes Net or Nets that entail such constraints.

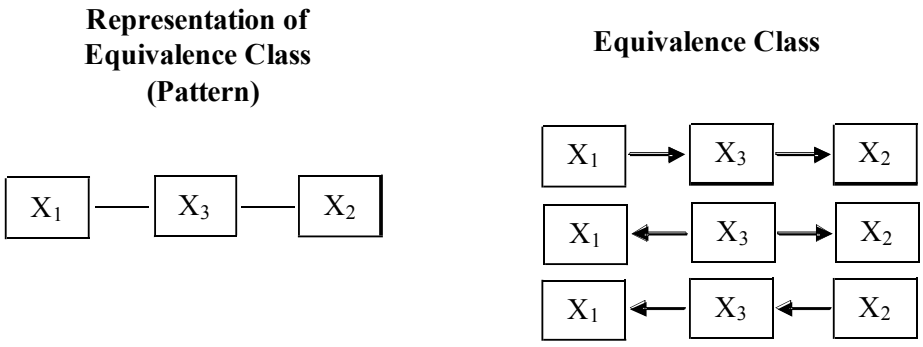


Figure 4: Equivalence Class for X1 _||_ X2 | X3

Suppose, for example, that we observe a sample of 100 individuals on variables X_1, X_2, and X_3, and after statistical inference conclude that X_1 and X_2 are statistically independent, conditional on X_3 (i.e., X1 _||_ X2 | X3). If we also assume that are no unobserved common causes for any pair of X_1, X_2, and X_3, then the PC algorithm (SGS, 2000) would output the **Pattern** shown on the left side of Figure 4. That pattern is a graphical object which represents the Markov equivalence class shown on the right side of Figure 4; all three graphs predict exactly the same set of unconditional and conditional independencies. In general, two causal graphs entail the same set of independencies if and only if they have the same adjacencies and same unshielded colliders, where X and Y are adjacent just in case X → Y or X ← Y, and Z is an unshielded collider between X and Y just in case X → Z ← Y and X and Y are *not* adjacent. Thus, in a Pattern, we need only represent the adjacencies and unshielded colliders. Constraint-based searches first compute the set of adjacencies for a set of variables and then try to "orient" these adjacencies, i.e., test for colliders among triples in which X and Y are adjacent, Y and Z are adjacent, but X and Z are not: X – Y – Z.

Testing high order conditional independence relations—relations that involve a large number of variables in the conditioning set—is computationally expensive and statistically unreliable, so the constraint-based approach sequences the tests to minimize the number of higher order conditional independence facts actually tested. Compared to other methods, constraint-based algorithms are extremely fast and under multivariate normal distributions (linear systmes) can handle hundreds of variables. Constraint-based algorithms can also handle models with unobserved common causes. Their drawback is that they are subject to errors if statistical decisions made early in the algorithm are incorrect.

If handed the independence relations true of a population, people could easily perform by hand the computations required by a constraint-based search, even for many causal structures with dozens of variables. Of course, people could not possibly compute all of the precise statistical tests of independence relations required, but they could potentially approximate a subset of such (unconditional and conditional) independence tests (see Danks, 2004 for one *very* tentative proposal).

In the score-based approach (Heckerman, 1995), we assign a "score" to a CBN that reflects both (i) the closeness of the CBN's "fit" of the data, and (ii) the plausibility of the CBN prior to seeing any data. We then search (in a variety of ways) among all the models consistent with background knowledge for the set that have the highest score. The most common scoring based approach is based on Bayesian principles: calculate a score based on the CBN's "prior" – the probability we assign to the model being true before seeing any data, and the model's likelihood – the probability of the observed data given this particular CBN.[12] Scoring based searches are very accurate, but are very slow, as calculating each model's score is very expensive. Given a flat prior over the models (i.e., equal probabilities on all models), the set of models that have the highest Bayesian score is identical to the Markov equivalence class of models output by a constraint-based algorithm.

Bayesian approaches are straightforwardly applied to standard psychological tasks. By computing the posterior over the models after each new sample point, we get a learning dynamics for that problem (as in, e.g., Danks, *et al.*, 2003; Griffiths, *et al.*, 2004; Steyvers, *et al.*, 2003; Tenenbaum & Griffiths, 2003). However, even if naïve subjects act like approximately rational Bayesian structure learners in cases involving 2 or 3 variables, they cannot possibly implement the approach precisely, nor can they possibly implement the approach for larger numbers of variables, e.g., 5-10. Hence, the Bayesian approach is not necessarily appropriate for teaching the normative theory.

3. The Causality Lab

Convinced that the qualitative story behind causal discovery should be taught to introductory level students either prior to or simultaneously with a basic course on statistical methods, a team[13] from Carnegie Mellon and the University of California, San Diego created enough online material for an entire semester's course in the basics of

[12]Strictly, the CBN with parameters set to the maximum-likelihood estimates.
[13]This team included Richard Scheines, Joel Smith, Clark Glymour, David Danks, Mara Harrell, Sandra Mitchell, Willie Wheeler, Joe Ramsey, and more recently, Matt Easterday.

CBNs. By the spring of 2004, over 2,600 students in over 70 courses at almost 30 different colleges or universities had taken all or part of our online course, which is available through Carnegie Mellon's Open Learning Initiative at www.cmu.edu/oli/.

Causal and Statistical Reasoning (CSR) involves three components: 1) 16 lessons, or concept modules; 2) a virtual laboratory for simulating social science experiments, the "Causality Lab"[14]; and 3) a bank of over 120 case studies: reports of "studies" by social, behavioral, or medical researchers. Each of the concept modules contains approximately the same amount of material as a text-book chapter. The Causality Lab embodies the normative theory by making explicit all the ideas we discussed above.

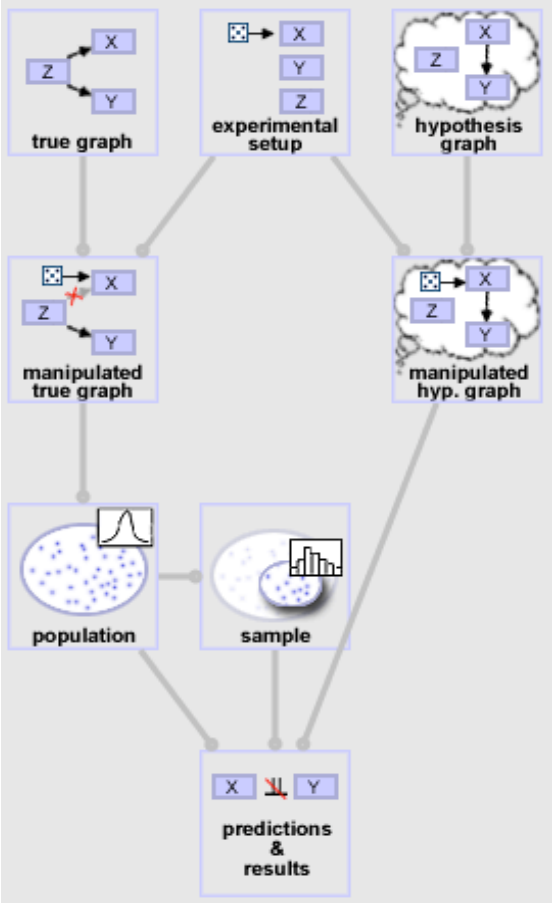


Figure 5: The Causality Lab Navigation Panel

Figure 5 shows the navigation panel for the lab. Each of the icons may be clicked to reveal and in some cases manipulate the contents of an object for a given exercise. The instructor creates the "true" CBN with an exercise building tool, and this constitutes the "true graph" to be discovered by the student. Of course, just as real scientists are confined

[14] The Causality Lab is available free at www.phil.cmu.edu/projects/causality-lab

to one side of the Humean curtain, so are students of the Causality Lab. In most exercises, they cannot access any of the icons in the left column, all of which represent on aspect of the truth to be discovered. Students cannot simply click and see the truth.

