Structural and Probabilistic Causality

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
judea@cs.ucla.edu

Abstract

Competing theories of causation based on statistical and power accounts are assessed and related to the normative theories of probabilistic causality and structural modeling. Recent advances in graphical models enable us to cast discussions of these theories in precise mathematical language, thus clarifying their strengths and limitations. The inferential machinery that accompanies the graphical language provides solutions to a number of enduring problems of causal inference, including the analysis of actions, and the processing of counterfactual utterances.

1 Introduction

The central aim of many empirical studies in the physical, behavioral, social, and biological sciences is the elucidation of cause-effect relationships among variables. It is through cause-effect relationships that we obtain a sense of a "deep understanding" of a given phenomenon, and it is through such relationships that we obtain a sense of being "in control," namely, that we are able to shape the course of events by deliberate actions or policies. It is for these two reasons, understanding and control, that causal thinking is so pervasive, popping up in everything from everyday activities to high-level decision making: a car owner wonders why an engine won't start; a cigarette smoker would like to know, given his/her specific characteristics, to what degree his/her health would be affected by refraining from further smoking; a policy maker would like to know to what degree anti-smoking advertising would reduce health care costs; and so on. Although a plethora of data has been collected on cars and on smoking and health, the appropriate methodology for extracting answers to such questions from the data has been fiercely debated, partly because some fundamental questions of causality have not been given fully satisfactory answers.

The two fundamental questions of causality are:

1. What empirical evidence is required for legitimate inference of cause-effect relationships?

2. Given that we are willing to accept causal information about a certain phenomenon, what inferences can we draw from such information, and how?

The primary difficulty is that we do not have a clear empirical semantics for causality; statistics teaches us that causation cannot be defined in terms of statistical associations, while any philosophical analysis of causation in terms of deliberate control quickly reaches metaphysical dead-ends over the meaning of free will. Indeed, Bertrand Russell (1913) noted that causation plays no role in physics proper and offered to purge the word from the language of science. Karl Pearson (1911) advocated such a purge from statistics which, regretfully, has been more successful than that envisioned by Russell.

Philosophical difficulties notwithstanding, scientific disciplines that must depend on causal thinking have developed paradigms and methodologies that successfully bypass the unsettled questions of causation and that provide acceptable answers to pressing problems of experimentation and inference. Social scientists, for example, have adopted path analysis and structural equation models, and programs such as LISREL have become common tools in social science research. Econometricians, likewise, have settled for stochastic simultaneous equations models as carriers of causal information and have focused most of their efforts on developing statistical techniques for estimating the parameters of these models. Statisticians, in contrast, have adopted Fisher's randomized experiment as the ruling paradigm for causal inference, with occasional excursions into its precursor – the Neyman-Rubin model of potential response [Neyman, 1923, Rubin, 1974].

None of these paradigms and methodologies can serve as an adequate substitute for a comprehensive theory of causation, one suitable for explaining the ways people infer and process causal relationships. The structural equations model is based largely on informal modeling assumptions and has hardly been applied beyond the boundaries of linear equations with Gaussian noise. The statisticians' paradigm of randomized experiments is too restrictive in natural, pre-scientific learning environment, and it does not allow for the integration of statistical data with the rich body of (previously acquired) causal knowledge that is available in ordinary discourse. And philosophers have essentially abandoned the quest for the empirical basis of causation. Early attempts to reduce causality to probabilities got entangled in circular definitions (see Subsection 2.2) and recent theories, based on processes [Salmon, 1994] or capacities [Cartwright, 1989, chapter 4], though conceptually appealing, have not been formalized with sufficient precision to describe how people learn, represent, and use causality in ordinary practice.

A new perspective on the problem of causation has recently emerged from a rather unexpected direction – artificial intelligence (AI). When encoding and processing causal relationships on digital machines became necessary, the problems and assumptions that other disciplines could keep dormant and implicit had to be explicated in great detail, so as to meet the levels of precision necessary in programming.

Explicating cause-effect relationships has become a concern central to several areas of AI: natural language processing, automated diagnosis, robot planning, qualitative physics, and database updates. In the area of robotics, for example, the two fundamental problems of causation were translated into concrete, practical questions:

1. How should a robot acquire causal information through interaction with its environment?

2. How should a robot process the causal information it receives from its creator-programmer?

Attempts to gloss over difficulties with causation quickly result in a programmer's nightmare. For example, when given the information "If the grass is wet, then the sprinkler must have been on" and "If I break this bottle, the grass will get wet," the computer will conclude "If I break this bottle, the sprinkler must have been on." The swiftness and concreteness with which such bugs surface has forced computer scientists to pinpoint loosely stated assumptions and then assemble new and more coherent theories of actions, causation, and change.

The purpose of this paper is to summarize recent advances in causal reasoning, to show how they clarify, unify, and enrich previous approaches in philosophy, economics, and statistics, and to relate these advances to two models of causal judgment that have been proposed in the psychological literature: the statistical contingency model [Jenkins and Ward, 1965, Cheng, 1992] and the power-based model [Shultz, 1982]. The statistical contingency model and its variants are grounded in the philosophical literature of probabilistic causality, to be described and assessed in Section 2. Related advancements in probabilistic causal discovery, based on the language of graphs [Pearl and Verma, 1991, Spirtes *et al.*, 1993] are described in Section 3. It is shown that graphs offer a powerful language for formulating and resolving some of the fundamental problems in probabilistic causality and, in addition, that graphs offer techniques of extracting causal relationships from intricate patterns of probabilistic dependencies, including distinct patterns created by unobserved factors. The power-based model takes after the structural equations models used in econometrics, in which causal relationships are defined in terms of hypothetical manipulative experiments. A general, non-parametric formulation of structural equations models is given in Section 4, and is shown to support a wide variety of causal relationships, including predictive, abductive, manipulative and counterfactual modes of reasoning.

Integrated models, in which causal judgment is shaped by both statistical data and preconceived notions of power [Cheng *et al.*, 1995, this volume], are more closely related to an action calculus formulated in [Pearl, 1994]. In this formulation, prior causal knowledge is encoded qualitatively in the form of a graph containing both observed and unobserved variables, and the magnitudes of causal forces in the domain are inferred from both the probability of the observed variables and the topological features of the graph. The calculus (described in Section 4.4) admits two types of conditioning operators: ordinary Bayes conditioning, $P(y|X = x)$, which represents the observation $X = x$, and causal conditioning, $P(y|do(X = x))$, read: the probability of $Y = y$ conditioned on holding X constant (at x) by deliberate action. Given a mixture of such observational and causal sentences, together with the topology of the causal graph, the calculus derives new conditional probabilities of both types, thus enabling one to quantify the effects of actions and observations, to specify conditions under which manipulative experiments are not necessary, and to suggest additional observations or auxiliary experiments from which the desired inferences can be obtained.

I propose this formalism as a basis for theories of causal learning and, in particular, how humans integrate information from diverse sources – passive observation, manipulative experimentation, and linguistic instruction – to synthesize a coherent causal picture of the environment.

2 Probabilistic Causality

Probabilistic causality is a branch of philosophy that attempts to explicate causal relationships in terms of probabilistic relationships. This attempt is motivated by several ideas and expectations. First and foremost, probabilistic causality promises a solution to the centuriesold puzzle of causal discovery, that is, how humans discover genuine causal relationships from bare empirical observations, free of any causal preconceptions. Given the Humean dictum that causal knowledge originates with human experience and the (less compelling but currently fashionable) assumption that human experience is encoded in the form of a probability function, it is natural to expect that causal utterances might be reducible to a set of relationships in some probability distribution that is defined over the variables of interest. Second, in contrast to deterministic accounts of causation, probabilistic causality offers substantial cognitive economy. Physical states and physical laws need not be specified in minute detail because instead they can be summarized in the form of probabilistic relationships among macro-states so as to match the granularity of natural discourse. Third, probabilistic causality is equipped to deal with the modern (i.e., quantum theoretical) conception of uncertainty, according to which determinism is merely an epistemic fiction, and nondeterminism is the fundamental feature of physical reality.

The formal program of probabilistic causality owes its inception to H. Reichenbach (1956) and I.J. Good (1961) and has subsequently been pursued by P. Suppes (1970), B. Skyrms (1980), R. Otte (1981), W. Spohn (1980), W.C. Salmon (1984), N. Cartwright (1989), and E. Eells (1991). The current state of this program is rather disappointing considering its original aspirations. Salmon has abandoned the effort altogether, concluding that "causal relations are not appropriately analyzable in terms of statistical relevance relations" [Salmon, 1984, page 185]; instead, he has proposed an analysis in which "causal processes" are the basic building blocks. More recent accounts by Cartwright and Eells have resolved some of the difficulties encountered by Salmon, but at the price of, on the one hand, complicating the theory beyond recognition and, on the other, compromising its original goals. The following is a brief account of the major achievements and difficulties of probabilistic causality, as elaborated in Cartwright (1989) and Eells (1991).

2.1 Temporal ordering

Standard probabilistic accounts of causality assume that, in addition to a probability function P, we are also given the temporal order of the variables in the analysis. This is understandable, considering that causality is an asymmetric relation, while statistical relevance is symmetric. Lacking temporal information, it would be impossible, for example, to decide which of two dependent variables is the cause and which the effect, since every joint distribution $P(x, y)$ induced by a model in which X is a cause of Y can also be induced by a model in which Y is the cause of X. Thus, any method of inferring that X is a cause of Y must also infer, by symmetry, that Y is a cause of X. By imposing the constraint that an effect never precede its cause, the symmetry is broken and causal inference can commence.

The reliance on temporal information has its price though, as it excludes a priori the analysis of cases in which the temporal order is not well defined, either because processes overlap in time or because they (appear to) occur instantaneously. For example, one must give up the prospect of determining (by uncontrolled methods) whether sustained physical exercise contributes to low cholesterol levels or, the other way around, low cholesterol levels enhance the urge to engage in physical exercise. Likewise, the philosophical theory of probabilistic causality would not attempt to distinguish between the claims "tall flag poles cause long shadows" and "long shadows cause tall flag poles," in which, for all practical purposes, the putative causes and effects occur simultaneously.

We shall see when we discuss graphical methods that some determination of causal directionality can be made from atemporal statistical information, albeit with a weakened set of guarantees.

2.2 Circularity

Despite the reliance on temporal precedence, the criteria that philosophers have devised for identifying causal relations suffer from glaring circularity: In order to determine whether an event C is a cause of event E, one must know in advance how other factors are causally related to C and E. Such circularity emerges from the need to define the "background context" under which a causal relation is evaluated, since the intuitive idea that causes should increase the probability of their effects must be qualified by the condition that other things are assumed equal. For example, striking a match is a cause for fire, but only when oxygen is present, when the match is dry, and so on. Thus, it seems natural to define

Definition 1 *C is causally relevant to E if there is at least one condition F in some background context K such that $P(E|C,F) > P(E|\neg C,F)$.*

But what kind of conditions should we include in the background context? On the one hand, insisting on a complete description of the physical environment would reduce probabilistic causality to deterministic physics (barring quantum-level considerations). On the other hand, ignoring background factors altogether, or describing them too coarsely, would introduce spurious correlations and other confounding effects. A natural compromise is to require that the background context itself be "causally relevant" to the variables in question, a move that is the source of circularity in the definition of statistical causality.

The dangers of describing the background too coarsely will be illustrated via two examples, one using the celebrated Simpson's paradox, the other the issue of interactive factors.

Simpson's paradox [Simpson, 1951], first encountered by Pearson in 1899 [Aldrich, 1994], refers to the phenomenon whereby an event C seems to increase the probability of E in a given population p and, at the same time, decrease the probability of E in every subpopulation of p. In other words, if F and $\neg F$ are two complementary events describing two subpopulations, we might well encounter the inequalities

$$P(E|C) > P(E|\neg C) \tag{1}$$

$$P(E|C,F) < P(E|\neg C,F) \tag{2}$$

$$P(E|C,\neg F) < P(E|\neg C,\neg F) \tag{3}$$

While such order reversal might not surprise students of probability, it may become paradoxical when given a causal interpretation. For example, if we associate C with taking a certain drug, E with recovery, and F with being a female, under the causal interpretation of Eqs. (1)-(3) the drug would be harmful to both males and females and beneficial to the population as a whole. Intuition deems such a result impossible, and correctly so.

The explanation for Simpson's paradox is that the inequality

$$P(E|C) > P(E|\neg C)$$

is interpreted erroneously. It is not a statement about C being a positive causal factor for E because the inequality may be due to spurious confounding factors that may cause both C

and E. In our example, for instance, the drug may appear beneficial on the average because the women, who recover (despite the drug) more often than the men, are also more likely than the men to use the drug.

