JUDEA PEARL

BAYESIANISM AND CAUSALITY, OR, WHY I AM ONLY A HALF-BAYESIAN

1 INTRODUCTION

I turned Bayesian in 1971, as soon as I began reading Savage's monograph *The Foundations of Statistical Inference* [Savage, 1962]. The arguments were unassailable: (i) It is plain silly to ignore what we know, (ii) It is natural and useful to cast what we know in the language of probabilities, and (iii) If our subjective probabilities are erroneous, their impact will get washed out in due time, as the number of observations increases.

Thirty years later, I am still a devout Bayesian in the sense of (i), but I now doubt the wisdom of (ii) and I know that, in general, (iii) is false. Like most Bayesians, I believe that the knowledge we carry in our skulls, be its origin experience, schooling or hearsay, is an invaluable resource in all human activity, and that combining this knowledge with empirical data is the key to scientific enquiry and intelligent behavior. Thus, in this broad sense, I am a still Bayesian. However, in order to be combined with data, our knowledge must first be cast in some formal language, and what I have come to realize in the past ten years is that the language of probability is not suitable for the task; the bulk of human knowledge is organized around causal, not probabilistic relationships, and the grammar of probability calculus is insufficient for capturing those relationships. Specifically, the building blocks of our scientific and everyday knowledge are elementary facts such as "mud does not cause rain" and "symptoms do not cause disease" and those facts, strangely enough, cannot be expressed in the vocabulary of probability calculus. It is for this reason that I consider myself only a half-Bayesian.

In the rest of the paper, I plan to review the dichotomy between causal and statistical knowledge, to show the limitation of probability calculus in handling the former, to explain the impact that this limitation has had on various scientific disciplines and, finally, I will express my vision for future development in Bayesian philosophy: the enrichment of personal probabilities with causal vocabulary and causal calculus, so as to bring mathematical analysis closer to where knowledge resides.

2 STATISTICS AND CAUSALITY: A BRIEF SUMMARY

The aim of standard statistical analysis, typified by regression and other estimation techniques, is to infer parameters of a distribution from samples drawn of that population. With the help of such parameters, one can infer associations among variables, estimate the likelihood of past and future events, as well as update the likelihood of events in light of new evidence or new measurements. These tasks are managed well by statistical analysis so long as experimental conditions remain the same. Causal analysis goes one step further; its aim is to infer aspects of the data generation process. With the help of such aspects, one can deduce not only the likelihood of events under static conditions, but also the dynamics of events under *changing conditions*. This capability includes predicting the effect of actions (e.g., treatments or policy decisions), identifying causes of reported events, and assessing responsibility and attribution (e.g., whether event x was necessary (or sufficient) for the occurrence of event y).

Almost by definition, causal and statistical concepts do not mix. Statistics deals with behavior under uncertain, yet static conditions, while causal analysis deals with changing conditions. There is nothing in the joint distribution of symptoms and diseases to tell us that curing the former would not cure the latter. In general, there is nothing in a distribution function that would tell us how that distribution would differ if external conditions were to change—say from observational to experimental setup—every conceivable difference in the distribution would be perfectly compatible with the laws of probability theory, no matter how slight the change in conditions.[1]

Drawing analogy to visual perception, the information contained in a probability function is analogous to a precise description of a three-dimensional object; it is sufficient for predicting how that object will be viewed from any angle outside the object, but it is insufficient for predicting how the object will be viewed if manipulated and squeezed by external forces. The additional properties needed for making such predictions (e.g., the object's resilience or elasticity) is analogous to the information that causal models provide using the vocabulary of directed graphs and/or structural equations. The role of this information is to identify those aspects of the world that remain invariant when external conditions change, say due to an action.

These considerations imply that the slogan "correlation does not imply causation" can be translated into a useful principle: one cannot substantiate causal claims from associations alone, even at the population level—behind every causal conclusion there must lie some causal assumption that is not testable in observational studies. Nancy Cartwright [1989] expressed this principle as "no causes in, no causes out", meaning we cannot convert statistical knowledge into causal knowledge.