Using the earlier example of room temperature and sweaters, suppose the true graph and conditional probability distributions are as given in Figure 1. To fully determine the population from which the student may draw a sample, however, he or she must also provide the (possibly null) experimental setup. Once the student specifies one or more experimental setups, he or she can "collect data" from any of them.

For example, suppose we clicked on the Experimental Setup icon and then created three distinct experimental setups (Figure 6). On the left, both Room Temperature and Sweater will be passively observed. In the middle, the value of Room Temperature will be randomly assigned (indicated by the icon of a die attached to Room_Temp), and the value of Sweater will be passively observed. On the right, the value of Sweater will be randomly assigned, and the value of Room Temperature will be passively observed.

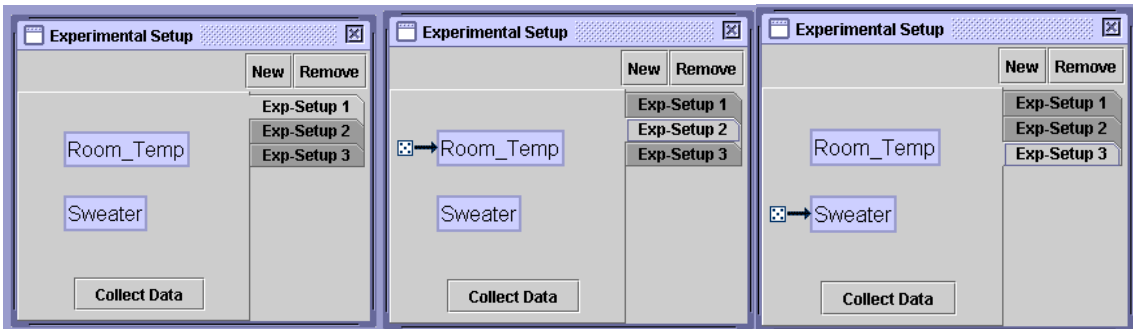


Figure 6: Three Experimental Setups

As the navigation panel in Figure 5 shows, it is the *combination* of the experimental setup and the true CBN that defines the manipulated system, which determines the population probability distribution. So if we click on "Collect Data" from Exp-Setup 1 (far left side of Figure 6), then we will be drawing a sample from the distribution shown at the top of **Figure 3**. If we collect data from Exp-Setup 2, then our sample will be drawn from the distribution shown in the middle of **Figure 3**, and so on. The fact that the sample population depends on both the experimental setup and the true CBN is a pillar of the normative theory, but this fact is rarely, if ever, taught.

Once a sample is pseudo-randomly drawn from the appropriate distribution, we may inspect it in any way we wish. To keep matters as qualitative as possible, however, the focus of the Causality Lab is on independence constraints—the normative theory's

primary connection between probability distributions and causal structure. In particular, the Predictions and Results window allows the student to inspect, for each experimental setup:

1. the independence relations that hold in the population[15]; and
2. the independence relations that cannot be rejected at $\alpha = .05$ by a statistical test applied to any sample drawn from that population

Figure 7: Independence Results

For example, Figure 7 shows the results of an experiment in which wearing a sweater is randomly assigned and a sample of 40 individuals was drawn from the resulting population. The Predictions and Results window indicates that, in the population, Room Temperature and Sweater Wearing are independent (notated as '_||_'). The lab also allows students to inspect histograms or scatterplots of their samples, and then enter their own guesses as to which independence relations hold in a given sample. In this example, a student used the histograms to guess that Room Temperature and Sweater Wearing were associated (not independent), though the statistical test applied to the sample of 40

[15] If the instructor writing the exercise allows the student to "see" the population.

could not reject the hypothesis of independence. Thus, one easy lesson for students is that statistical tests are sometimes better at determining independence relations than students who eyeball sample summaries.

Students can also create hypotheses and then compare the predictions of their hypotheses to the results of their experiments. For example, we may rebel against common sense and hypothesize that wearing a sweater causes the room temperature. The Causality Lab helps the students learn that their hypothetical graph only makes testable predictions about independence in combination with an experimental setup, which leads to a manipulated hypothetical graph (see Figure 5).

Causal Discovery in the Lab

Equipped with the tools of the Causality Lab, we can decompose the causal discovery task into the following steps:

1. Enumerate all the hypotheses that are consistent with background knowledge.
2. Create an experimental setup and collect a sample of data.
3. Make statistical inferences about the independences that hold in the population from the sample
4. Eliminate or re-allocate confidence in hypotheses on the basis of the results from step 3.
5. If no unique model emerges, go back to step 2.

Steps 1 (enumeration) and 3 (statistics) are interesting, though only necessary if one is following a constraint-based approach. The interesting action is in steps 2 and 4. As operationalized in the Causality Lab and defined in the normative theory, the first part of step 2 (experimental design) amounts to determining, for each variable under study, whether that variable will be observed passively or have its values assigned randomly.

Depending upon the hypotheses still under consideration, experimental setups differ in the informativeness of the experiment's results. For example, suppose the currently active hypotheses include: 1) X → Y → Z and 2) X ← Y → Z. An experimental setup (call it ES1) in which X is randomized and Y and Z are passively observed will uniquely determine the correct graph no matter the outcome.[16] A different experiment (call it ES2) in which Z is randomized and X and Y passively observed will tell us nothing, again regardless of the outcome of the experiment. The difference in the experiments'

[16] Assuming, of course, that the statistical inferences are correct.

informativeness arises because the *manipulated graphs* are distinguishable in ES1, but not in ES2 (Figure 8). In ES1, the two possibilities have different adjacencies (X → Y in one, and no edges in the other) and thus entail different sets of independencies. In ES2, however, the two manipulated graphs are indistinguishable; they have the same adjacencies.