The standard method for dealing with potential confounders of this kind is to "hold them fixed,"[1] namely, to condition the probabilities on any factor that might cause both C and E. In our example, if being a female (F) is perceived to be a cause for both recovery (E) and drug usage (C), then the effect of the drug needs to be evaluated separately for men and women (as in Eqs. (2)-(3)) and averaged accordingly.

Here we see the emergence of circularity: In order to determine the causal role of C relative to E (e.g., the effect of the drug on recovery), we must first determine the causal role of every factor F (e.g., gender) relative to C and E. More crucial, we must make sure that C is not causally relevant to F, or else no C would ever qualify as a cause of E, because we can always find factors F that are intermediaries between C and E which screen off E from C.[2]

Factors affecting both C and E can be rescued from circularity by conditioning on *all* factors preceding C but, unfortunately, other factors that cannot be identified through temporal ordering alone must also be weighed. Consider the following example. I must bet heads or tails on the outcome of a fair coin toss; I win if I guess correctly, lose if I don't. Naturally, once the coin is tossed (and while the outcome is still unknown), the bet is deemed causally relevant to winning, even though the probability of winning is the same whether I bet heads or tails. To reveal the causal relevance of the bet (C), we must include the outcome of the coin (F) in the background context, even though F does not meet the common-cause criterion – it does not affect my bet (C) nor is it causally relevant to winning (E) (unless we first proclaim the bet relevant to winning). Worse yet, we cannot justify including F in the background context by virtue of its occurring earlier than C because whether the coin is tossed before or after my bet is totally irrelevant to the problem at hand. We conclude that temporal precedence alone is insufficient for identifying the background context, and we must refine the definition of the background context to include what Eells (1991) calls "interacting causes," namely, (simplified) factors F that (i) are not affected causally by C and (ii) jointly with C (or $\neg C$) increase the probability of E.

Due to the circularity inherent in all definitions of causal relevance, probabilistic causality cannot be regarded as a program for extracting causal relations from temporal-probabilistic information but, rather, as a program for validating whether a proposed set of causal relationships is consistent with the available temporal-probabilistic information. More formally, suppose someone gives us a probability distribution P and a temporal order O on a (complete) set of variables V. Furthermore, any pair of variable sets, X and Y, in V is annotated by a symbol R or I, where R stands for "causally relevant" and I for "causally irrelevant." Probabilistic causality deals with testing whether the proposed R and I labels are consistent with the pair $< P, O >$ and that cause should both precede and increase the probability of

[1]The phrases "hold F fixed" or "control for F," used by both philosophers (e.g., [Eells, 1991]) and statisticians (e.g., [Pratt and Schlaifer, 1988]), connote external interventions and may, therefore, be misleading (see later sections on acting vs. seeing). In standard probability language, all one can do is to simulate "holding F fixed" by considering cases with equal values of F, namely, "conditioning" on F and $\neg F$, an operation I will call "adjusting for F."

[2]F "screens off" E from C if C and E are conditionally independent, given both F and $\neg F$; equivalently, if equalities hold in Eqs. (2) and (3).

effect.

Currently, the most advanced consistency test is the one based on Eells' criterion of relevance [Eells, 1991], which translates into:

Consistency test: For each pair of variables labeled $R(X, Y)$, test whether

(i) X precedes Y in O, and

(ii) there exist x, x', y such that $P(y|x, z) > P(y|x', z)$ for some z in Z, where Z is a set of variables in the background context K, such that $I(X, Z)$ and $R(Z, Y)$.

This now raises additional questions:

A. Is there a consistent label for every pair $< P, O >$?

B. When is the label unique?

C. Is there a procedure for finding a consistent label when it exists?

While some insights into these questions are provided by the graphical methods to be discussed in Section 3, the point to notice is that, due to circularity, the mission of probabilistic causality has been altered: from discovery to that of consistency testing.

2.3 The closed-world assumption

By far the most critical and least defensible paradigm underlying probabilistic causality rests on the assumption that a probability function exists on all variables relevant to a given domain of discourse. This assumption absolves the analyst from worrying about unmeasured spurious causes, which might (physically) affect several variables in the analysis and still remain obscure to the analyst. It is well known that the presence of such "confounders" may reverse or negate any causal conclusion that might be drawn from probabilities. For example, observers might conclude that "bad air" is the cause of malaria if they are not aware of the role of mosquitoes, or that falling barometers are the cause of rain, or that speeding to work is the cause of being late to work, and so on. Because they are unmeasured, or even unsuspected, the confounding factors in such examples cannot be neutralized by conditioning or by "holding them fixed." Thus, taking Hume's program of extracting causal information from raw data seriously entails coping with the problem that the validity of any such information is predicated on the untestable assumption that all relevant factors have been accounted for.

Similar problems affect psychological theories that use statistical relevance to explain how children extract causal information from experience. The proponents of such theories cannot ignore the fact that the child never operates in a closed, isolated environment. Unnoticed external conditions govern the operation of every learning environment, and these conditions often have the potential to confound cause and effect in unexpected and clandestine ways.

Fortunately, that children do not grow in closed, sterile environments like those in statistical textbooks has its advantages too. Aside from passive observations, a child possesses two valuable sources of causal information which are not available to the ordinary statistician: manipulative experimentation and linguistic advice. Manipulation subjugates the putative causal event to the sole influence of a known mechanism, thus overruling the influence of uncontrolled factors which might also produce the putative effect. "The beauty of independent manipulation is, of course, that other factors can be kept constant without their being identified" [Cheng, 1992]. The independence is accomplished by subjecting the object of interest to the whims of one's volition, to ensure that the manipulation is not influenced by any environmental factor likely to produce the putative effect. Thus, for example, a child can infer that shaking a toy can produce a rattling sound, because it is the child's hand, governed solely by the child's volition, that brings about the shaking of the toy and the subsequent rattling sound. The whimsical nature of free manipulation replaces the statistical notion of randomized experimentation and serves to filter sounds produced by the child's actions from those produced by uncontrolled environmental factors.

But manipulative experimentation cannot explain all of the causal knowledge that humans acquire and possess, simply because most variables in our environment are not subject to direct manipulation. The second valuable source of causal knowledge is linguistic advice, namely, explicit causal sentences about the workings of things which we obtain from parents, friends, teachers, and books, and which encodes manipulative experience of past generations. As obvious and uninteresting as this source of causal information might appear, it probably accounts for the bulk of our causal knowledge, and understanding how this transference of knowledge works is far from trivial. In order to comprehend and absorb causal sentences such as "The glass broke because you pushed it," the child must already possess a causal schema within which such inputs make sense. To further infer that pushing the glass will make someone angry at you and not at your brother, even though he was responsible for all previous breakage, requires a truly sophisticated inferential machinery. In most children, this machinery is probably innate.

Note, however, that linguistic input is by and large qualitative; we rarely hear parents explaining to children that placing the glass at the edge of the table increases the probability of breakage by a factor of 2. Yet, quantitative assessments of the effects of one's actions must be made in any decision-making situation, and the question arises, How does one combine quantitative empirical data with qualitative causal relations to deduce quantitative causal assessments? The problem is especially critical in situations in which empirical data is available on only a small part of the causal field, while the bulk of that field is represented as rudimentary statements of what affects what in the domain. By analogy, this resembles the task of figuring out how to fix a TV set when given only a general understanding of the principles of television electronics combined with empirical data on five knobs and one screen. This problem will be dealt with in Section 4.

2.4 Singular vs. general causes

Wayne A. Davis (1988, page 145) summarizes the distinction between singular and general causes as follows:

> A *general causal statement*, like "Drinking hemlock causes death," asserts that one general event causes another. A *singular causal statement*, like "Socrates' drinking hemlock caused his death," asserts that one singular event caused another. The relationship between singular and general causation is not simple. From the fact that being poisoned causes death, we cannot infer that Alan's being poisoned caused his death (he might have died of a bullet wound first). And

> even though Jim Fixx's last run caused his death, it is too strong to say that going for a a run causes death.

The account of probabilistic causality provided so far (Subsections 2.1–2.3) addresses only general causal statements. Whether probabilistic information suffices for asserting singular causal statements, and where knowledge about singular causes comes from if it doesn't, further exacerbates the problems of probabilistic causality.[3] The next example demonstrates that singular causes require knowledge in the form of counterfactual or functional relationships. Such knowledge is not needed for general causes, nor can it be extracted from bare statistical data even under controlled experimentation. It requires a higher level of inductive generalization, one capable of extracting temporal invariants.

> My son Danny feeds the dog whenever I ask him to, with a few exceptions. Ten percent of the time he feeds the dog even when I do not ask him to, and 10% of the time he does not feed the dog even when asked to. Today I asked Danny to feed the dog, which he did, and I wonder, Did he do it *because* I asked him to or was he about to do it anyway?

Let C and E stand for "asking" and "feeding," respectively. The story above can be summarized by two conditional probability statements,

$$P(E|C) = 0.90 \quad P(\neg E|\neg C) = 0.90 \tag{4}$$

which, together with the prior probability $P(C)$, fully specify the joint probability on the variables in question. Moreover, we can safely assume that C is the only relevant cause of E in the story, and that C and E are not confounded by any hidden common cause, so the same probabilities would prevail if C (vs. $\neg C$) were chosen by randomized experiment, hence the outcome associated with interventions or decisions is likewise determined by Eq. (4). For example, the probability that Danny will feed the dog tomorrow if I decide to ask him to is unequivocally 0.90.

Still, whether today's request was the *actual cause* for today's feeding is difficult, in fact impossible, to determine, given the information at hand. The difficulty stems from the ambiguity concerning the *mechanism* underlying Danny's occasionally abnormal behavior. We will show two alternative mechanisms, both compatible with the probabilistic behavior of Eq. (4), yet each giving a different answer to the singular causal query, "Was my asking the *actual cause* of today's feeding?" (equivalently: "Would E have been true had C been false?").

Consider two competing models:

A. 20% of the time, Danny is in an absent-minded trance; he would feed the dog at random with 50% probability, regardless of whether he was asked to.

B. 10% of the time, Danny is in a rebellious mood; he would feed the dog if he were not asked to, and would not feed the dog if he were asked to do so.

[3] Eells's (1991, chapter 6) analysis of token-level causation and Cartwright's (1989, chapter 3) argument for "singular causes first" (rejected by Eells) both presuppose knowledge of how the occurrence of one singular event raises the probability of another, and thus only beg the question of where that extra knowledge comes from and how it is encoded in the mind.

It is easy to see that Models A and B are both compatible with the probabilistic information given in Eq. (4), while they differ on the counterfactual query. In Model B, Danny's feeding the dog today rules out the possibility that he is in a rebellious mood; hence, he would not have fed the dog if not asked, and we can rest assured that my asking was the *actual cause* of today's feeding. In Model A, however, today's feeding still leaves uncertain whether Danny is in an alert state of mind or in one of those absent-minded trances (giving a 8:1 chance to each possibility). If alert, Danny would not have fed the dog had I not asked him to; if in a trance, there is a 50% chance that he would still have fed the dog. Thus, the probability that my asking *actually caused* today's feeding is 100% in Model B, and less than 100% (8/9) in Model A.

We see now that probabilistic information, even enriched with information about temporal ordering and causal relevance, is insufficient for answering counterfactual queries; the task requires the specification of the *functional* relationship between the putative cause and the putative effect.

This deficiency of the probabilistic account cannot be dismissed as metaphysical, that is, on the grounds that counterfactual sentences are, by definition, empirically untestable, hence meaningless. Counterfactual statements do in fact have an empirical content, but only when coupled with assumptions of persistence (perhaps this is what Hume meant by "regularity"). For example, assume that Danny's state of being in an abnormal mood persists for not one but several days. Our counterfactual query would then translate into a sharp empirical question of whether we can count on the dog being fed tomorrow. In fact, the ingredient that makes counterfactual probabilities hard to compute is not the counterfactual phrasing of the query but rather the fact that the query is accompanied with information that renders the event in question unique, unlike any other event summarized in the probability distribution P. Before we find out that Danny in fact fed the dog today, we have no problem answering the counterfactual query, "Would he feed the dog if he were not asked to?" The difficulty stems from observing Danny feeding the dog today, thus making this day singular, unlike any other day summarized in P. In other words, Danny was not (and can never be) observed under both conditions, being asked and not being asked, on the very same day. Thus, it is only when we observe the persistence of some mechanism (Danny's abnormality) for several successive observations that we can substantiate a counterfactual claim, and it is due to such persistence that counterfactual statements acquire their empirical content and their unique role in planning and knowledge communication. We shall see in Section 4.5 that counterfactual knowledge is essential for predicting the effect of actions when measurements are available about conditions that are likely to be affected by those actions.