The demarcation line between causal and statistical concepts is thus clear and crisp. A statistical concept is any concept that can be defined in terms of a distribution (be it personal or frequency-based) of observed variables, and a causal con-

[1] Even the theory of stochastic processes, which provides probabilistic characterization of certain dynamic phenomena, assumes a fixed density function over time-indexed variables. There is nothing in such a function to tell us how it would be altered if external conditions were to change. If a parametric family of distributions is used, we can represent some changes by selecting a different set of parameters. But we are still unable to represent changes that do not correspond to parameter selection; for example, restricting a variable to a certain value, or forcing one variable to equal another.

cept is any concept concerning changes in variables that cannot be defined from the distribution alone. Examples of statistical concepts are: correlation, regression, dependence, conditional independence, association, likelihood, collapsibility, risk ratio, odd ratio, and so on.[2] Examples of causal concepts are: randomization, influence, effect, confounding, disturbance, spurious correlation, instrumental variables, intervention, explanation, attribution, and so on. The purpose of this demarcation line is not to exclude causal concepts from the province of statistical analysis but, rather, to make it easy for investigators and philosophers to trace the assumptions that are needed for substantiating various types of scientific claims. Every claim invoking causal concepts must be traced to some premises that invoke such concepts; it cannot be derived or inferred from statistical claims alone.

This principle may sound obvious, almost tautological, yet it has some far reaching consequences. It implies, for example, that any systematic approach to causal analysis must acquire new mathematical notation for expressing causal assumptions and causal claims. The vocabulary of probability calculus, with its powerful operators of conditionalization and marginalization, is simply insufficient for expressing causal information. To illustrate, the syntax of probability calculus does not permit us to express the simple fact that "symptoms do not cause diseases", let alone draw mathematical conclusions from such facts. All we can say is that two events are dependent—meaning that if we find one, we can expect to encounter the other, but we cannot distinguish statistical dependence, quantified by the conditional probability $P(disease|symptom)$ from causal dependence, for which we have no expression in standard probability calculus.[3] Scientists seeking to express causal relationships must therefore supplement the language of probability with a vocabulary for causality, one in which the symbolic representation for the relation "symptoms cause disease" is distinct from the symbolic representation of "symptoms are associated with disease." Only after achieving such a distinction can we label the former sentence "false," and the latter "true."

The preceding two requirements: (1) to commence causal analysis with untested,[4] judgmentally based assumptions, and (2) to extend the syntax of probability calculus, constitute, in my experience, the two main obstacles to the acceptance of causal analysis among statisticians, philosophers and professionals with traditional training in statistics. We shall now explore in more detail the nature of these two barriers, and why they have been so tough to cross.

[2]The term 'risk ratio' and 'risk factors' have been used ambivalently in the literature; some authors insist on a risk factor having causal influence on the outcome, and some embrace factors that are merely associated with the outcome.

[3]Attempts to define causal dependence by conditioning on the entire past (e.g., Suppes, 1970) violate the statistical requirement of limiting the analysis to "observed variables", and encounter other insurmountable difficulties (see Eells [1991], Pearl [2000a], pp. 249-257).

[4]By "untested" I mean untested using frequency data in nonexperimental studies.

2.1 The Barrier of Untested Assumptions

All statistical studies are based on some untested assumptions. For examples, we often assume that variables are multivariate normal, that the density function has certain smoothness properties, or that a certain parameter falls in a given range. The question thus arises why innocent causal assumptions, say, that symptoms do not cause disease or that mud does not cause rain, invite mistrust and resistance among statisticians, especially of the Bayesian school.

There are three fundamental differences between statistical and causal assumptions. First, statistical assumptions, even untested, are testable in principle, given sufficiently large sample and sufficiently fine measurements. Causal assumptions, in contrast, cannot be verified even in principle, unless one resorts to experimental control. This difference is especially accentuated in Bayesian analysis. Though the priors that Bayesians commonly assign to statistical parameters are untested quantities, the sensitivity to these priors tends to diminish with increasing sample size. In contrast, sensitivity to priors of causal parameters, say those measuring the effect of smoking on lung cancer, remains non-zero regardless of (nonexperimental) sample size.