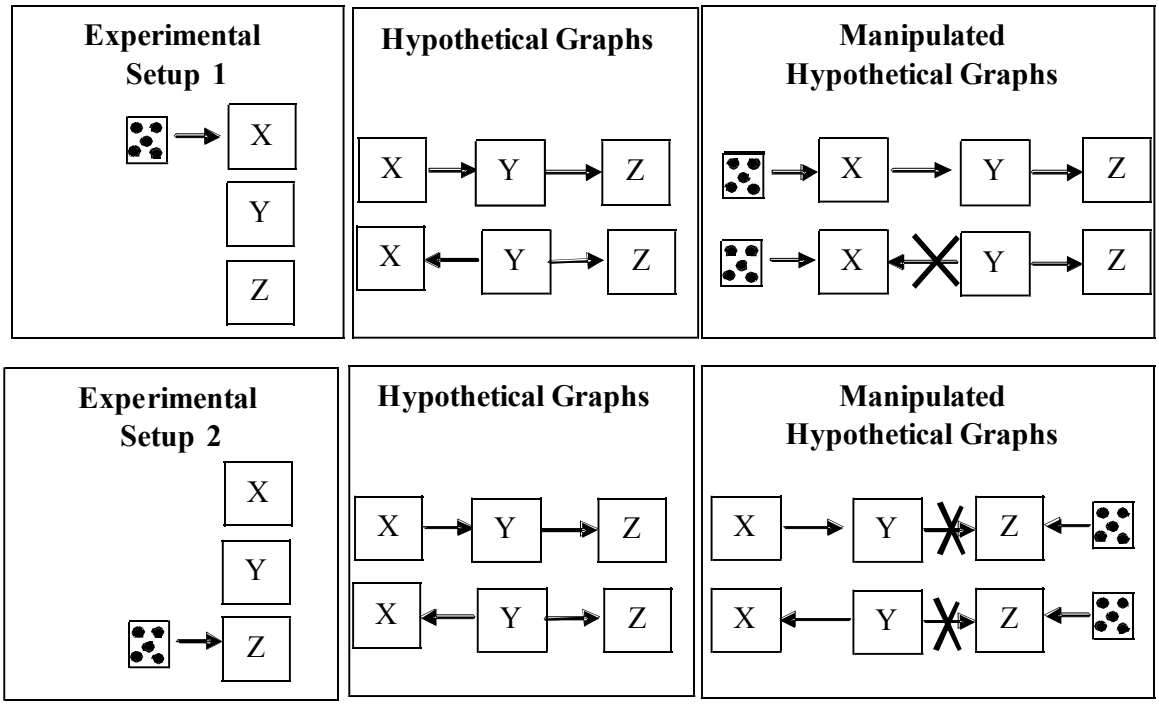


Figure 8: Informative and Uninformative Experimental Setups

From this perspective, the causal discovery task involves determining, for each possible experimental setup one might use, the set of manipulated hypothetical graphs and whether they are (partially) distinguishable. This is a challenging task. What are the general principles for experimental design, if any? When the goal is to parameterize the dependence of one effect on several causes, then there is a rich and powerful theory of experimental design from the statistical literature (Berger, 2005; Cochran and Cox, 1957). When the goal is to discover which among many possible causal structures are true, however, the theory of optimal experimental design is much less developed. From a Bayesian perspective, we must first specify a prior distribution over the hypothetical graphs. Given such a distribution, each experimental setup has an expected gain in information (reduction in uncertainty), and one should thus pick the experiment that would most reduce uncertainty (Murphy, 2001; Tong & Koller, 2001). Computing this gain is intractable for all but the simplest of cases, though Steyvers et al, (2003) argue

that naïve subjects approximate just this sort of behavior. Regardless of the descriptive question, a theory of so-called "active learning" provides normative guidance as to the optimal sequencing of experiments. Taking a constraint-based approach, Eberhardt, Glymour, and Scheines (2005) have shown that for N variables, N-1 experiments that randomize at most a single variable are always sufficient to identify the correct graph, and in the worst case that many are necessary.

Although there is not yet a graphical characterization of the best experiment given a set of active hypotheses, we do have a few powerful heuristics. For example, passive observation is sufficient, under a constraint-based approach, to identify all the adjacencies among a set of variables. Given the adjacencies, an intervention on X will orient all the edges adjacent to X. Suppose X and Z are adjacent. If X and Z are independent after an intervention on X, then the edge is X ← Z; if X and Z are associated, then the edge must be X → Z.

4. Pilot Studies

An obvious question about teaching the normative theory is: does learning it improve student's performance on causal learning tasks? In the fall of 2004, one of us (Scheines) taught an upper level seminar at Carnegie Mellon on Causation and Social Policy. For about a month in the middle of the class, the students went through the CSR material and learned the rudiments of the representational theory of CBNs. The class covered the idea of causation, causal graphs, manipulations, manipulated models, independence, conditional independence, and d-separation,[17] but included no instruction on model equivalence, and no instruction on a procedure for causal discovery. All fifteen of the students in the class agreed to participate in a pilot study in which they were given four discovery tasks. The students all worked for a little over an hour in a computer cluster. We were unable to enforce strict silence between students, and thus the results of our pilot study cannot be considered rigorous. They are, nevertheless, interesting and suggestive.

In all of our experiments, participants were allowed to see the full independence relations that hold in the population defined by an experimental setup of their choice, and so no statistical judgments were required. We recognize that this is different from the standard presentation in psychological experiments, but our intent was to focus on the skills involved in causal discovery from known facts about the population, as opposed to

[17] D-separation enables us to compute the independence relations entailed by a causal graph.

making statistical inferences from samples. To provide familiarity with the Causality Lab interface, all participants were provided a simple training problem. In the training task, the students were instructed to (i) do a passive observation, then (ii) eliminate all the models they could, and finally (iii) determine the true graph using the fewest number of experiments.

Experiment 1

In Experiment 1, we asked students to determine which model in Figure 9 was the true graph in the minimum number of experiments. Students were randomly assigned to a model, and there was no effect of condition.

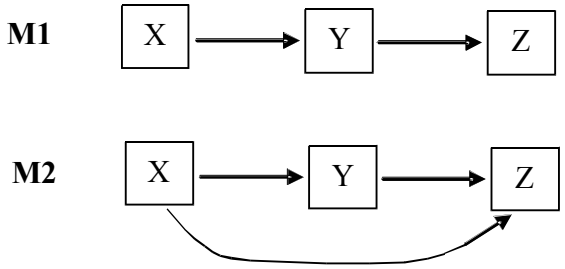


Figure 9: Choices in Experiment 1

The experiment explored whether students understood the difference between direct and indirect causation. All 15 students learned the correct model in a single experiment. We were also interested in the students' choices of experimental targets. Table 1 shows the independence relations entailed by both models in every possible experimental setup, as well as whether M1 and M2 can be distinguished in that experiment. From a normative point-of-view, no one should choose to randomize Z, since that experiment will not distinguish between these two models. Randomizing Y is optimal, as under that intervention the two models make different predictions about both X _||_ Z and X _||_ Z | Y. Steyvers, *et al*. (2003) report a source bias in choosing interventions: people prefer to intervene on variables believed to have no edges into them (i.e., no causes in the system). If this bias holds, then people should prefer to randomize on X, when they randomize on any variable at all. Note that the source bias refers only to choices *among experiments*; no prediction was made about whether people will prefer to experiment or passively observe.

Table 1: Independencies Implied by M1 and M2

Experimental Setup	X _||_ Y	X _||_ Z	X _||_ Z | Y	M1 and M2 Distinguishable?
Passive Observation	Neither	Neither	M1, not M2	Yes
Randomize X	Neither	Neither	M1, not M2	Yes
Randomize Y	Both	M1, not M2	M1, not M2	Yes
Randomize Z	Neither	Both		No

Figure 10 shows the frequency with which each experiment was chosen first. All students were normatively correct; no one chose to randomize on Z. Our students preferred the passive observation, which can be explained by its use in the training experiment. And in contrast to the results reported in Steyvers, *et. al*. (2003), students exhibited no source bias whatsoever: six of the seven who chose to intervene did so on the mediating variable Y.