Proponents of probabilistic causality may argue that by introducing new hypothetical variables into the analysis and stretching the notion of a "factor" to include counterfactual strategies, singular causes can still be treated in the probabilistic framework. For example, in the situation above, we could introduce Danny's "mood" or "mode of behavior" as a factor in the background context and, by conditioning the outcome E on this new factor, the correct answer would obtain using ordinary probabilistic computations. In general, accessing the degree to which an event C *actually causes* an observed event E involves considering as factors each of the four possible functions from $\{C, \neg C\}$ to $\{E, \neg E\}$, for which the term "mood" is merely indexical.[4] However, introducing these new factors seems like a roundabout way of squeezing meta-probabilistic causal and counterfactual information into the probabilistic vocabulary, and it is a far cry from Hume's program of inferring causes from probabilities, because there is no way to distinguish Model A from Model B solely on the basis of statistical observations without either going into a deeper analysis of Danny's state of mind or assuming that whatever mood Danny is in persists unaltered for at least few trials. Such specification is accomplished more naturally in the structural equations framework, to be described in Section 4.

3 The Language of Causal Graphs

Causal graphs appear sporadically in the writings of Simon, Reichenbach, Cartwright, and Eells, where they are used primarily for mnemonic or display purposes. The use of graphs as a formal mathematical language for defining and processing causal relationships is relatively recent. We shall see that graphs offer a powerful language for expressing and resolving some fundamental questions in probabilistic causality, as well as a plausible hypothesis of how causal relationships are organized in the human mind.

3.1 Direct causes and Bayesian networks

A convenient starting point for introducing causal graphs is through the notion of *Markovian parents.*

Definition 2 *Let $V = \{X_1, \ldots, X_n\}$ be an ordered set of variables, and let $P(v)$ be the joint probability distribution on these variables. A set of variables PA_j is said to be Markovian parents of X_j if PA_j is a minimal set of predecessors of X_i that renders X_j independent of all its other predecessors. In other words, PA_j is any subset of $\{X_1, \ldots, X_{j-1}\}$ satisfying*

$$P(x_j|pa_j) = P(x_j|x_1, \ldots, x_{j-1}) \tag{5}$$

such that no proper subset of PA_j satisfies Eq. (5).

Definition 2 assigns to each variable X_j a select set of parent variables which are sufficient for determining the probability of X_j; knowing the values of other preceding variables is redundant once we know the values pa_j of the parent set PA_j. This assignment can be represented in a form of a directed acyclic graph (DAG) in which variables are represented by nodes and where arrows are drawn from each node of the parent set PA_j toward the child node X_j. Definition 2 also suggests a simple recursive method of constructing such a DAG: Starting with the pair (X_1, X_2), we draw an arrow from X_1 to X_2 if the two variables are dependent. Continuing to X_3, we draw no arrow in case X_3 is independent of $\{X_1, X_2\}$; otherwise, we examine whether X_2 screens off X_3 from X_1 or X_1 screens off X_3 from X_2. In the first case, we draw an arrow from X_2 to X_3; in the second, we draw an arrow from X_1 to X_3. If no screening condition is found, we draw arrows to X_3 from both X_1 and X_2. In general, at the ith stage of the construction, we select any minimal set of X_i's predecessors

[4] Many more functions need be considered in cases where C interacts with other factors of E (see[Balke and Pearl, 1994]).

that shield X_i from its other predecessors (as in Definition 2), call this set PA_i (connoting "parents"), and draw an arrow from each member in PA_i to X_i. The result is a directed acyclic graph, called a "Bayesian network" in [Pearl, 1988], in which an arrow from X_i to X_j assigns X_i as a Markovian parent of X_j, consistent with Definition 2.

Figure 1 illustrates a simple yet typical Bayesian network. It describes relationships among the season of the year (X_1), whether rain falls (X_2) during the season, whether the sprinkler is on (X_3) during that season, whether the pavement would get wet (X_4), and whether the pavement would be slippery (X_5). All variables in this figure are binary, taking a value of either true or false, except the root variable X_1, which can take one of four values: Spring, Summer, Fall, or Winter. The network was constructed in accordance with Definition 2, using causal intuition as guide. The absence of a direct link between X_1 and X_5, for example, captures our understanding that the influence of seasonal variations on the slipperiness of the pavement is mediated by other conditions (e.g., the wetness of the pavement). This intuition coincides with the independence condition of Eq. (5), since knowing X_4 renders X_5 independent of $\{X_1, X_2, X_3\}$.

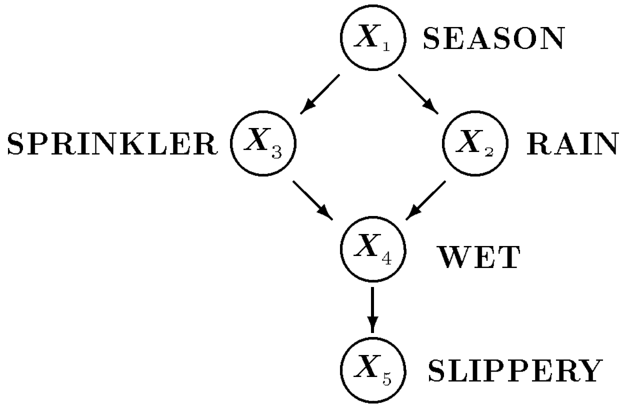

Figure 1: A Bayesian network representing causal influences among five variables.

How do graphs enter our discussion of causality? To see the connection, we need to make three additional steps. First, we identify the ordering of the variables with their temporal order. Second, we make the closed-world assumption, namely, that $V = X_1, ..., X_n$ include *all* relevant variables for the phenomenon under study. Finally, we make a smooth transition from events to variables as the basic objects of causal relationships. This will enable us to say, for example, that force causes (or influences) acceleration, without specifying precisely what magnitude of force (an event) accounts for what level of acceleration (an event).[5]

With these provisions in mind, it is natural to identify the Markovian parents PA_i as "direct causes" of X_i; "causes", because they exhibit the temporal-probabilistic features of causal relevance described in Section 2, and "direct" because they are not mediated (or screened off) by any other group of variables, especially when the parent set is unique.

Definition 3 (direct causes) *Let $V = \{X_1, \ldots, X_n\}$ be a complete set of temporally ordered variables, and let $P(v)$ be the joint probability distribution on these variables. We say that X_i is a direct cause of X_j if X_i is a member of the parent set PA_j in a Bayesian network of $P(v)$ constructed along the temporal order.*

[5][Spohn, 1980, Mulaik, 1986] are among the few who advocated this transition in the philosophical literature, though it has been used routinely in path analysis [Wright, 1921] economics [Simon, 1953] and artificial intelligence [Kim and Pearl, 1983]

Definition 3 provides a natural generalization of deterministic causality, in the spirit of Mulaik, (1986). If deterministic causes are defined as a set of conditions sufficient for determining the value of X_j, regardless of other eventualities, then Definition 3 merely substitutes probability determination for value determination: once the direct causes of X_j are known, the probability of X_j is completely determined; no other event preceding X_j would modify this probability. In Section 3.2 we will see that this invariance is in fact stronger; no other event except consequences of X_j would modify the probability of X_j. However, this invariance still falls short of the absolute invariance induced by deterministic causes, where the value of X_j remains determined against both past and future eventualities. A probabilistic version of such absolute invariance will be achieved through the manipulative and counterfactual accounts of causation, to be discussed in Section 4.

Definition 3 is also compatible with that of Eells (Section 3.2), which was based on causal relevance among events. A direct cause X_i of X_j must contain an event $X_i = x_i$ that is causally relevant to at least one event $X_j = x_j$ (i.e., $P(x_j|x_i, F) \neq P(x_j|x_i', F)$ for at least two values, x_i and x_i' of X_i) because, otherwise, there would be some set of factors F which screens off X_j from X_i, thus violating the minimality of PA_j. Conversely, if any event $X_i = x_i$ is deemed causally relevant to event $X_j = x_j$, then X_i must be either a direct cause of X_j or causally relevant to some direct cause of X_j, because if X_i satisfies neither of these possibilities, it is not an ancestor of X_j in the graph; X_i would then be screened off from X_j by some of its X_i's predecessors, and that would imply that X_i is not causally relevant to X_j after all.

However, the main advantage of commencing discussion of causality with the notion of direct causes is that the problem of circularity disappears (since each parent set is assigned independently of the others), the questions of consistency and uniqueness are resolved, and, not the least important, Definition 3 invites the language of graphs, with the help of which much harder questions of causality can be formulated and resolved.

Consider first the question of consistency. Assume we are given the pair $< P, O >$ as before and we wish to find a consistent labeling (D) on pairs of variables, such that a pair (X_i, X_j) is labeled D iff "X_i is a direct cause of X_j" in accordance with Definition 3. It is a simple matter to find such a labeling by constructing a Bayesian network along O, and associating the labels with the links of the resulting DAG. Thus, the question of consistency is answered in the affirmative; every pair $< P, O >$ has a labeling consistent with Definition 3, as given by the links of the constructed graph. The question of uniqueness also has a simple solution; if $P(v) > 0$ for every configuration v, then the parent sets PA_i are unique [Pearl, 1988, page 119] and, hence, there will be a unique set of direct causes for every variable. Causal ambiguities emerge when some configurations obtain zero probability, representing deterministic constraints. For example, a chain $X_1 \rightarrow X_2 \rightarrow X_3$ of necessary and sufficient causes cannot be distinguished from a fork $X_2 \leftarrow X_1 \rightarrow X_3$ of necessary and sufficient causes, because $\{X_2\}$ and $\{X_1\}$ each is sufficient for determining $\{X_3\}$, hence, each can serve as a Markovian parent of X_3. This is precisely the ambiguity noticed in probabilistic causality [Otte, 1981]: even given complete specification of temporal ordering, probabilistic information fails to distinguish genuine from spurious causes when causal connections degenerate into deterministic, necessary and sufficient relationships.[6] Definition 3

[6] We will argue later, in discussing the structural definition of causation, that neither causal chains nor causal forks can consist of strictly necessary and sufficient causes, because the meaning of the sentence "X_2 is a cause of X_3" rests in the claim that manipulating X_2, independently of events preceding X_2, would

confines the occurrence of such ambiguities to cases where deterministic constraints permit multiple minimal parent sets in the construction of the network.

An immediate beneficiary of graph language is the simplification and clarification of the notion of *background context* (sometimes called *causal field*) namely, the set K of variables which one should assume constant in assessing the causal relevance of one variable to another (see Definition 1). Subsection 2.2 summarizes the difficulties which philosophers have encountered in defining the appropriate background context for such assessment. The graphical concepts established by Definition 2 permit a constructive, noncircular definition of K, as follows:

Background context: In assessing the causal role of X relative to Y, the appropriate background context consists of all variables which are

1. direct parents of Y or of any intermediate variable between X and Y, and

2. nondescendants of X.

Since the notions of *parents*, *intermediate* and *descendants* are defined unambiguously in the graph, and the graph is defined constructively from the pair $< P, O >$, the background context, likewise, is well defined. Thus, one can now test systematically whether any event $X = x$ is a positive, negative or a mixed cause of another event $Y = y$, by constructing the Bayesian network, identifying the variables in K and, finally, comparing the probabilities $P(y|x, F)$ and $P(y|x', F)$ for each realization F of K.[7] For example, in assessing the causal role of having the sprinkler on ($X_3 = ON$) on having a slippery pavement ($X_5 = true$) in Figure 1, the relevant background context consists of a single variable: $Rain$ (X_2), and one needs to compare the quantities:

$$P(X_5 = true|X_3 = ON, X_2 = true) \text{ vs. } P(X_5 = true|X_3 = OFF, X_2 = true)$$

and,

$$P(X_5 = true|X_3 = ON, X_2 = false) \text{ vs. } P(X_5 = true|X_3 = OFF, X_2 = false)$$

Since inequality holds in the first pair and equality in the second, we conclude that $Sprinkler = ON$ is a positive cause of $Slippery$. We need not (and, in fact, should not) adjust for X_4 because it is a descendant of X_2. There is no harm, however, in including additional variables (such as X_1) in K, as long as they are nondescendants of X_2.