Second, statistical assumptions can be expressed in the familiar language of probability calculus, and thus assume an aura of scholarship and scientific respectability. Causal assumptions, as we have seen before, are deprived of that honor, and thus become immediate suspect of informal, anecdotal or metaphysical thinking. Again, this difference becomes illuminated among Bayesians, who are accustomed to accepting untested, judgmental assumptions, and should therefore invite causal assumptions with open arms—they don't. A Bayesian is prepared to accept an expert's judgment, however esoteric and untestable, so long as the judgment is wrapped in the safety blanket of a probability expression. Bayesians turn extremely suspicious when that same judgment is cast in plain English, as in "mud does not cause rain." A typical example can be seen in Lindley and Novick's [1981] treatment of Simpson's paradox.

Lindley and Novick showed that decisions on whether to use conditional or marginal contingency tables should depend on the story behind the tables, that is, on one's assumption about how the tables were generated. For example, to decide whether a treatment $X = x$ is beneficial ($Y = y$) in a population, one should compare $\Sigma_z P(y|x, z)$ to $\Sigma_z P(y|x', z)$ if Z stands for the gender of patients. In contrast, if Z stands for a factor that is affected by the treatment (say blood pressure), one should compare the marginal probabilities, $P(y|x)$ vis-$\grave{a}$-vis $P(y|x')$, and refrain from conditioning on Z (see [Pearl, 2000a; pp. 174-182] for details). Remarkably, instead of attributing this difference to the causal relationships in the story, Lindley and Novick wrote: "We have not chosen to do this; nor to discuss causation, because the concept, although widely used, does not seem to be well-defined" (p. 51). Thus, instead of discussing causation, they attribute the change in strategy to another untestable relationship in the story—exchangeability [DeFinetti, 1974] which is cognitively formidable yet, at least formally, can be cast in a probability expression. In Section 4.2, we will return to discuss this trend among Bayesians of equating "definability" with expressibility in probabilistic language.

The third resistance to causal (vis-à-vis statistical) assumptions stems from their intimidating clarity. Assumptions about abstract properties of density functions or about conditional independencies among variables are, cognitively speaking, rather opaque, hence they tend to be forgiven, rather than debated. In contrast, assumptions about how variables cause one another are shockingly transparent, and tend therefore to invite counter-arguments and counter-hypotheses. A co-reviewer on a paper I have read recently offered the following objection to the causal model postulated by the author:

> "A thoughtful and knowledgeable epidemiologist could write down two or more equally plausible models that leads to different conclusions regarding confounding."

Indeed, since the bulk of scientific knowledge is organized in causal schema, scientists are incredibly creative in constructing competing alternatives to any causal hypothesis, however plausible. Statistical hypotheses in contrast, having been several levels removed from our store of knowledge, are relatively protected from such challenges.

I conclude this subsection with a suggestion that statisticians' suspicion of causal assumptions, vis-à-vis probabilistic assumptions, is unjustified. Considering the organization of scientific knowledge, it makes prefect sense that we permit scientists to articulate what they know in plain causal expressions, and not force them to compromise reliability by converting to the "higher level" language of prior probabilities, conditional independence and other cognitively unfriendly terminology.[5]

2.2 *The Barrier of New Notation*

If reluctance to making causal assumptions has been a hindrance to causal analysis, finding a mathematical way of expressing such assumptions encountered a formidable mental block. The need to adopt a new notation, foreign to the province of probability theory, has been traumatic to most persons trained in statistics; partly because the adaptation of a new language is difficult in general, and partly because statisticians have been accustomed to assuming that all phenomena, processes, thoughts, and modes of inference can be captured in the powerful language of probability theory.[6]

[5]Similar observations were expressed by J. Heckman [2001].

[6]Commenting on my $set(x)$ notation [Pearl, 1995a, b], a leading statistician wrote: "Is this a concept in some new theory of probability or expectation? If so, please provide it. Otherwise, 'metaphysics' may remain the leading explanation." Another statistician, commenting on the $do(x)$ notation used in *Causality* [Pearl, 2000a], insisted: "...the calculus of probability is the calculus of causality."

Not surprisingly, in the bulk of the statistical literature, causal claims never appear in the mathematics. They surface only in the verbal interpretation that investigators occasionally attach to certain associations, and in the verbal description with which investigators justify assumptions. For example, the assumption that a covariate is not affected by a treatment, a necessary assumption for the control of confounding [Cox, 1958], is expressed in plain English, not in a mathematical equation.