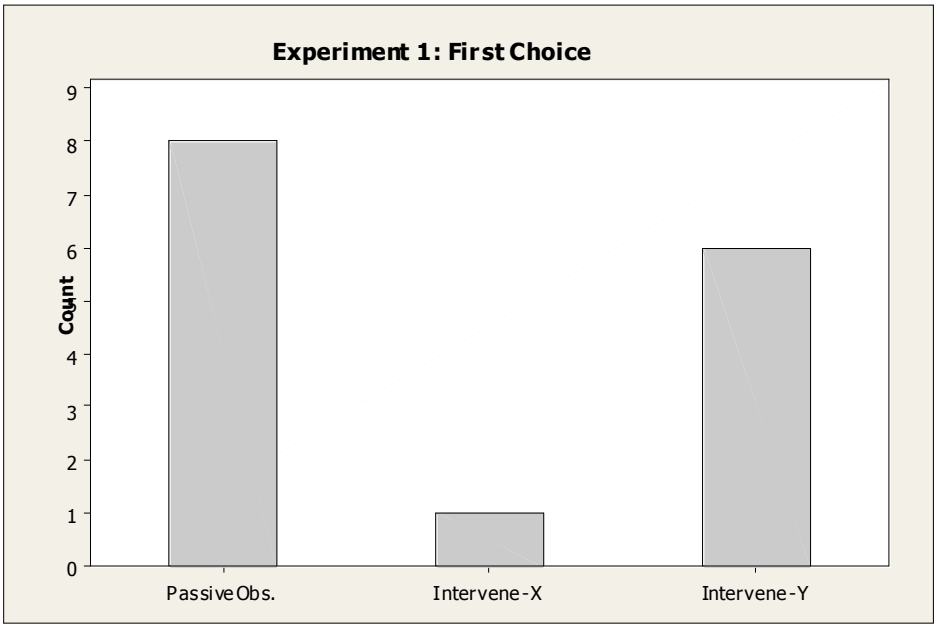


Figure 10: Choice of Experiments in Experiment 1

Experiment 2

In the second experiment, the students had to choose among four possibilities (**Figure 11**). They were again told to find the true graph in the minimum number of experiments, though they understood that they were not required to do the passive observation experiment first. Since M3 and M4 are essentially the same *a priori*, we randomized the students to a true graph of either M1, M2, or M3.

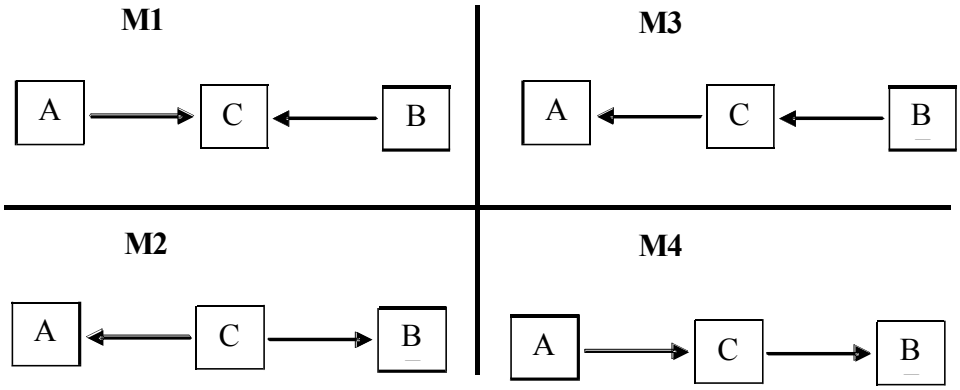


Figure 11: Possibilities in Experiment 2

This experiment aimed to determine whether students could choose an informative intervention; in this problem, the choice of experimental setup matters a lot, as shown in Table 2. For example, if we passively observe all variables, then we can tell only whether M1 is the true model or not the true model (i.e., that the true model is one of: {M2, M3, M4}). The normatively optimal experiment to perform is the one in which the middle variable C is randomized. That experiment is guaranteed to uniquely identify the correct model, regardless of outcome.

Table 2: Distinguishable Models by Intervention Choice

Experimental Setup	Distinguishable?
Passive Observation	M1 from {M2, M3, M4}
Randomize A	M1 from {M2, M3} from M4
Randomize B	M1 from {M2, M4} from M3
Randomize C	M1 from M2 from M3 from M4

Again, students were quite successful in the overall task: 14 out of 15 correctly identified the model. The number of experiments it took to arrive at an answer varied considerably: two experiments was the mode, but several students used three or four. Figure 13 shows the students' first experimental choice (top graph), and the target of the

first intervention they performed regardless of when that first intervention experiment occurred (bottom graph). Clearly, students preferred passive observation as a first choice, but the first choice for an intervention was overwhelmingly the mediator C as opposed to either endpoint variables A or B.

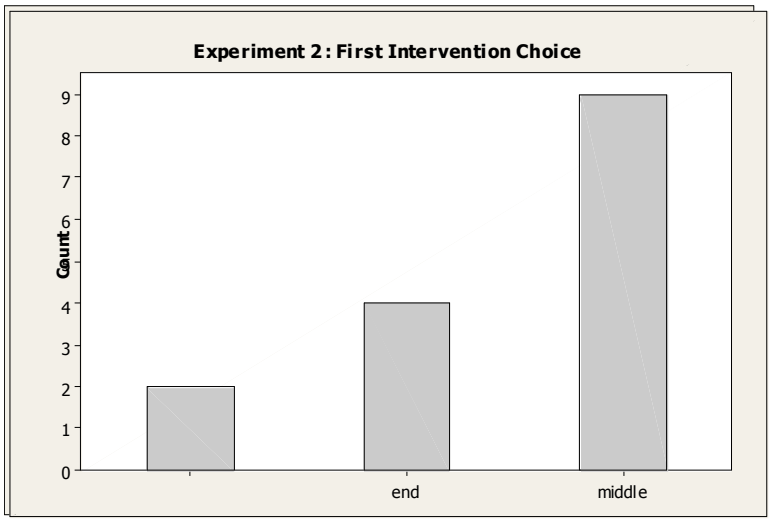


Figure 12: Results of Experiment 2

Experiment 3

In the third experiment, students were told that the true model was one of the models in Figure 13, and we randomly assigned students to have either the Single Edge model or the Chain model (both highlighted in Figure 13) as the true underlying causal structure. (Students were not told that those were the only two possibilities.) All participants were

required to (i) begin with the passive observation experiment, (ii) eliminate as many models as possible after each experiment, and (iii) find the true model in the minimum number of experiments. Students recorded the experimental design used to eliminate each model except the final one. Students did not create use the hypothetical graph window of the Causality Lab, and so had no computational aids to calculate the independencies implied by each hypothesis under a given experimental setup.