We will see later (Section 4) that, the variables in K possess a unique feature; it does not matter if we "hold K fixed" by external intervention or we "condition on K being constant." Although the two interpretations are generally not equivalent, they yield the same result for the relation between X and Y whenever K is selected by the graphical criterion above. This might explain why philosophers and statisticians, who generally ignore the distinction between "fixing" and "conditioning" (see footnote 1), often manage to escape the paradoxical consequences that such confusion may produce.

change X_3; the very existence of such manipulation rules out X_1 as being necessary and sufficient for X_2.

[7] If variable X is not an ancestor of variable Y then, clearly, event $X = x$ must be causally irrelevant to event $Y = y$. If X is an ancestor of Y, then $X = x$ may still be causally irrelevant to $Y = y$, since the causal relevance between X and Y shown in the graph may be due to other states of X and Y.

3.2 Implied independencies and observational equivalence

The construction implied by Definition 2 defines a Bayesian network as a carrier of conditional independence information relative to a specific temporal order O. Since temporal information is not always available (see Section 2.1) and since variable ordering, in general, is a meta-probabilistic notion, one may ask whether the independence information conveyed by the graph can be communicated without making an explicit reference to the ordering O. This information would then impose constraints on the possible ordering of the variables, and would open the possibility of inferring, or ruling out, causal relations from P alone.

Assume that a Bayesian network G was constructed from a probability distribution P and ordering O. It is interesting to ask what features of P characterize all those distributions that are capable, under some ordering of the variables, to produce a Bayesian network identical to G. To answer this question, we recall that the essential property of P used in the construction of G was Eq. (5), and that every distribution satisfying Eq. (5) can be decomposed (using the chain rule of probability calculus) into the product

$$P(x_1, ..., x_n) = \prod_i P(x_i \mid pa_i) \tag{6}$$

where pa_i are the values of the parents (PA_i) of X_i in G. For example, the DAG in Figure 1 induces the decomposition

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)\ P(x_2|x_1)\ P(x_3|x_1)\ P(x_4|x_2, x_3)\ P(x_5|x_4) \tag{7}$$

The product decomposition in Eq. (6) is no longer order-specific since, given P and G, we can test whether P decomposes into the product given by Eq. (6) without making any reference to variable ordering. Moreover, for every distribution decomposed as Eq. (6) one can find an ordering O that would produce G as a Bayesian network. We therefore conclude that a necessary and sufficient condition for a probability distribution P to induce a DAG G is that P admits the product decomposition dictated by G, as given in Eq.(6). If P satisfies this condition, we say that G *represents* P.

A convenient way of characterizing the set of distributions represented by a DAG G is to list the set of (conditional) independencies that each such distribution must satisfy. These independencies can be read off the DAG by using a graphical criterion called d-separation [Pearl, 1988]. To test whether X is independent of Y given Z in the distributions represented by G, we need to examine G and test whether the nodes corresponding to variables Z d-separate all paths from nodes in X to nodes in Y. By *path* we mean a sequence of consecutive edges (of any directionality) in the DAG.

Definition 4 (d-separation) *A path p is said to be d-separated (or blocked) by a set of nodes Z iff:*

(i) p contains a chain $i \longrightarrow j \longrightarrow k$ or a fork $i \longleftarrow j \longrightarrow k$ such that the middle node j is in Z, or,

(ii) p contains an inverted fork $i \longrightarrow j \longleftarrow k$ such that neither the middle node j nor any of its descendants (in G) are in Z.

If X, Y, and Z are three disjoint subsets of nodes in a DAG G, then Z is said to d-separate X from Y, denoted $(X \perp\!\!\!\perp Y|Z)_G$, iff Z d-separates every path from a node in X to a node in Y.

The intuition behind d-separation is simple: In chains $X \rightarrow Z \rightarrow Y$ and forks $X \leftarrow Z \rightarrow Y$, the two extreme variables are dependent (marginally) but become independent of each other (i.e., blocked) once we know the middle variable. Inverted forks $X \rightarrow Z \leftarrow Y$ act the opposite way; the two extreme variables are independent (marginally) and become dependent (i.e., unblocked) once the value of the middle variable (i.e., the common effect) or any of its descendants is known. For example, finding that the pavement is wet or slippery (see Figure 1) renders Rain and Sprinkler dependent, because refuting one of these explanations increases the probability of the other.

In Figure 1, for example, $X = \{X_2\}$ and $Y = \{X_3\}$ are d-separated by $Z = \{X_1\}$; the path $X_2 \leftarrow X_1 \rightarrow X_3$ is blocked by $X_1 \in Z$, while the path $X_2 \rightarrow X_4 \leftarrow X_3$ is blocked because X_4 and all its descendants are outside Z. Thus $(X_2 \perp\!\!\!\perp X_3|X_1)_G$ holds in G. However, X and Y are not d-separated by $Z' = \{X_1, X_5\}$, because the path $X_2 \rightarrow X_4 \leftarrow X_3$ is unblocked by virtue of X_5, a descendant of X_4, being in Z'. Consequently, $(X_2 \perp\!\!\!\perp X_3|\{X_1, X_5\})_G$ does not hold; in words, learning the value of the consequence X_5 renders its causes X_2 and X_3 dependent, as if a pathway were opened along the arrows converging at X_4.

Theorem 1 [Verma and Pearl, 1990, Geiger *et al.*, 1990]. *For any three disjoint subsets of nodes (X, Y, Z) in a DAG G, Z d-separates X from Y in G if and only if X is independent of Y conditional on Z in every distribution represented by G.*

The d-separation criterion can be tested in time linear in the number of edges in G. Thus, a DAG can be viewed as an efficient scheme for representing Markovian independence assumptions and for deducing and displaying all the logical consequences of such assumptions.

Note that the ordering with which the graph was constructed does not enter into the d-separation criterion; it is only the topology of the resulting graph that determines the set of independencies that the probability P must satisfy. Indeed, the following theorem can be proven [Pearl, 1988, page 120].

Theorem 2 *If a Bayesian network G is constructed recursively along some ordering O (as in Definition 2), then a construction along any ordering O' consistent with the direction of arrows in G would yield the same network. Consequently, any variable in a Bayesian network is independent of all its nondescendants, conditional on its parents.*

An important property that follows from the d-separation characterization is a criterion for determining whether two given DAGs are observationally equivalent, that is, whether every probability distribution that is represented by one of the DAGs is also represented by the other.

Theorem 3 [Verma and Pearl, 1990] *Two DAGs are observationally equivalent iff they have the same sets of edges and the same sets of v-structures, that is, two converging arrows whose tails are not connected by an arrow.*

Observational equivalence places a limit on our ability to infer causal directionality from probabilities alone. Two networks that are observationally equivalent cannot be distinguished without resorting to manipulative experimentation or temporal information. For example, reversing the direction of the arrow between X_1 and X_2 in Figure 1 does not introduce any new v-structure. Therefore, this reversal yields an observationally equivalent network, and the directionality of the link $X_1 \rightarrow X_2$ cannot be determined from probabilistic information. The arrows $X_2 \rightarrow X_4$ and $X_4 \rightarrow X_5$, however, are of different nature; there is no way of reversing their directionality without creating a new v-structure. Thus, we see that some probability functions P (such as the one responsible for the construction of the Bayesian network in Figure 1), unaccompanied by temporal information, can constrain the directionality of some arrows, and hence the directionality of the causal relationships among the corresponding variables. The precise meaning of such directionality constraints will be discussed in the next subsection.

Additional properties of DAGs and their applications to evidential reasoning are discussed in [Geiger, 1990, Lauritzen and Spiegelhalter, 1988, Spiegelhalter *et al.*, 1993, Pearl, 1988, Pearl, 1993, Pearl *et al.*, 1990].

3.3 Causal discovery

The interpretation of DAGs as carriers of independence assumptions does not necessarily imply causation and will in fact be valid for any set of Markovian independencies along any ordering (not necessarily causal or chronological) of the variables. However, the patterns of independencies portrayed in a DAG are typical of causal organizations, and some of these patterns can only be given meaningful interpretation in terms of causation. Consider, for example, the following *intransitive* pattern of dependencies among three events: E_1 and E_3 are dependent, E_3 and E_2 are dependent, yet E_1 and E_2 are independent. If you ask a person to supply an example of three such events, the example invariably portrays E_1 and E_2 as two independent causes and E_3 as their common effect, namely, $E_1 \rightarrow E_3 \leftarrow E_2$. Fitting this dependence pattern by using E_3 as the cause and E_1 and E_2 as the effects, although mathematically feasible, is very unnatural indeed (the reader is encouraged to try this exercise).

Such thought experiments teach us that certain patterns of dependency, totally void of temporal information, are conceptually characteristic of certain causal directionalities and not others. Reichenbach (1956) has suggested that this temporal asymmetry is a characteristic of Nature, reflective of the second law of thermodynamics. Pearl and Verma (1991) have offered a more subjective explanation, attributing the asymmetry to choice of language and to certain assumptions (e.g., Occam's razor) prevalent in scientific induction. Regardless of the origins of this asymmetry, exploring whether it provides a significant source of causal information (or at least causal clues) in human learning is an interesting topic for research [Waldmann *et al.*, 1995].

The distinction between transitive and intransitive dependencies has become the basis for algorithms aimed at extracting causal structures from raw statistical data. Several systems that systematically search and identify causal structures from empirical data have been developed [Pearl, 1988, page 387-397] and [Pearl and Verma, 1991, Spirtes *et al.*, 1993]. Technically, because these algorithms rely solely on conditional independence relationships, the structures found are valid only if one is willing to accept forms of guarantees that are weaker than those obtained through controlled randomized experiments – namely, minimality and stability [Pearl and Verma, 1991]. Minimality guarantees that any other structure compatible with the data is necessarily less specific, and hence less falsifiable and less trustworthy, than the one(s) inferred. Stability ensures that any alternative structure compatible with the data must be less stable than the one(s) inferred; in other words, slight fluctuations in experimental conditions will render the alternative structure incompatible with the data. With these forms of guarantees, the algorithms can provide criteria for identifying genuine and spurious causes, with or without temporal information.

Minimality can be easily illustrated in Figure 1: if one draws all graphs that are observational equivalent to the one shown in the figure (there are exactly three such graphs) one finds that they all contain an arrow directed from X_2 to X_4. This still does not make X_2 a genuine cause of X_4, because the specific data at hand, summarized in P, could in fact be generated by another graph, say G', which is not observationally equivalent to G, and in which an arrow is directed the other way around, from X_4 to X_2. For example, one choice of G' would be a complete DAG (i.e., one containing a link between every pair of nodes) rooted at X_4; although G' contains an arrow from X_4 to X_2, it could be made (with the proper choice of parameters) to represent any probability distribution whatsoever, including P. Is there a rationale, then, for preferring G on G', given that both represent P precisely (in the sense of Eq. (6))? There is! Having the potential of fitting any data means that G' is empirically nonfalsifiable, that P is overfitted, hence, that G' is less trustworthy than G. This preference argument can be advanced not merely to complete DAGs but against any DAG G' that can be made to fit more experimental data (i.e., probability functions) than G. Indeed, it can be shown that the set of probabilities representable by any DAG G' which fits P and contains an arrow from X_4 to X_2 would necessarily be a superset of those represented by G.

The minimality argument above rests on the closed world assumption, and would fail if hidden variables are permitted. For example, the DAG $X \leftarrow a \rightarrow Z \leftarrow b \rightarrow Y$ imposes the same set of independencies on the observed variables X, Y, Z as the v-structure $X \rightarrow Z \leftarrow Y$, yet the former does not present X as a cause of Z. The remarkable thing about minimality, however, is that it uniquely determines the directionality of some arrows even when we dispose of the closed-world assumption and allow for the presence of hidden variables. The arrow from X_4 to X_5 in Figure 1 is an example of such occasion. Among all DAGs that fit P, including DAGs containing unobserved variables, those which do not include an arrow from X_4 to X_5 are nonminimal, i.e., each fits a superset of the probability distributions (on the observables) represented by G. It is this feature that encouraged Pearl and Verma (1991) to label certain links in the DAG "genuine causes", to be distinguished from "potential causes" and "spurious associations". The latter identifies certain associations as non-causal (i.e., no link exists between the corresponding nodes in all minimal DAG's that fit the data) implying that the observed association must be attributed to a hidden common cause between the corresponding variables. Criteria and algorithms for identifying genuine causes, potential causes, and spurious associations are described in Pearl and Verma (1991) and Spirtes et al (1993).

Alternative methods of identifying causal structures in data assign prior probabilities to the parameters of the network and use Bayes' rule to score the degree to which a given network fits the data [Cooper and Herskovits, 1990, Heckerman *et al.*, 1994]. These methods have the advantage of operating well under small-sample conditions, but they encounter difficulties in coping with hidden variables.