In some applications (e.g., epidemiology), the absence of notational distinction between causal and statistical dependencies seemed unnecessary, because investigators were able to keep such distinctions implicitly in their heads, and managed to confine the mathematics to conventional probability expressions. In others, as in economics and the social sciences, investigators rebelled against this notational tyranny by leaving mainstream statistics and constructing their own mathematical machinery (called Structural Equations Models). Unfortunately, this machinery has remained a mystery to outsiders, and eventually became a mystery to insiders as well.[7]

But such tensions could not remain dormant forever. "Every science is only so far exact as it knows how to express one thing by one sign," wrote Augustus de Morgan in 1858 — the harsh consequences of not having the signs for expressing causality surfaced in the 1980-90's. Problems such as the control of confounding, the estimation of treatment effects, the distinction between direct and indirect effects, the estimation of probability of causation, and the combination of experimental and nonexperimental data became a source of endless disputes among the users of statistics, and statisticians could not come to the rescue. [Pearl, 2000a] describes several such disputes, and why they could not be resolved by conventional statistical methodology.

3 LANGUAGES FOR CAUSAL ANALYSIS

3.1 The language of diagrams and structural equations

How can one express mathematically the common understanding that symptoms do not cause diseases? The earliest attempt to formulate such relationship mathematically was made in the 1920's by the geneticist Sewall Wright [1921]. Wright used a combination of equations and graphs to communicate causal relationships. For example, if X stands for a disease variable and Y stands for a certain symptom of the disease, Wright would write a linear equation:

(1) $y = ax + u$

supplemented with the diagram $X \longrightarrow Y$, where x stands for the level (or severity) of the disease, y stands for the level (or severity) of the symptom, and u stands

[7] Most econometric texts in the last decade have refrained from defining what an economic model is, and those that attempted a definition, erroneously view structural equations models as compact representations of probability density functions (see [Pearl, 2000a, pp. 135-138]).

for all factors, other than the disease in question, that could possibly affect Y (U is called "exogenous", "background", or "disturbance".) The diagram encodes the possible existence of (direct) causal influence of X on Y, and the absence of causal influence of Y on X, while the equation encodes the quantitative relationships among the variables involved, to be determined from the data. The parameter a in the equation is called a "path coefficient" and it quantifies the (direct) causal effect of X on Y; given the numerical value of a, the equation claims that, ceteras paribus, a unit increase in X would result in an a-unit increase of Y. If correlation between X and U is presumed possible, it is customary to add a double arrow between X and Y.

The asymmetry induced by the diagram renders the equality sign in Eq. (1) different from algebraic equality, resembling instead the assignment symbol (:=) in programming languages. Indeed, the distinctive characteristic of structural equations, setting them apart from algebraic equations, is that they stand for a value-assignment process — an autonomous mechanism by which the value of Y (not X) is determined. In this assignment process, Y is committed to track changes in X, while X is not subject to such commitment.[8]

Wright's major contribution to causal analysis, aside from introducing the language of path diagrams, has been the development of graphical rules for writing down (by inspection) the covariance of any pair of observed variables in terms of path coefficients and of covariances among disturbances. Under certain causal assumptions, (e.g. if $Cov(U, X) = 0$), the resulting equations may allow one to solve for the path coefficients in terms of observed covariance terms only, and this amounts to inferring the magnitude of (direct) causal effects from observed, non-experimental associations, assuming of course that one is prepared to defend the causal assumptions encoded in the diagram.

The causal assumptions embodied in the diagram (e.g, the absence of arrow from Y to X, or $Cov(U, X) = 0$) are not generally testable from nonexperimental data. However, the fact that each causal assumption in isolation cannot be tested does not mean that the sum total of all causal assumptions in a model does not have testable implications. The chain model $X \longrightarrow Y \longrightarrow Z$ for example, encodes seven causal assumptions, each corresponding to a missing arrow or a missing double-arrow between a pair of variables. None of those assumptions is testable in isolation, yet the totality of all those assumptions implies that Z is unassociated with X, conditioned on Y. Such testable implications can be read off the diagrams (see [Pearl 2000a, pp. 16–19]), and these constitute the only opening through which the assumption embodies in structural equation models can be tested in observational studies. Every conceivable statistical test that can be applied to the model is entailed by those implications.

[8]Clearly, if we intervene on X, Y would continue to track changes in X. Not so when we intervene on Y, X will reman unchanged. Such intervention (on Y) would alter the assignment mechanism for Y and, naturally, would cause the equality in Eq. (1) to be violated.