In our experiment, over two thirds of participants (11 of 15) answered correctly, and success was independent of condition. Including the passive observation, students averaged just under 3 experiments before reaching a final answer, and the number of experiments was also independent of condition. As one would expect, the 11 students who got the answer right averaged fewer significantly fewer experiments than the 4 who got it wrong. For the remaining analyses, we restrict our attention to the participant responses after only the initial passive observation.

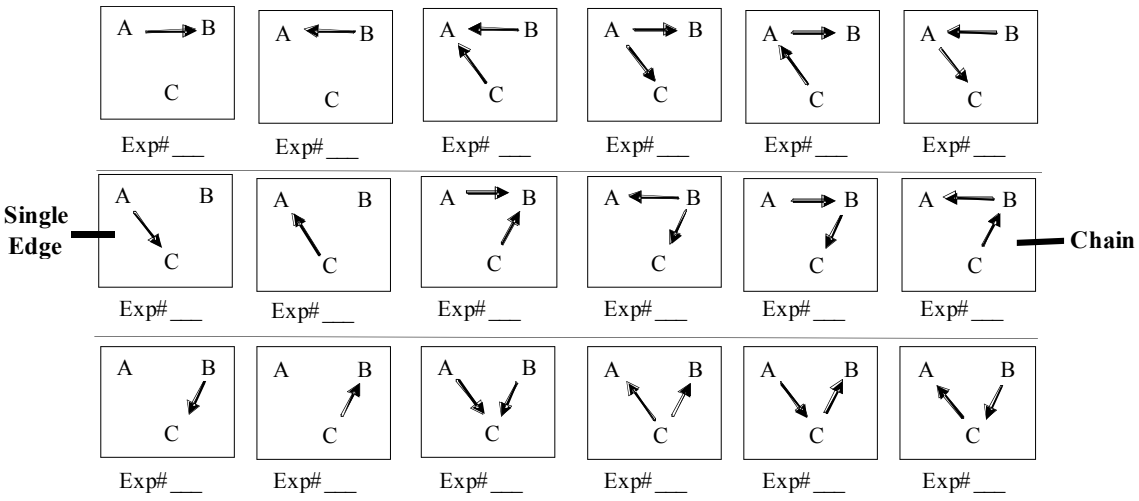


Figure 13: Possibilities for Experiment 3

One question behind our experiment was whether students acted as if they understood the concept of Markov equivalence classes (MECs): sets of models that are indistinguishable by passive observation, since they imply the same set of independence relations. In Figure 14 we show again the 18 possible models, but group them in boxes corresponding to the nine Markov equivalence classes.

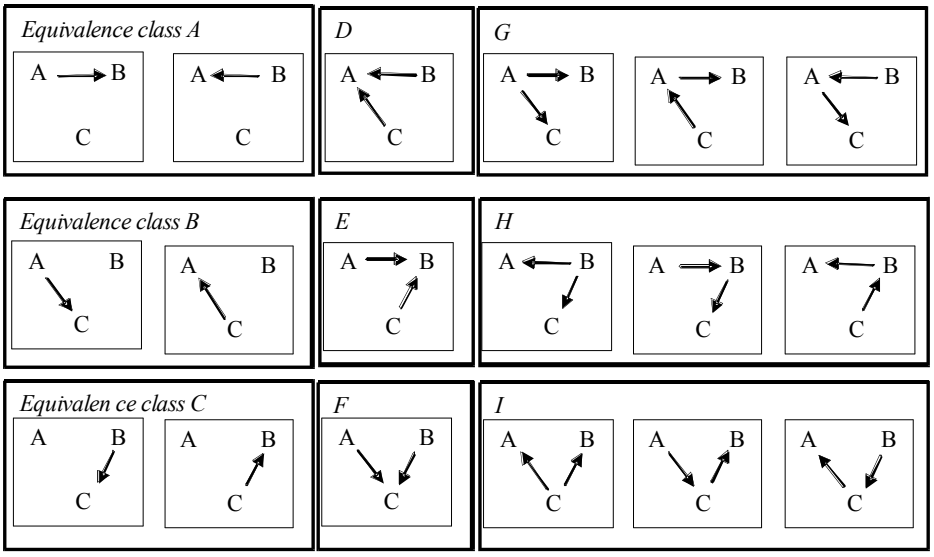


Figure 14: Equivalence Classes for the Passive Observation in Experiment 3

Individuals who (act as if they) understand the idea of Markov equivalence classes should, for every equivalence class, either keep or remove all its members together after the passive observation stage. For equivalence classes D, E, and F, which have only a single member, this necessarily happens, so we exclude those classes. We then define a (weighted) MEC "integrity" score as:

$$\textit{MEC-Integrity} = \frac{\sum_{\text{mec} \in \{A,B,C,G,H,I\}} \begin{cases} |\textbf{mec}| : \text{if all models in mec were included or all excluded:} \\ \textbf{0} \quad : \text{otherwise} \end{cases}}{15}$$

The weighting captures the fact that it is more challenging to have MEC integrity for equivalence classes G, H, and I, which have three members, than it is for equivalence classes A, B, or C, which have two. If a participant always keeps or removes members of a MEC together, then MEC-Integrity equals 1; if members of a MEC are never kept or removed together, then MEC-Integrity equals 0. Figure 16 shows that students exhibited an extremely high degree of MEC integrity: twelve of fifteen participants were perfect, and only one student was massively confused.

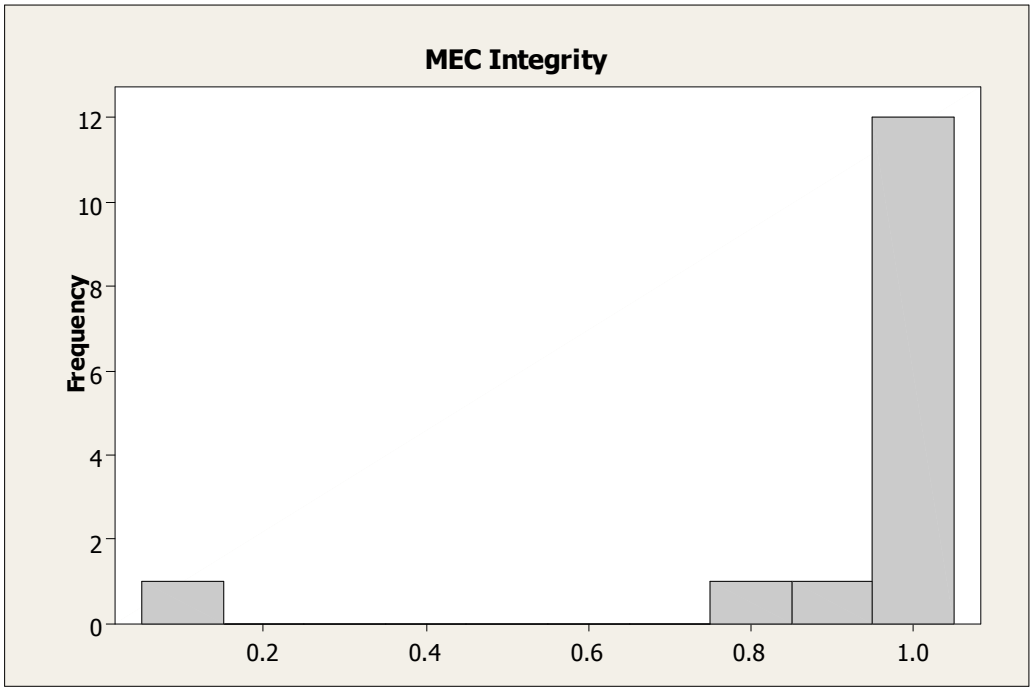


Figure 15: MEC Integrity

Even if someone exhibited perfect MEC integrity, they might still be retaining or excluding the wrong graphs (or the wrong MECs), given the data they received. To measure whether they are including too many graphs, we computed the percentage of commission errors:

$$\textit{Commission Error} = \frac{\#\text{ of graphs retained by student, but not in correct MEC}}{\#\text{ of graphs not in correct MEC}}$$