4 Structural Causality

While Bayesian networks capture patterns of independencies that are characteristic of causal organizations, they still leave open the question of how these patterns relate to the more basic notions associated with causation, such as influence, manipulation, and control, which reside outside the province of probability theory. Manipulations are unquestionably central to the analysis of causal thinking. Even generative accounts of causality, according to which causal inquiries aim merely at gaining an "understanding" of how data are generated, are not totally divorced from notions of manipulation, albeit hypothetical. In the final analysis, the quest for understanding "how data is generated" or "how things work" is merely a quest for predictions of what could be expected if things were taken apart and reconfigured in various ways, that is, for expectations under various hypothetical manipulations.

An inspection of the Bayesian network depicted in Figure 1 reveals that the network does in fact provide an effective representation for certain kinds of manipulations and changes of configuration. Any local reconfiguration of the mechanisms in the environment can be translated, with only minor modification, into an isomorphic reconfiguration of the network topology. For example, to represent a disabled sprinkler, we simply delete from the network all links incident to the node "Sprinkler"; to represent a pavement covered by a tent, we simply delete the link between "Rain" and "Wet." This flexibility is often cited as the ingredient that marks the division between deliberative and reactive agents, and that enables the former to manage novel situations instantaneously, without requiring training or adaptation. How then are these extra-probabilistic notions of reconfiguration and manipulation connected to the strictly probabilistic notion of conditional independence, which forms the standard basis for Bayesian networks and the entire study of probabilistic causality?

The connection is made through the structural account of causation, according to which probabilistic dependencies are but a surface phenomenon of more fundamental relationships – functional dependencies among stable, or autonomous, mechanisms. The roots of this account go back to path analysis in genetics [Wright, 1921] and structural equation models in econometrics [Haavelmo, 1943, Simon, 1953], and it can justly be regarded as the mathematical basis for the power models used in the psychological literature. The basic idea behind the structural account was extended in [Pearl and Verma, 1991] for defining general probabilistic causal theories, as follows. Each child-parents family in a DAG G represents a deterministic function

$$X_i = f_i(pa_i, \epsilon_i) \tag{8}$$

where pa_i are the parents of variable X_i in G, and where ϵ_i, $0 < i < n$, are mutually independent, arbitrarily distributed random disturbances. Characterizing each child-parent relationship as a deterministic function, instead of as the usual conditional probability $P(x_i \mid pa_i)$, imposes equivalent independence constraints on the resulting distributions and leads to the same recursive decomposition that characterizes DAG models (see Eq. (6)). However, the functional characterization $X_i = f_i(pa_i, \epsilon_i)$ also specifies how the resulting distributions would change in response to external interventions, since each function is presumed to represent a stable mechanism in the domain and therefore remains constant unless specifically altered. Thus, once we know the identity of the mechanisms altered by an intervention and the nature of the alteration, the overall effect of an intervention can be predicted by modifying the appropriate equations in the model of Eq. (8) and using the modified model to compute a new probability function of the observables.

The simplest type of external intervention is one in which a single variable, say X_i, is forced to take on some fixed value x_i'. Such *atomic* intervention amounts to replacing the old functional mechanism $X_i = f_i(pa_i, \epsilon_i)$ with a new mechanism $X_i = x_i'$ which represents the external force that sets the value x_i'. If we imagine that each variable X_i could potentially be subject to the influence of such an external force, then we can view each Bayesian network as an efficient code for predicting the effects of atomic interventions and of myriad combinations of such interventions, without encoding these interventions explicitly. What is more remarkable yet is that it is possible, under certain conditions, to predict the effect of interventions without knowing the functions $\{f_i\}$; the topology of the graph combined with the probability of the observables suffice. This means that it should be possible to infer causal influences, in the presence of unmeasured variables, from a combination of statistical data and qualitative linguistic assertions about the general workings of mechanisms. The following subsection presents these ideas in a formal setting.

4.1 Causal theories and actions

Definition 5 *A causal theory is a four-tuple*

$$T =< V, U, P(u), \{f_i\} >$$

where

(i) $V = \{X_1, \ldots, X_n\}$ *is a set of observed variables,*

(ii) $U = \{U_1, \ldots, U_m\}$ *is a set of exogenous (often unmeasured) variables that represent disturbances, abnormalities, or assumptions,*

(iii) $P(u)$ *is a distribution function over* $U_1, \ldots, U_m$, *and*

(iv) $\{f_i\}$ *is a set of* n *deterministic functions, each of the form*

$$X_i = f_i(PA_i, u) \quad i = 1, \ldots, n \tag{9}$$

where PA_i *is a subset of variables in* V *not containing* X_i.

We will assume that the set of equations in (iv) has a unique solution for $X_i, \ldots, X_n$, given any value of the disturbances $U_1, \ldots, U_m$. Therefore, the distribution $P(u)$ induces a unique distribution on the observables, which we denote by $P_T(v)$. The structural parent sets, PA_i, are again considered the direct causes of X_i and they define a directed graph G which may, in general, be cyclic. However, unlike the Markovian parents defined in Subsection 2.1 (see Definition 2), PA_i is selected from V by considering outcomes of manipulative experiments (according to Lemma 4 below), not by conditional independence considerations, as in probabilistic causality. The result of encoding this manipulative information in the equations will be a major relaxation of the small-world assumption (Subsection 2.3); the analysis of actions will require only rudimentary, qualitative assumptions about the structure of the unmeasured U variables.

Consider the example depicted in Figure 1. The corresponding theory consists of five functions, each representing an autonomous mechanism:

$$\begin{aligned} X_1 &= U_1 \\ X_2 &= f_2(X_1, U_2) \\ X_3 &= f_3(X_1, U_3) \\ X_4 &= f_4(X_3, X_2, U_4) \\ X_5 &= f_5(X_4, U_5) \end{aligned} \tag{10}$$

The disturbances $U_1, \ldots, U_5$ are not shown explicitly in the graph of Figure 1, but are understood to govern the uncertainties associated with the causal relationships. A typical specification of the functions $\{f_1, \ldots, f_5\}$ and the disturbance terms is given by the Boolean theory below:

$$\begin{aligned} x_2 &= [(X_1 = \text{Winter}) \vee (X_1 = \text{Fall}) \vee ab_2] \wedge \neg ab'_2 \\ x_3 &= [(X_1 = \text{Summer}) \vee (X_1 = \text{Spring}) \vee ab_3] \wedge \neg ab'_3 \\ x_4 &= (x_2 \vee x_3 \vee ab_4) \wedge \neg ab'_4 \\ x_5 &= (x_4 \vee ab_5) \wedge \neg ab'_5 \end{aligned} \tag{11}$$

where x_i stands for $X_i = true$, and ab_i and ab'_i stand, respectively, for triggering and inhibiting abnormalities.[8] For example, ab_4 stands for (unspecified) events which might cause the ground to get wet (x_4) when the sprinkler is off ($\neg x_2$) and it does not rain ($\neg x_3$), while $\neg ab'_4$ stands for events which will keep the ground dry despite the rain, the sprinkler and ab_4, say covering the ground with plastic sheet.

As stated in the introductory subsection, the main role of structural causal theories is to facilitate the analysis of actions. We will consider local concurrent actions of the form $do(X = x)$, where $X \subseteq V$ is a set of variables and x is a set of values from the domain of X. In other words, $do(X = x)$ represents a combination of direct actions that forces the variables in X to attain the values x.

Definition 6 (effect of actions) *The effect of the action $do(X = x)$ on a causal theory T is given by a subtheory T_x of T, where T_x obtains by deleting from T all equations corresponding to variables in X and substituting the equations $X = x$ instead.*

For example, to represent the action "turning the sprinkler ON," $do(X_3 = \text{ON})$, we delete the equation $X_3 = f_3(X_1, U_3)$ from the theory of Eq. (10), and replace it with $X_3 = \text{ON}$. The resulting subtheory, $T_{X_3=\text{ON}}$, contains all the information needed for computing the effect of the actions on other variables. It is easy to see from this subtheory that the only variables affected by the action are X_4 and X_5, that is, the descendants, of the manipulated variable X_3. This is to be expected, since nondescendants of X_3 (i.e., season and rain) are presumed to be causally irrelevant to X_3, yet it stands in marked contrast to the operation

[8]Goldszmidt and Pearl (1992, 1995) describe a qualitative method of causal analysis based on attributing infinitesimal probabilities to the ab predicates.

of probabilistic conditionalization (on X_3) which may potentially influence (the beliefs in) every variable in the network. The mathematics underlying these two operations, and the conditions that enable us to predict the effects of actions without specifying $\{f_i\}$, will be discussed in the next two subsections.

Definition 6 should be taken as an integral part of Definition 5, because it assigns meaning to each individual equation in T. Specifically, it dictates what hypothetical experiments of the type $do(X = x)$ must be considered by the author of the structural equations in deciding which variables PA_i should enter into the r.h.s of each equation. By writing $X_4 = f_4(X_2, X_3, u)$, for example, the analyst defines X_2 and X_3 as the direct causes of X_4 which, according to Definition 6, means that holding X_2 and X_3 fixed determines the value of X_4 regardless of changes in the season (X_1) and regardless of any direct action we might take to make the ground slippery (X_5). In general, Definition 6 endows PA_i with the following meaning: PA_i is a set of variables that, if held fixed, would determine (for any u) the value of X_i regardless of any other action $do(Z = z)$ that one may perform, where Z is any set of variables not containing X_i or any member of PA_i. Moreover, no proper subset of PA_i possesses that quality.

Lemma 4 provides a succinct summary of this property, and can also be viewed as the structural definition of direct causes.

Lemma 4 *Let $Y(x; u)$ stand for the solution of Y under subtheory T_x, as in Definition 6. The direct causes of variable X_i are the minimal set of variables PA_i which satisfy*

$$X_i(pa_i, z; u) = X_i(pa_i; u) \tag{12}$$

for every u and for every set Z not containing X_i or any member of PA_i. (pa_i denotes a specific instantiation of PA_i).

Clearly, if a causal theory is given explicitly, as in Definition 5, then the direct causes PA_i can be identified syntactically, as the arguments of each f_i. However, if the theory is represented implicitly in a form of a function F: Actions $\times\ U \rightarrow V$, (as is often assumed in decision theory [Savage, 1954, Heckerman and Shachter, 1995]), then Lemma 4 can be used to identify, given F, the unique set of direct causes for each variable X_i.[9]

We see that the distinctive characteristic of structural equations, which sets them apart from ordinary algebraic equations, is that meaning is attached to any subset of equations from T. Mathematically, this characteristic does not show up explicitly in the equations, but rather implicitly, in the understanding that T stands for not one but 2^n sets of equations. This restricts, of course, the type of algebraic transformations admissible on T to those that preserve the solution of not one but each of the 2^n sets.

The framework provided by Definitions 5 and 6 permits the coherent formalization of many nuances and subtle concepts found in causal conversation, including causal influence, causal effect, causal relevance, average causal effect, identifiability, counterfactuals, and exogeneity. Examples are:

[9] Likewise, the local operator $do(X_i = x_i)$ can be identified from F as the unique action A for which the equality $F(A, u)_i = F(A \text{ and } B, u)_i$ holds for every action B compatible with A. In words, $do(X_i = x_i)$ is the only action which keeps the value of X_i invariant to any other action that can be implemented at the same time.

- X **influences** Y **in context** u if there are two values of X, x and x', such that $Y(x; u) \neq Y(x'; u)$. In other words, the solution for Y under $U = u$ and $do(X = x)$ is different from the solution under $U = u$ and $do(X = x')$.

  We say, for example, that the weather (X_2) influences the wetness of the pavement (X_4) in a context u where the pavement is uncovered, and the sprinkler controller is at off position, because a change in weather from not-rain to rain is accompanied with a change in pavement condition from dry to wet. This definition interprets causal influence as the transference of change from X to Y triggered by the local intervention $do(X = x)$. Although the word "influence" is sometimes used with no intervention in mind (as in the case of the weather), the hypothetical operator $do(X = x)$ ensures that the change in Y is attributable only to changes in X, and not to spurious side effects (e.g., strange people who turn their sprinklers on whenever they see clouds in the sky.)

- X **can potentially influence** Y in context $U = u$ if there exists a subtheory T_z of T in which X influences Y.

  The difference between influence and potential influence is that the latter requires an additional intervention, $do(Z = z)$, to reveal the effect of X on Y. In our earlier example, we find it plausible to maintain that, although the weather does not influence wetness in a context (u) where the sprinkler controller is stuck at ON position, it nevertheless can potentially influence wetness, at u, as is revealed when the action $do(Sprinkler\ =\ \text{OFF})$ is implemented, say, by manual intervention. Along the same vein, we may say that seasonal variations (X_1) have potential influence on wetness, even though their influence through rain may perfectly cancel their influence through sprinkler; This potential would surface when we hold *Sprinkler* fixed (at either ON or OFF position).[10]

- **Event** $X = x$ **is the (singular) cause of event** $Y = y$ if (i) $X = x$ and $Y = y$ are true and (ii) in every context u compatible with $X = x$ and $Y = y$, and for all $x' \neq x$, we have $Y(x'; u) \neq y$.

  This definition reflects the counterfactual explication of a singular cause: "$Y = y$ would be false if it were not for $X = x$," as used in Section 2.4. A separate analysis of counterfactuals will be given in Section 4.5.

4.2 Probabilistic causal effects and identifiability

The definitions above are deterministic. Probabilistic causality emerges when we define a probability distribution $P(u)$ for the U variables. Under the assumption that the set of equations $\{f_i\}$ and every subset thereof has a unique solution, $P(u)$ induces a unique distribution $P_{T_x}(v)$ on the endogenous variables for each combination of atomic interventions $do(X = x)$. This leads to a natural probabilistic definition of causal effects.

[10] The standard example in the philosophical literature [Cartwright, 1989] involves the potential positive influence of birth-control pills on thrombosis, which might be masked by its negative effect on pregnancy (another cause of thrombosis). Cartwright proposal (rejected by Eells), that the influence of the pill be assessed by considering separately the population of women that would get pregnant (or remain non-pregnant) regardless of the pill, amounts to considering a subtheory T_z in which pregnancy (Z) is held fixed.

Definition 7 (causal effect) *Given two disjoint subsets of variables, $X \subseteq V$ and $Y \subseteq V$, the causal effect of X on Y, denoted $P_T(y|do(x))$ or $P_T(y|\hat{x})$, gives the distribution of Y induced by the action $do(X = x)$, that is,*

$$P_T(y|\hat{x}) = P_{T_x}(y) \tag{13}$$

for each realization x of X.

The probabilistic notion of causal effect is much weaker than its deterministic counterparts of causal influence and potential causal influence. For example (from Subsection 2.2), if U is the outcome of a fair coin, X is my bet, and Y stands for winning a dollar iff $X = U$, then the causal effect of X on Y is nil, because $P(y|do(X = Tail)) = P(y|do(X = Head) = \frac{1}{2}$. At the same time, X will qualify as having an influence on Y in every possible context, $U = Head$ and $U = Tail$. Note that causal effects are defined relative to a given causal theory T, though the subscript T is often suppressed for brevity.

Definition 8 (identifiability) *Let $Q(T)$ be any computable quantity of a theory T. Q is* identifiable *in a class M of theories if for any pairs of theories T_1 and T_2 from M, $Q(T_1) = Q(T_2)$ whenever $P_{T_1}(v) = P_{T_2}(v)$.*

Identifiability is essential for integrating statistical data (summarized by $P(v)$) with incomplete prior causal knowledge of $\{f_i\}$, as it enables the reasoner to estimate quantities Q from P alone, without specifying the details of T, so that the general characteristics of the class M suffice.[11] For the purpose of our analysis, the quantity Q of interest is the causal effect $P_T(y|\hat{x})$ which is certainly computable from a given theory T (using Eq. (13)), but which we will now attempt to compute from incomplete specification of T, in the form of general characteristics such as the identities of the parent sets PA_i and the independencies embedded in $P(u)$. We will therefore consider a class M of theories which have the following characteristics in common:

(i) they share the same parent-child families (i.e., the same causal graph G),

(ii) they share the same set of independencies in $P(u)$, and,

(iii) they induce positive distributions on the endogenous variables,[12] i.e., $P(v) > 0$.

Relative to such classes we now define:

Definition 9 (causal-effect identifiability) *The* causal effect *of X on Y is said to be* identifiable *in M if the quantity $P(y|\hat{x})$ can be computed uniquely from the probabilities of the observed variables, that is, if for every pair of theories T_1 and T_2 in M such that $P_{T_1}(\mathrm{v}) = P_{T_2}(\mathrm{v})$, we have $P_{T_1}(y|\hat{x}) = P_{T_2}(y|\hat{x})$.*

[11] The notion of identifiability is central to much work in econometrics, where it has become synonymous to the identification of the functions $\{f_i\}$ or some of their parameters [Koopman and Reiersol, 1950], mostly under conditions of additive Gaussian noise. Definition 8, which does not assume any parametric representation of the functions $\{f_i\}$, extends the notion of identifiability to quantities Q that do not require the precision of parametric models. In particular, it permits one (see Definition 9) to dispose with the identification of functional parameters altogether, and deal directly with causal effects $P(y|\hat{x})$ – the very purpose of identifying parameters in policy-analysis applications.

[12] This requirement ensures that the disturbances U are sufficiently rich to simulate a "natural experiment", that is, an experiment in which conditions change by natural phenomena rather than a human experimenter.

The identifiability of $P(y|\hat{x})$ ensures that it is possible to infer the effect of action $do(X = x)$ on Y from two sources of information:

(i) passive observations, as summarized by the probability function $P(v)$,

(ii) the causal graph, G, which specifies, qualitatively, which variables make up the stable mechanisms in the domain or, alternatively, which variables participate in the determination of each variable in the domain.

Simple examples of identifiability will be discussed in the next subsection.

4.3 Inferring consequences of actions from passive observations

The probabilistic analysis of actions becomes particularly simple when two conditions are satisfied:

1. The theory is recursive, that is, there exists an ordering of the variables $V = \{X_1, \ldots, X_n\}$ such that each X_i is a function of a subset PA_i of its predecessors

$$X_i = f_i(PA_i, U_i) \qquad PA_i \subseteq \{X_1, \ldots, X_{i-1}\} \tag{14}$$

2. The disturbances $U_1, \ldots, U_n$ are mutually independent, which implies (from the exogeneity of the U_i's)

$$U_i \perp\!\!\!\perp \{X_1, \ldots, X_{i-1}\} \tag{15}$$

These two conditions, also called Markovian, are the basis of the independencies embodied in Bayesian networks (Section 3.2), and they enable us to compute causal effects directly from the conditional probabilities $P(x_i|pa_i)$, without specifying either the functional form of the functions f_i or the distributions $P(u_i)$ of the disturbances [Pearl, 1993, Spirtes *et al.*, 1993]. This is seen immediately from the following observations: On the one hand, the distribution induced by any Markovian theory T is given by the product in Eq. (6),

$$P_T(x_1, \ldots, x_n) = \prod_i P(x_i|pa_i) \tag{16}$$

where pa_i are (values of) the parents of X_i in the diagram representing T. On the other hand, the subtheory $T_{x'_j}$, representing the action $do(X_j = x'_j)$, is also Markovian; hence, it also induces a product-like distribution

$$P_{T_{x'_j}}(x_1, \ldots, x_n) \;=\; \begin{cases} \prod_{i \neq j} P(x_i|pa_i) = \frac{P(x_1,\ldots,x_n)}{P(x_j|pa_j)} & \text{if } \; x_j = x'_j \\ 0 & \text{if } \; x_j \neq x'_j \end{cases} \tag{17}$$

where the partial product reflects the surgical removal of the equation $X_j = f_j(pa_j, U_j)$ from the theory of Eq. (14). Thus, we see that both the pre-action and the post-action distributions depend only on observed conditional probabilities but are independent of the particular functional form of $\{f_i\}$ and of the distributions $P(u)$ that generate those probabilities. This is the essence of identifiability as given in Definition 9, which stems from the Markovian assumptions (14) and (15). Section 4.4 will demonstrate that certain, though not all, causal effects are identifiable even when the Markovian property is destroyed by introducing dependencies among the disturbance terms.

In the example of Figure 1, the pre-action distribution is given by the product

$$P_T(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_4) \tag{18}$$

while the surgery corresponding to the action $do(X_3 = \text{ON})$ amounts to deleting the link $X_1 \to X_3$ from the graph and fixing the value of X_3 to ON, yielding the post-action distribution

$$P_T(x_1, x_2, x_4, x_5|do(X_3 = \text{ON})) = P(x_1)\ P(x_2|x_1)\ P(x_4|x_2, X_3 = \text{ON})\ P(x_5|x_4) \tag{19}$$

Note the difference between the action $do(X_3 = \text{ON})$ and the observation $X_3 = \text{ON}$. The latter is encoded by ordinary Bayesian conditioning,

$$P_T(x_1, x_2, x_4, x_5|X_3 = \text{ON}) = \frac{P(x_1)\ P(x_2|x_1)\ P(x_3 = \text{ON}|x_1)P(x_4|x_2, X_3 = \text{ON})P(x_5|x_4)}{P(X_3 = \text{ON})}$$

The former is obtained by conditioning a mutilated graph, with the link $X_1 \to X_3$ removed. This mirrors indeed the difference between seeing and doing: after observing that the sprinkler is ON, we wish to infer that the season is dry, that it probably did not rain, and so on; no such inferences should be drawn in evaluating the effects of the deliberate action "turning the sprinkler ON." The excision of $X_3 = f_3(X_1, U_3)$ from (10) ensures the suppression of any abductive inferences from the action, as well as from any of its consequences.

Generalization to multiple actions and conditional actions is straightforward. Multiple actions $do(X = x)$, where X is a compound variable, result in a distribution similar to (17), except that all factors corresponding to the variables in X are removed from the product in (16). Stochastic conditional strategies [Pearl, 1994] of the form

$$do(X_j = x_j) \text{ with probability } P^*(x_j|pa_j^*) \tag{20}$$

where PA_j^* is the support set of the decision strategy, also result in a product decomposition similar to (16), except that each factor $P(x_j|pa_j)$ is *replaced* with $P^*(x_j|pa_j^*)$.

4.4 A calculus of acting and seeing

The identifiability of causal effects demonstrated in Section 4.3 relies critically on the Markovian assumptions given in (14) and (15). If a variable that has two descendants in the graph is unobserved, the disturbances in the two equations are no longer independent, the Markovian property (14) is violated, and identifiability may be destroyed. This can be seen easily from Eq. (17); if any parent of the manipulated variable X_j is unobserved, one cannot estimate the conditional probability $P(x_j|pa_j)$, and the effect of the action $do(X_j = x_j)$ may not be predictable from the observed distribution $P(x_1, \ldots, x_n)$. Fortunately, certain causal effects are identifiable even in situations where members of pa_j are unobservable [Pearl, 1993]. Moreover, polynomial tests are now available for deciding when $P(x_i|\hat{x}_j)$ is identifiable and for deriving closed-form expressions for $P(x_i|\hat{x}_j)$ in terms of observed quantities [Galles and Pearl, 1995].

These tests and derivations are based on a symbolic calculus [Pearl, 1994b, 1995], to be described in the sequel, in which interventions, side by side with observations, are given explicit notation and are permitted to transform probability expressions. The transformation rules of this calculus reflect the understanding that interventions perform "local surgeries" as described in Definition 6, namely, they overrule equations that tie the manipulated variables to their pre-intervention causes.

Let X, Y, and Z be arbitrary disjoint sets of nodes in a DAG G. We say that X and Y are independent given Z in G, denoted $(X \perp\!\!\!\perp Y|Z)_G$, if the set Z d-separates X from Y in G. We denote by $G_{\overline{X}}$ the graph obtained by deleting from G all arrows pointing to nodes in X. Likewise, we denote by $G_{\underline{X}}$ the graph obtained by deleting from G all arrows emerging from nodes in X. To represent the deletion of both incoming and outgoing arrows, we use the notation $G_{\overline{X}\underline{Z}}$. Finally, the expression $P(y|\hat{x}, z) \triangleq P(y, z|\hat{x})/P(z|\hat{x})$ stands for the probability of $Y = y$ given that $Z = z$ is observed and X is held constant at x.

Theorem 5 Let G be the DAG associated with a Markovian causal theory, and let $P(\cdot)$ stand for the probability distribution induced by that theory. For any disjoint subsets of variables X, Y, Z, and W we have:

Rule 1 Insertion/deletion of observations

$$P(y|\hat{x}, z, w) = P(y|\hat{x}, w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}} \tag{21}$$

Rule 2 Action/observation exchange

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, z, w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}} \tag{22}$$

Rule 3 Insertion/deletion of actions

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, w) \text{ if } (Y \perp\!\!\!\perp Z|X,\ W)_{G_{\overline{X},\ \overline{Z(W)}}} \tag{23}$$

where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$.