Similarly, to measure whether they are excluding graphs equivalent to the truth, we computed the percentage of omission errors:

$$\textit{Omission Error} = \frac{\#\text{ of graphs in the correct MEC omitted by the student}}{\#\text{ of graphs in the correct MEC}}$$

Not surprisingly, students were not as good on the accuracy of their inferences. Figure 16 shows that, although their omission error was quite low (very few correct graphs were

left out), students often retained more graphs than were consistent with the passive observation.

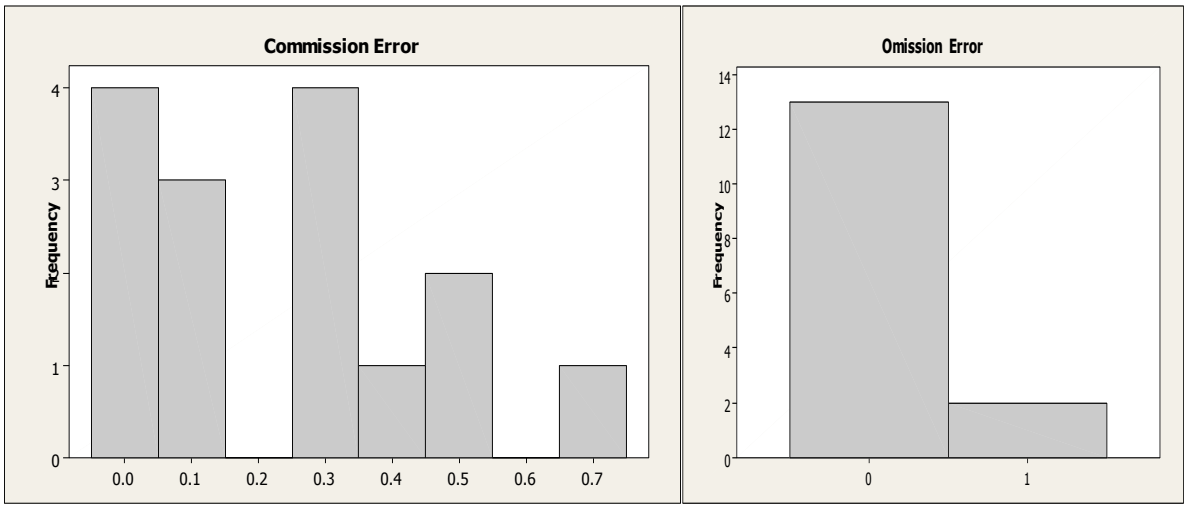


Figure 16: Commission and Omission Error

Interestingly, we think we can explain why. Although we did not include equivalence classes D, E, and F in our computation of MEC-Integrity (because they each have only one graph as a member), we did include those graphs in our calculations of omission and commission error. These graphs each have the same *adjacencies* as some equivalence class, though they differ from the class in edge orientation. In Figure 14, classes D and G share the same adjacencies, as do E and H, and F and I. If, for example, the true graph was C → B → A (part of equivalence class H) and I included every graph in classes E and H, then I would have a perfect score on MEC-Integrity, but a non-zero commission error. In general, if I attend only to adjacencies and ignore orientations, I will (provably) always receive a perfect score on MEC-Integrity, even though I might make a number of commission errors.

After looking at the data we hypothesized that students were quite good at determining the correct adjacencies, but not very good at determining the correct orientations. To explore this, we first computed participants' *Adjacency-Integrity* to determine whether the students included or excluded graphs that share adjacencies as a unit.

$$\textit{Adjacency-Integrity} = \frac{\sum_{adj \in \{A,B,C,\ D+G,E+H,F+I\}} \begin{cases} |\mathbf{adj}| : \text{if all models in adj were included or all excluded:} \\ \mathbf{0} : \text{otherwise} \end{cases}}{18}$$

The histogram in **Figure 17** shows that students had relatively high Adjacency Integrity, suggesting that the high MEC-Integrity scores were due (at least in part) to people keeping/removing graphs with the same adjacencies, and not necessarily those that made the identical observational predictions.

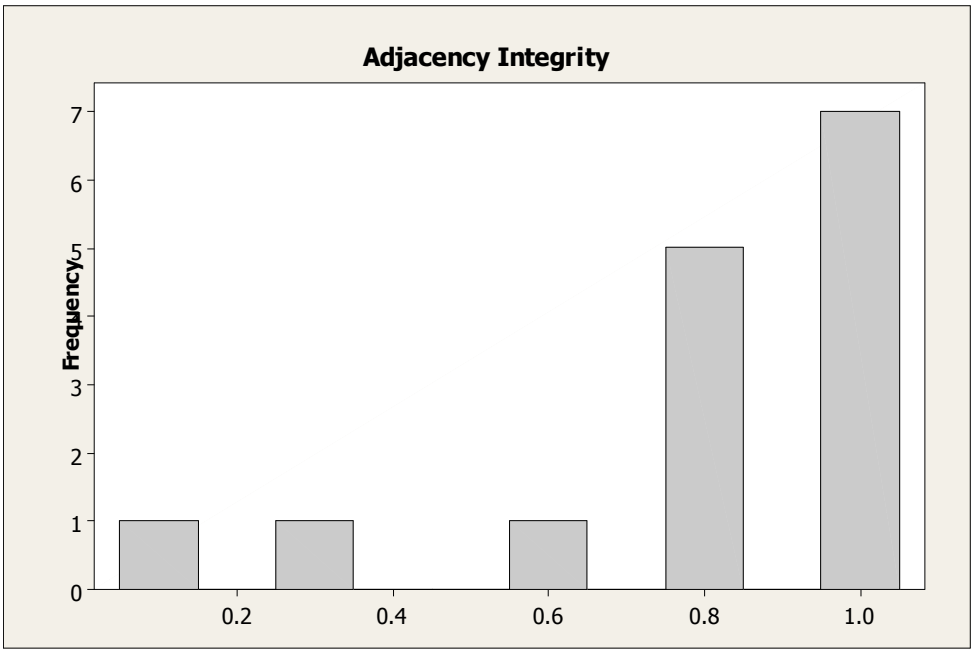


Figure 17: Adjacency-Integrity

This explanation does not completely account for students' performance. Many included graphs that were neither Markov nor adjacency equivalent to the truth. But not all mistakes are quite the same. Suppose the truth is A → B → C. Including the graph A → B ← C is arguably a less severe mistake than including the graph B → C → A. In the former case, the adjacencies were correctly learned, though not the orientations. In the latter case, however, a true adjacency (A — B) was excluded and a false adjacency (C — A) was included. We will say that a graph G is *adjacency consistent* with a graph H if either G's adjacencies are a subset of H's, or vice versa. The former error in this example is adjacency consistent with the truth; the latter error is not.