Each of the inference rules above follows from the basic interpretation of the $\hat{x}$ operator as a replacement of the causal mechanism that connects X to its pre-action parents by a new mechanism $X = x$ introduced by the intervening force.

Corollary 1 *A causal effect Q: $P(y_1, ..., y_k|\hat{x}_1, ..., \hat{x}_m)$ is identifiable in a model characterized by a graph G if there exists a finite sequence of transformations, each conforming to one of the inference rules in Theorem 5, which reduces q into a standard (i.e., hat-free) probability expression involving observed quantities.*

Although Theorem 5 and Corollary 1 require the Markovian property, they can also be applied to non-Markovian, recursive theories, because such theories become Markovian if we consider the unobserved variables as part of the analysis and represent them as nodes in the graph. To illustrate: Assume that variable X_1 in Figure 1 is unobserved, rendering the disturbances U_3 and U_2 dependent since these terms now include the common influence of X_1. Theorem 5 tells us that the causal effect $P(x_4|\hat{x}_3)$ is identifiable, because

$$P(x_4|\hat{x}_3) = \sum_{x_2} P(x_4|\hat{x}_3, x_2)P(x_2|\hat{x}_3) \tag{24}$$

Rule 3 permits the deletion

$$P(x_2|\hat{x}_3) = P(x_2) \tag{25}$$

because $(X_2 \perp\!\!\!\perp X_3)_{G_{\overline{X_3}}}$, while Rule 2 permits the exchange

$$P(x_4|\hat{x}_3, x_2) = P(x_4|x_3, x_2) \tag{26}$$

because $(X_4 \perp\!\!\!\perp X_3|X_2)_{G_{\underline{X_3}}}$. This gives

$$P(x_4|\hat{x}_3) = \sum_{x_2} P(x_4|x_3, x_2)P(x_2) \tag{27}$$

which is a hat-free expression, involving only observed quantities.

The reader might recognize Eq. (27) as the standard formula for covariate adjustment (also called "stratification"), which is used in experimental design both for improving precision and for minimizing confounding bias. However, a formal, general criterion for deciding whether a set of covariates Z (X_2 in our example) qualifies for adjustment has long been wanting [Smith, 1957, Wainer, 1991, Shafer, 1995].[13] Theorem 5 provides such a criterion (called the "back-door criterion" in [Pearl, 1993]) which reads:

Definition 10 *Z is an admissible set of covariates relative to the effect of X on Y if:*

(i) *no node in Z is a descendant of X, and*

(ii) *Z d-separates X from Y along any path containing an arrow into X (equivalently, $(Y \perp\!\!\!\perp X|Z)_{G_{\underline{X}}}$).*

We see, for instance, that X_2 and X_1 (or both) qualify as admissible covariates relative the effect of X_3 on X_4, but X_5 will not qualify. The graphical definition of admissible covariates replaces statistical folklore with formal procedures, and should enable analysts to systematically select an optimal set of observations, namely, a set Z that minimizes measurement cost or sampling variability.

In general, it can be shown [Pearl, 1995] that:

1. The effect of interventions can often be identified (from nonexperimental data) without resorting to parametric models.

[13] Most of the statistical literature is satisfied with informal warnings that "Z should be quite unaffected by X" [Cox, 1958, page 48], which is necessary but not sufficient, or that X should not precede Z [Shafer, 1995, page 294], which is neither necessary nor sufficient. In some academic circles, a criterion called "ignorability" is invoked [Rosenbaum and Rubin, 1983], which merely paraphrases the problem in the language of counterfactuals. Simplified, it reads: Z is an admissible covariate relative to the effect of X on Y if, for every x, the value that Y would obtain had X been x is conditionally independent of X, given Z.

2. The conditions under which such nonparametric identification is possible can be determined by simple graphical criteria.

3. When the effect of interventions is not identifiable, the causal graph may suggest nontrivial experiments which, if performed, would render the effect identifiable.

While the ability to assess the effect of interventions from nonexperimental data has immediate applications in the medical and social sciences, such assessments are also important in psychological learning theory: they explain how agents can predict the effect of the next action (e.g.,turning the sprinkler on) on the basis of past experience, where that action has never been enacted out of free will, but only in response to environmental needs (e.g., dry season) or to other agents' requests.

4.5 Processing counterfactuals

A counterfactual sentence has the form

If A were true, then C would have been true, given O

where A, the counterfactual antecedent, specifies an event that is contrary to one's real-world observations O, and C, the counterfactual consequent, specifies a result that is expected to hold in an alternative world where the antecedent is true. A typical example is "If Oswald were not to have shot Kennedy, then Kennedy would still be alive," which presumes the factual knowledge of Oswald's assassination of Kennedy, contrary to the antecedent of the sentence.

The majority of the philosophers who have examined the semantics of counterfactual sentences have resorted to some version of Lewis' "closest world" approach: "C if it were A" is true, if C is true in worlds that are "closest" to the real world yet consistent with the counterfactual antecedent A [Lewis, 1973]. While the closest world approach leaves the precise specification of the closeness measure almost unconstrained, causal knowledge imposes very specific preferences as to which worlds should be considered closest to any given world. For example, consider an array of domino tiles standing close to each other. The manifestly closest world consistent with the statement "tile i is tipped to the right" would be a world in which just tile i is tipped, while all the others remain erect. Yet, we all accept the counterfactual sentence "Had tile i been tipped to the right, tile $i+1$ would be tipped as well" as plausible and valid. Thus, distances among worlds are not determined merely by surface similarities but require a distinction between explained and unexplained dissimilarities. The local surgery paradigm expounded in Section 4.1 offers a concrete explication of the closest-world approach which respects such causal considerations. A world w_1 is "closer" to w than a world w_2 is, if the set of atomic surgeries needed for transforming w into w_1 is a proper subset of those needed for transforming w into w_2. In the domino example, finding tile i tipped and $i+1$ erect requires the alteration of two basic mechanisms (i.e., two unexplained actions or "miracles" [Lewis, 1973]) compared with one altered mechanism for the world in which all j tiles, $j > i$, are tipped. This paradigm conforms to our perception of causal influences and lends itself to economical machine representation.

The structural equations framework, coupled with the surgical operator $do(X = x)$, also offers the syntactic machinery for counterfactual analysis, while leaving the closest-world interpretation implicit. The basis for this analysis is the potential response function $Y(x;u)$ invoked in Lemma 4, which we take as the formal explication of the English phrase "the value that Y would obtain in context u, had X been x".

Definition 11 (potential response) *Given a causal theory T the potential response of Y to X in a context u, denoted $Y(x;u)$ or $Y_x(u)$, is the solution for Y under $U=u$ in the subtheory T_x.* [14]

Note that this definition allows for the context $U=u$ and the proposition $X=x$ to be incompatible in T. For example, if T describes a logic circuit with input U, it may well be reasonable to assert the counterfactual: "Given $U=u$, voltage Y would be high if current X were low," even though the input $U=u$ may preclude X from being low. It is for this reason that one must invoke some notion of intervention (alternatively, a theory change or a "miracle"[Lewis, 1973]) in the definition of counterfactuals. This is further attested by the suppression of abductive arguments in counterfactual reasoning; for example, the following sentence would be deemed unacceptable: "Had I done my homework, I would have felt miserable, because I always do my homework after my father beats me up." The reason we do not accept this argument is that it conflicts with the common understanding that the counterfactual antecedent "done my homework" should be considered an external willful act, totally free of normal inducements (e.g., beatings), as modeled by the surgical subtheory T_x.

Counterfactual sentences rarely specify a complete context u. Instead they imply a partial description of u in the form of a set o of (often implicit) facts or observations. Thus, a general counterfactual sentence would have the format $x \rightarrow y|o$, read "Given factual knowledge o, Y would obtain the value y had X been x." For example, the sentence "If Oswald were not to have shot Kennedy, then Kennedy would still be alive" would be formulated:

$$\neg Shot(Oswald, Kennedy) \rightarrow Alive(Kennedy) \mid Dead(Kennedy), Shot(Oswald, Kennedy)$$

The truth of such a sentence in a theory T can be defined in terms of the potential response $Y(x;u)$ as follows:

Definition 12 (counterfactual assertability) *The sentence $x \rightarrow y|o$ is true in T if $Y(x;u) = y$ for every u compatible with o.*

This definition parallels Lewis's closest world approach, with u playing the role of a possible world. Note the difference between the treatments of o and x; the former insists on direct compatibility between u and o, while the latter tolerates a surgical face-lift where x and u are incompatible.

If U is treated as a random variable, then the value of the counterfactual $Y(x;u)$ becomes a random variable as well, denoted $Y(x)$ or Y_x. Moreover, the distribution of this random variable is easily seen to coincide with the causal effect $P(y|\hat{x})$:

$$P((Y(x) = y) = P(y|\hat{x})$$

[14] The term *unit* instead of *context* is often used in the statistical literature [Rubin, 1974], where it normally stands for the identity of a specific individual in a population, namely, the set of attributes u that characterize that individual. In general, u may include the time of day, the experimental conditions under study, and so on. Practitioners of the counterfactual notation do not explicitly mention the notions of "solution" or "intervention" in the definition of $Y(x;u)$. Instead, the phrase "the value that Y would take in unit u, had X been x," viewed as basic, is posited as the definition of $Y(x;u)$.

Thus, the probability of a counterfactual conditional $x \rightarrow y \mid o$ may be evaluated by the following procedure:

- Use the observations o to update $P(u)$, thus forming a revised causal theory $T^o =< V, U, \{f_i\}, P(u|o) >$
- Form the mutilated theory T^o_x (by deleting from T^o the equation corresponding to variables in X) and compute the probability $P_{T^o}(y|\hat{x})$ that T^o_x induces on Y.

In Subsection 2.4 we have demonstrated that, unlike causal-effect queries, counterfactual queries may not be identifiable in Markovian theories, but require that the functional form of $\{f_i\}$ be specified. However, the example also shows that the counterfactual probabilities computed under two different functional forms produced almost the same answer to a counterfactual query. This is no coincidence. In [Balke and Pearl, 1994], a method is devised for computing sharp bounds on counterfactual probabilities, and, under certain circumstances, those bounds may collapse to point estimates. This method has been applied to the evaluation of causal effects in studies involving noncompliance and to determination of legal liability.

Counterfactual reasoning is at the heart of many cognitive abilities, especially real-time planning. For example, when a planner discovers that the current state of affairs deviates from the one expected, a "plan repair" activity will be invoked to determine what went wrong and how the error can be rectified. This activity amounts to an exercise in counterfactual thinking, as it calls for rolling back the natural course of events and determining, based on the factual observations at hand, whether the culprit resides in previous decisions or in some unexpected, external eventualities. Moreover, in reasoning forward to determine whether things would have been different, a new model of the world must be consulted, one that embodies hypothetical changes in decisions or eventualities, hence, a breakdown of the old model or theory. The surgical semantics expounded in this section offers a formal account of such breakdown.

The capacity to mentally simulate theory breakdowns is required whenever one wishes to evaluate the merit of actions on the basis of the past performance. The odd statement: "Had I done my homework, I would have felt miserable, because I always do my homework after my father beats me up" demonstrates the consequences of failing to exercise this capacity. A person aware of the signals triggering the past actions, must devise a method for selectively ignoring the influence of those signals from the evaluation process. In fact, the very essence of *evaluation* is having the freedom to imagine and compare trajectories in various counterfactual worlds, where each world or trajectory is created by a hypothetical implementation of actions that are free of the very pressures that compelled the implementation of such actions in the past.

The task of inferring singular causes (Subsection 2.4), also requires counterfactual reasoning. Finding the probability that $X = x$ is the actual cause for effect E, also amounts to answering a counterfactual query: "Given effect E and observations O, find the probability that E would not have been realized, had X not been x." The technique developed in Balke and Pearl (1995) permits the evaluation of such queries in the framework of Definition 11.

4.6 Historical remarks

An explicit translation of interventions to "striking out" equations from linear econometric models was first proposed by Strotz and Wold (1960) and later used in Fisher (1970) and Sobel (1990). Extensions to action representation in nonmonotonic reasoning and statistical analysis were reported in [Goldszmidt and Pearl, 1992, Pearl, 1993]. Graphical ramifications of this translation were explicated first in Spirtes et al. (1993) and later in Pearl (1993b). A related formulation of causal effects, based on event trees and counterfactual analysis, was developed by Robins (1986, pp. 1422-1425). Shafer (1995) offers a novel formulation of probabilistic causation, based also on event trees. Calculi for actions and counterfactuals based on surgery semantics are developed in [Pearl, 1994] and [Balke and Pearl, 1994], respectively.

5 Conclusions

Statistical contingency models of causal induction have had two major advantages over their power-based rivals. First, statistics-based models are grounded in direct experience and, hence, promise to explicate the evidence and the processes responsible for acquiring cause-effect relationships from raw data. Second, statistics-based models enjoy the symbolic machinery of probability calculus, which enables researchers to posit hypotheses, communicate ideas, and make predictions with mathematical precision. In comparison, as well as skirting the issue of causal induction by presuming the pre-existence of a causal structure, power-based theories have lacked an adequate formal language in which to cast assumptions, claims, and predictions.

This chapter offers a formal setting, based on mechanisms, structures and surgeries, which accommodates both the statistical and the power components of causal inference. It has shown how pre-existing causal knowledge, cast qualitatively in the form of a graph, can combine with statistical data to produce new causal knowledge, that is both qualitative and quantitative in nature. It has also shown how the formal setting of structural causality provides not only a semantics for distinguishing subtle nuances in causal discourse, but also an inferential machinery for processing actions, observations, and counterfactuals.

Returning to the problem of induction, the question of how knowledge about mechanisms is acquired in the first place remains unanswered. Mechanisms, however, are nothing but ordinary physical laws, cast in the form of deterministic equations. Therefore, the acquisition of causal relationships is no different from the acquisition, using controlled experimentation, of physical laws such as Hooke's law of suspended springs or Newton's law of acceleration. The asymmetry associated with causal relations, which is normally absent from physical laws, is partly a by-product of the distinction we make between endogenous and exogenous variables, namely, between variables we choose to analyze within the system and those we prefer to take as given (see Simon, 1953), partly due to the distinction we perceive between manipulable and nonmanipulable variables, and partly due to inherent asymmetries induced when closed physical systems (described by symmetric equations) are placed in contact with powerful external influences, e.g., wetting the ground does not make the sprinkler turn on, moving cars do not turn ignition keys, and so on.

The explication of causal relationships in terms of mechanisms and physical laws is not meant to imply that the induction of physical laws is a solved, trivial task. It implies, however, that the problem of causal induction, once freed of the mysteries and suspicions that normally surround discussions of causality, can be formulated as part of the more familiar problem of scientific induction.

Acknowledgments

The research was partially supported by Air Force grant #F49620-94-1-0173, NSF grant #IRI-9420306, and Northrop/Rockwell Micro grant #94-100.

References

[Aldrich, 1994] J. Aldrich. Correlations genuine and spurious in Pearson and Yule. Technical report, University of Southampton, Department of Economics, UK, October 1994.

[Balke and Pearl, 1994] A. Balke and J. Pearl. Counterfactual probabilities: Computational methods, bounds, and applications. In R. Lopez de Mantaras and D. Poole, editors, *Uncertainty in Artificial Intelligence 10*, pages 46–54. Morgan Kaufmann, San Mateo, CA, 1994.

[Balke and Pearl, 1995] A. Balke and J. Pearl. Counterfactuals and policy analysis in structural models. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 11–18. Morgan Kaufmann, San Francisco, CA, 1995.

[Cartwright, 1989] N. Cartwright. *Nature's Capacities and Their Measurement.* Clarendon Press, Oxford, 1989.

[Cheng *et al.*, 1995] P.W. Cheng, J. Park, A. Yarlas, and K. Holyoak. *Psychology of Learning and Motivation*, chapter A causal-power theory of focal sets. 1995.

[Cheng, 1992] P.W. Cheng. Separating causal laws from causal facts: Pressing the limits of statistical relevance. *Psychology of Learning and Motivation*, 30:215–264, 1992.

[Cooper and Herskovits, 1990] G.F. Cooper and E. Herskovits. A Bayesian method for constructing Bayesian belief networks from databases. *Proceedings of the Conference on Uncertainty in AI*, pages 86–94, 1990.

[Cox, 1958] D.R. Cox. *The Planning of Experiments.* John Wiley and Sons, NY, 1958.

[Davis, 1988] W.A. Davis. Probabilistic theories of causation. In James H. Fetzer, editor, *Probability and Causality*, pages 133–160. D. Reidel, Dordrecht, 1988.

[Eells, 1991] E. Eells. *Probabilistic Causality.* Cambridge University Press, Cambridge, MA, 1991.

[Fisher, 1970] F.M. Fisher. A correspondence principle for simultaneous equations models. *Econometrica*, 38:73–92, 1970.

[Galles and Pearl, 1995] D. Galles and J. Pearl. Testing identifiability of causal effects. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 185–195. Morgan Kaufmann, San Francisco, CA, 1995.

[Geiger *et al.*, 1990] D. Geiger, T.S. Verma, and J. Pearl. Identifying independence in Bayesian networks. In *Networks*, volume 20, pages 507–534. John Wiley and Sons, Sussex, England, 1990.

[Geiger, 1990] D. Geiger. Graphoids: A qualitative framework for probabilistic inference. PhD thesis, University of California, Los Angeles, 1990.

[Goldszmidt and Pearl, 1992] M. Goldszmidt and J. Pearl. Default ranking: A practical framework for evidential reasoning, belief revision and update. *Proceedings of the Third International Conference on Knowledge Representation and Reasoning*, pages 661–672, 1992.

[Goldszmidt and Pearl, 1995] M. Goldszmidt and J. Pearl. Qualitative probabilities for default reasoning, belief revision, and causal modeling. Technical Report R-161-L, Computer Science Department, UCLA, 1995. Forthcoming in *Artificial Intelligence*.

[Good, 1961] I.J. Good. A causal calculus, I-II. *British Journal for the Philosophy of Science*, 11:305–318, 12:43–51, 1961. Errata and corrigenda, 13:88. Reprinted in I.J. Good's *Good Thinking*, University of Minnesota Press, Minnesota, 1983.

[Haavelmo, 1943] T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11:1–12, 1943.

[Heckerman and Shachter, 1995] D. Heckerman and R. Shachter. A definition and graphical representation for causality. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 262–273, San Mateo, CA, 1995. Morgan Kaufmann.

[Heckerman *et al.*, 1994] D. Heckerman, D. Geiger, and D. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence,* Seattle, WA, pages 293–301, San Mateo, CA, July 1994. Morgan Kaufmann.

[Jenkins and Ward, 1965] H. Jenkins and W. Ward. Judgement of contingency between responses and outcomes. *Psychological Monographs*, 7:1–17, 1965.

[Kim and Pearl, 1983] J.H. Kim and J. Pearl. A computational model for combined causal and diagnostic reasoning in inference systems. In *Proceedings IJCAI-83*, pages 190–193, Karlsruhe, Germany, 1983.

[Koopman and Reiersol, 1950] T.C. Koopman and O. Reiersol. The identification of structural characteristics. *Annals of Mathematical Statistics*, 21:165–181, 1950.

[Lauritzen and Spiegelhalter, 1988] S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems(with discussion). *Journal of the Royal Statistical Society, Series B*, 50(2):157–224, 1988.

[Lewis, 1973] D. Lewis. *Counterfactuals*. Basil Blackwell, Oxford, 1973.

[Mulaik, 1986] S.A. Mulaik. Toward a synthesis of deterministic and probabilistic formulations of causal relations by the functional relation concept. *Philosophy of Science*, 53:313–332, September 1986.

[Neyman, 1923] J. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–480, 1923.

[Otte, 1981] R. Otte. A critque of suppes' theory of probabilistic causality. *Synthese*, 48:167–189, 1981.

[Pearl and Verma, 1991] J. Pearl and T. Verma. A theory of inferred causation. In J.A. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 441–452, San Mateo, CA, 1991. Morgan Kaufmann.

[Pearl *et al.*, 1990] J. Pearl, D. Geiger, and T. Verma. The logic of influence diagrams. In R.M. Oliver and J.Q. Smith, editors, *Influence Diagrams, Belief Nets and Decision Analysis*, pages 67–87. John Wiley and Sons, Inc., New York, NY, 1990.

[Pearl, 1988] J. Pearl. *Probabilistic Reasoning in Intelligence Systems.* Morgan Kaufmann, San Mateo, CA, 1988. (Revised 2nd printing, 1992).

[Pearl, 1993] J. Pearl. From Bayesian networks to causal networks. In *Proceedings of the Adaptive Computing and Information Processing Seminar*, pages 25–27, Brunel Conference Centre, London, January 1993. See also Statistical Science, 8(3):, 266–269, 1993.

[Pearl, 1994] J. Pearl. A probabilistic calculus of actions. In R. Lopez de Mantaras and D. Poole, editors, *Uncertainty in Artificial Intelligence 10*, pages 454–462. Morgan Kaufmann, San Mateo, CA, 1994.

[Pearl, 1995] J. Pearl. Causal diagrams for experimental research. Technical Report R-218-B, Computer Science Department, UCLA, 1995. To appear in *Biometrika*, December 1995.

[Pearson, 1911] K. Pearson. *Grammar of Science, 3rd ed.* A. and C. Black Publishers, 1911.

[Pratt and Schlaifer, 1988] J.W. Pratt and R. Schlaifer. On the interpretation and observation of laws. *Journal of Econometrics*, 39:23–52, 1988.

[Reichenbach, 1956] H. Reichenbach. *The Direction of Time.* University of California Press, Berkeley, 1956.

[Robins, 1986] J.M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period - applications to control of the healthy workers survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.

[Rosenbaum and Rubin, 1983] P. Rosenbaum and D. Rubin. The central role of propensity score in observational studies for causal effects. *Biometrica*, 70:41–55, 1983.

[Rubin, 1974] D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[Russell, 1913] B. Russell. On the notion of cause. *Proceedings of the Aristotelian Society*, 13:1–26, 1913.

[Salmon, 1984] W.C. Salmon. *Scientific Explanation and the Causal Structure of the World.* Princeton University Press, Princeton, 1984.

[Salmon, 1994] W. Salmon. Causality without counterfactuals. *Philosophy of Science Association*, 61:297–312, 1994.

[Savage, 1954] L.J. Savage. *The Foundations of Statistics.* John Wiley and Sons, Inc., New York, 1954.

[Shafer, 1995] G. Shafer. *The Art of Causal Conjecture.* MIT Press, Cambridge, MA, 1995. Forthcoming.

[Shultz, 1982] T.R. Shultz. Rules of causal attribution. *Monographs of the Society for Research in Child Development*, 47(1), 1982.

[Simon, 1953] H.A. Simon. Causal ordering and identifiability. In W.C. Hood and T.C. Koopmans, editors, *Studies in Econometric Method*, pages 49–74. John Wiley and Sons, New York, 1953.

[Simpson, 1951] E.H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*,13:238–241, 1951.

[Skyrms, 1980] B. Skyrms. *Causal Necessity.* Yale University Press, New Haven, 1980.

[Smith, 1957] H.F. Smith. Interpretation of adjusted treatment means and regressions in analysis of covariates. *Biometrics*, 13:282–308, 1957.

[Sobel, 1990] M.E. Sobel. Effect analysis and causation in linear structural equation models. *Psychometrika*, 55(3):495–515, 1990.

[Spiegelhalter *et al.*, 1993] D.J. Spiegelhalter, S.L. Lauritzen, P.A. Dawid, and R.G. Cowell. Bayesian analysis in expert systems. *Statistical Science*, 8:219–247, 1993.

[Spirtes *et al.*, 1993] P. Spirtes, C. Glymour, and R. Schienes. *Causation, Prediction, and Search.* Springer-Verlag, New York, 1993.

[Spohn, 1980] W. Spohn. Stochastic independence, causal independence, and shieldability. *Journal of Philosophical Logic*, 9:73–99, 1980.

[Strotz and Wold, 1960] R.H. Strotz and H.O.A. Wold. Causal models in the social sciences. *Econometrica*, 28:417–427, 1960.

[Suppes, 1970] P. Suppes. *A Probabilistic Theory of Causation.* North-Holland, Amsterdam, 1970.

[Verma and Pearl, 1990] T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Uncertainty in Artificial Intelligence*, pages 6,220–227, Cambridge, MA, 1990. Elsevier Science Publishers.

[Wainer, 1991] H. Wainer. Adjusting for differential base-rates: Lord's paradox again. *Psychological Bulletin*, 109:147–151, 1991.

[Waldmann *et al.*, 1995] M.R. Waldmann, K.J. Holyoak, and A. Fratiannea. Causal models and the acquisition of category structure. *Journal of Experimental Psychology*, 124:181–206, 1995.

[Wright, 1921] S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.