To better understand the severity of the students' errors, we computed the proportion of the commission errors that were adjacency consistent with the true MEC.

$$\textit{Adjacency Consistent Error} = \frac{\textbf{\# of graphs committed that are adjacency consistent}}{\textbf{\# of graphs committed}}$$

Figure 19 shows that students' errors tend to be adjacency consistent; the majority of their mistakes involved keeping a graph that was either a subgraph or supergraph of the truth

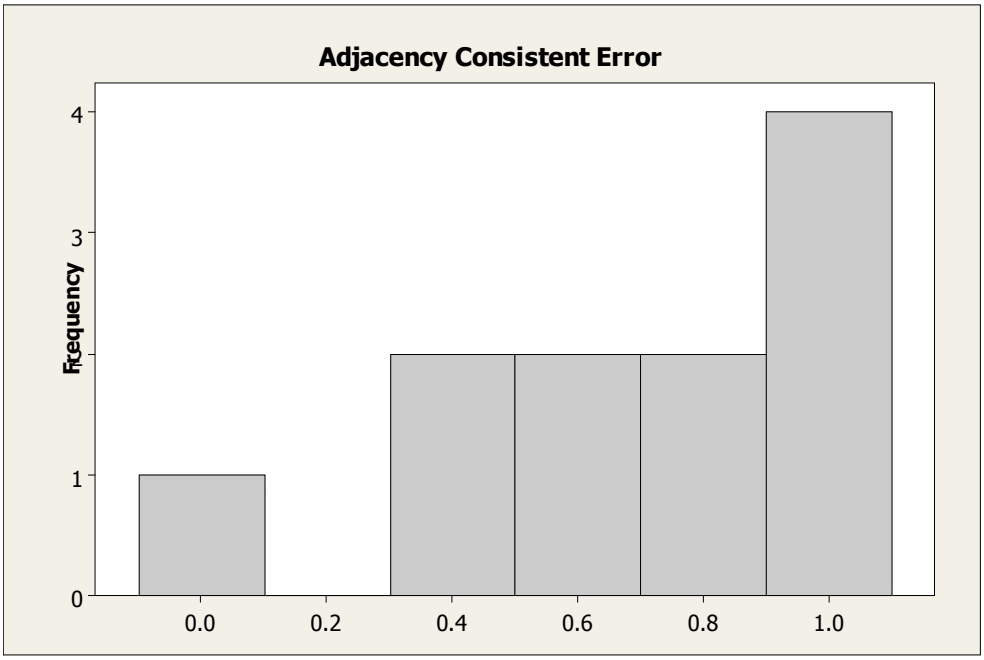


Figure 18: Adjacency Consistent Error

Of course, this high percentage could arise if most graphs are adjacency consistent with the truth (though this is not actually the case in this experiment). To normalize for the number of errors that *could* be adjacency consistent or inconsistent, we also computed:

$$\textit{Adjacency Consistent Inclusion} = \frac{\textbf{\# of committed graphs that are adjacency consistent}}{\textbf{\# of committable graphs that are adjacency consistent}}$$

$$\textit{Adjacency Inconsistent Inclusion} = \frac{\textbf{\# of committed graphs that are adjacency inconsistent}}{\textbf{\# of committable graphs that are adjacency inconsistent}}$$

If students were indifferent between adjacency consistent and adjacency inconsistent errors, then the within-student difference between these two measures should center around 0. As Figure 19 shows, it clearly does not.

These results seem to indicate that:

1. Students have very high Adjacency-Integrity (**Figure 17**);
2. A large fraction of the graphs committed are adjacency consistent (**Error! Reference source not found.**); and
3. The fraction of the committable adjacency consistent graphs that are actually committed is much higher than the fraction of committable adjacency inconsistent graphs that are actually committed (Figure 19).

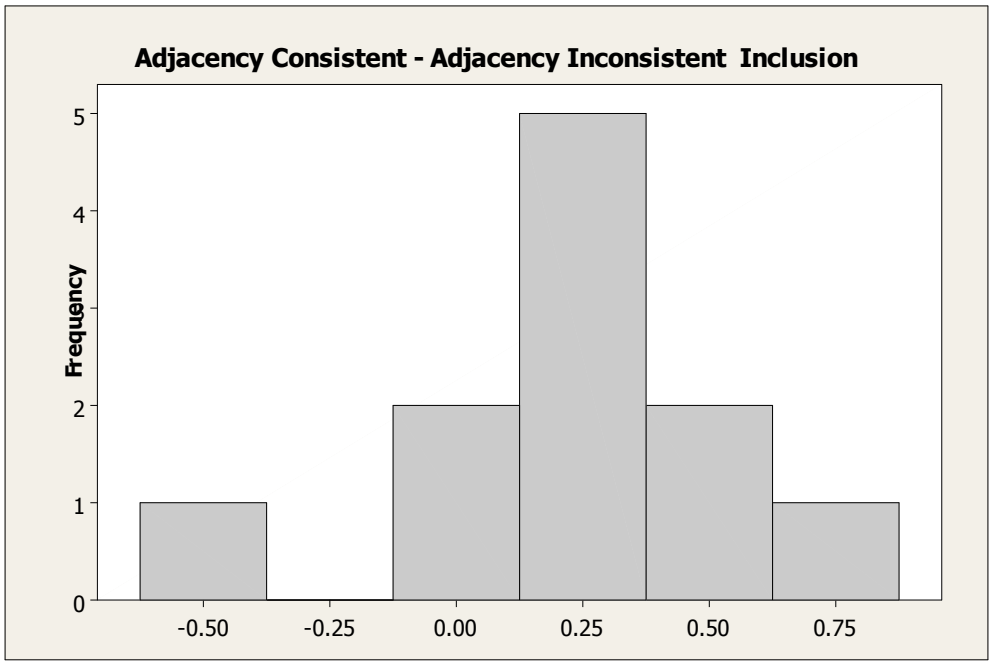


Figure 19: Adjacency Consistent - Adjacency Inconsistent Inclusion

We interpret these results to mean that, like constraint-based algorithms, and consistent with Danks (2004), students are using one cognitive strategy for detecting when two variables are adjacent, and another for detecting how the adjacencies are oriented, especially in the case of data collected from passive observation. Detecting whether X and Y are adjacent is as simple as detecting whether X and Y are independent conditional on any set. Detecting whether X – Y – Z is oriented as: X → Y ← Z or as one of: {X → Y → Z, X ← Y → Z , X ← Y ← Z } is much more difficult.

5. Conclusions

The pilot studies discussed here are suggestive, but still quite prelimary. Subjects had direct access to the independence data true of the population, and in several of our experiments the choices they confronted were limited. Nevertheless, these studies suggest that there is a lot to be learned from comparing naïve subjects to those trained even for a short time on the normative theory of Causal Bayes Networks. For whatever reason, trained subjects can reliably differentiate between direct and indirect causation, and many can do so with an optimal strategy for picking interventions. Indeed, our first experiment suggests that trained students are not subject to source bias in picking interventions, even though they were never trained in this particular skill. We speculate that simple training in the normative theory sensitizes subjects to the connection between conditional independence and indirect causation, and attending to the mediating variable, which is the conditioning variable, leads subjects to intervene on the mediator instead of the source. Our pilot studies also suggest that only minimal training in the normative theory is needed to exhibit sensitivity to model equivalence, a core idea in the normative theory. Finally, they suggest that students pursue a strategy by which they find which pairs of variables are adjacent and then attempt to find in which direction the causal relations obtain.

Strategies for automatically learning causal structures in the normative theory divide into "constraint-based" and "score-based" methods.[18] In constraint based methods, one decides on individual constraints, e.g., independence or conditional independence facts, in order to decide on local parts of the model, e.g., whether a given pair of variables are adjacent or not. In score based searches, one computes a score reflecting the goodness of

[18] For a detailed but accessible primer, read the chapter on score-based vs. constrain-based methods in Glymour and Cooper, 1999.

fit of the entire model. Human subjects, both naïve and trained, arguably execute a simple version of a constraint-based search. Our subjects used particular independence relations to decide on questions of adjacency, and were reliable at this, and then used interventions to decide on orientation for local fragments of the model, and were moderately reliable at this. None judged models as a whole and attempted to maximize some global score. As it turns out, constraint-based approaches are much more efficient, but less accurate in the face of noisy data. Our conjecture is that human subjects employ a constraint-based approach because it allows a sequence of decisions, each involving a potentially very simple computation, like whether two variables are independent or not. I

In systems of more than toy complexity, that is, systems involving more than two or three variables, a score-based strategy would become computationally prohibitive for a human cognitive agent, while a constraint-based approach would still be viable. Since a constraint-based approach also lends itself to an anytime approach, that is, using only the simplest constraints first and then stopping "anytime" the constraints under test become too complicated to compute or to trust statistically, it is also well suited to systems with severe computational or memory constraints, e.g., human learners.

Nevertheless, we do not claim that evolution has trained humans to execute anything like the theoretically correct version of a constraint-based search for causal structure. Even minimally trained subjects using a constraint-based approach well suited for toy systems but not theoretically correct might quickly be overcome by the complexity of a five variable system. In informal observation this is exactly what happens. Even on systems involving four variables, if subjects are given no background knowledge whatsoever about which variables are prior to which others, e.g., which variable is the "outcome" variable, then they become quickly lost in the more than fifty models in their search space. In future experiments, we will investigate the discontinuities in performance for trained subjects as a function of system complexity. We will train subjects to execute a modified version of a constraint-based approach that would handle much larger systems, and see if this will help students to become truly more reliable causal learners.

6. References

Berger, Martijn (2005). *Applied Optimal Designs*. Wiley.

Blalock, H. (1961). *Causal Inferences in Nonexperimental Research*. University of North Carolina Press, Chapel Hill, NC.

Bowden, R. and Turkington, D. (1984). *Instrumental variables*. Cambridge University Press, NY

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

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, 99, 365-382.

Cochran, W., and Cox, G. M. (1957). *Experimental Designs*, 2nd Edition. Wiley.

Danks, D. (2004). Constraint-based human causal learning. In M. Lovett, C. Schunn, C. Lebiere, & P. Munro, eds. *Proceedings of the 6th International Conference on Cognitive Modeling* (ICCM-2004). Mahwah, N.J.: Lawrence Erlbaum Associates. pp. 342-343.

Danks, D. (2005). Causal Learning from Observations and Manipulations, in Lovett, M., Shah, P. eds. *Thinking with Data*, Lawrence Erlbaum Associates.

Danks, D., Griffiths, T. L., & Tenenbaum, J. B. (2003). Dynamical causal learning. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 67-74). Cambridge, Mass.: MIT Press.

Eberhardt, F., Glymour, C., and Scheines, R. (2005). "N-1 Experiments Suffice to Determine the Causal Relations Among N Variables," *Technical Report No. CMU_PHIL-161*, Dept. of Philosophy, Carnegie Mellon University, Pittsburgh, PA, 15213

Glymour, C. (1998). Learning causes: psychological explanations of causal explanation. *Minds and Machines*, 8, 39-60.

Glymour, C. (2000). Bayes nets as psychological models. In F.C. Keil & R.A. Wilson (Eds.), *Explanation and cognition*. Cambridge, Mass.: The MIT Press.

Glymour, C., and Cooper, G. (1999). *Computation, Causation, and Discovery*. AAAI Press and MIT Press.

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, 3-32.

Gopnik, A., Sobel, D. M., Schulz, L. E., Glymour, C. (2001). Causal learning mechanisms in very young children: Two, three, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, 37, 620-629.

Griffiths, T.L., Baraff, E.R., & Tenenbaum, J.B. (2004). Using physical theories to infer hidden causal structure. *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society*

Koedinger, K. R., & Anderson, J. R. (1998). Illustrating principled design: The early evolution of a cognitive tutor for algebra symbolization. *Interactive Learning Environments*, 5, 161-180.

Lagnado, D., & Sloman, S.A. (2002). Learning causal structure. *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*, Maryland.

Lagnado, D., & Sloman, S.A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 856-876.

Murphy, K. (2001). Active learning of causal Bayes net structure. *Technical report*, Computer Science Division, University of California-Berkeley.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo.

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

Perales, J. C., & Shanks, D. R. (2003). Normative and descriptive accounts of the influence of power and contingency on causal judgement. *The Quarterly Journal of Experimental Psychology*, 56A, 977-1007.

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.

Simon, H. (1953). Causal ordering and identifiability. *Studies in Econometric Methods*. Hood and Koopmans (eds). 49-74.Wiley, NY.

Sloman, S. A., & Lagnado, D. (2002). Counterfactual undoing in deterministic causal reasoning. *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*, Maryland.

Sobel, D. M. & Kushnir, T. (2004). Do it, or watch it done: The importance of decision demands in causal learning from interventions. Manuscript submitted for publication, Brown University.

Spiegelhalter, D. and Lauritzen, S. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:579--605.

Spirtes, P., Glymour, C., Scheines R., (2000). *Causation, Prediction and Search, 2nd Edition*, MIT Press, Cambridge, MA.

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

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In T. Leen, T. Deitterich, & V. Tresp (Eds.), *Advances in Neural Information Processing 13* (pp. 59-65). Cambridge, Mass.: The MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal inference. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15* (pp. 35-42). Cambridge, Mass.: The MIT Press.

Tenenbaum, J. B. & Niyogi, S. (2003). Learning causal laws. In *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*.

Tong, S., & Koller, D. (2001). Active learning for structure in Bayesian networks. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Waldmann, M. R., & Hagmayer, Y. (in press). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Waldmann, M. R., & Martignon, L. (1998). A Bayesian network model of causal learning. In M.A. Gernsbacher & S.J. Derry (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Wright, S. (1934). The method of path coefficients. *Ann. Math. Stat*. 5, 161-215.