3.2 From path-diagrams to do-calculus

Structural equation modeling (SEM) has been the main vehicle for causal analysis in economics, and the behavioral and social sciences [Goldberger 1972; Duncan 1975]. However, the bulk of SEM methodology was developed for linear analysis and, until recently, no comparable methodology has been devised to extend its capabilities to models involving discrete variables, nonlinear dependencies, or situations in which the functional form of the equations is unknown. A central requirement for any such extension is to detach the notion of "effect" from its algebraic representation as a coefficient in an equation, and redefine "effect" as a general capacity to transmit *changes* among variables. One such extension, based on simulating hypothetical interventions in the model, is presented in Pearl [1995a, 2000a]

The central idea is to exploit the invariant characteristics of structural equations without committing to a specific functional form. For example, the non-parametric interpretation of the chain model $Z \longrightarrow X \longrightarrow Y$ corresponds to a set of three functions, each corresponding to one of the variables:

$$\begin{aligned} z &= f_Z(w) \\ x &= f_X(z, v) \\ y &= f_Y(x, u) \end{aligned} \tag{2}$$

together with the assumption that the background variables W,V, U (not shown in the chain) are jointly independent but, otherwise, arbitrarily distributed. Each of these functions represents a causal process (or mechanism) that determines the value of the left variable (output) from those on the right variables (input). The absence of a variable from the right hand side of an equation encodes the assumption that it has no direct effect on the left variable. For example, the absence of variable Z from the arguments of f_Y indicates that variations in Z will leave Y unchanged, as long as variables U and X remain constant. A system of such functions are said to be *structural* (or *modular*) if they are assumed to be autonomous, that is, each function is invariant to possible changes in the form of the other functions [Simon 1953; Koopmans 1953].

This feature of invariance permits us to use structural equations as a basis for modeling actions and counterfactuals. This is done through a mathematical operator called $do(x)$ which simulates physical interventions by deleting certain functions from the model, replacing them by constants, while keeping the rest of the model unchanged. For example, to represent an intervention that sets the value of X to x_0 the model for Eq. (2) would become

$$\begin{aligned} z &= f_Z(w) \\ x &= x_0 \\ y &= f_Y(x, u) \end{aligned} \tag{3}$$

The distribution of Y and Z calculated from this modified model characterizes the effect of the action $do(X = x_0)$ and is denoted as $P(y, z|do(x_0))$. It is not hard to show that, as expected, the model of Eq. (2) yields $P(y|do(x_0)) = P(y|x_0)$ and $P(z|do(x_0)) = P(z)$ regardless of the functions f_X, f_Y and f_Z. The general rule is simply to remove from the factorized distribution $P(x,y,z) = P(z)P(x|z)P(y|x)$ the factor that corresponds to the manipulated variable (X in our example) and to substitute the new value of that variable (x_0 in our example) into the truncated expression — the resulting expression then gives the post-intervention distribution of the remaining variables [Pearl, 2000a; section 3.2]. Additional features of this transformation are discussed in the Appendix; see [Pearl, 2000a; chapter 7] for full details.

The main task of causal analysis is to infer causal quantities from two sources of information: (i) the assumptions embodied in the model, and (ii) the observed distribution $P(x,y,z)$, or from samples of that distribution. Such analysis requires mathematical means of transforming causal quantities, represented by expressions such as $P(y|do(x))$, into *do*-free expressions derivable from $P(z,x,y)$, since only *do*-free expressions are estimable from non-experimental data. When such a transformation is feasible, we say that the causal quantity is *identifiable*. A calculus for performing such transformations, called *do*-calculus, was developed in [Pearl, 1995a]. Remarkably, the rules governing this calculus depend merely on the topology of the diagram; it takes no notice of the functional form of the equations, nor of the distribution of the disturbance terms. This calculus permits the investigator to inspect the causal diagram and

1. Decide whether the assumptions embodied in the model are sufficient to obtain consistent estimates of the target quantity;

2. Derive (if the answer to item 1 is affirmative) a closed-form expression for the target quantity in terms of distributions of observed quantities; and

3. Suggest (if the answer to item 1 is negative) a set of observations and experiments that, if performed, would render a consistent estimate feasible.

4 ON THE DEFINITION OF CAUSALITY

In this section, I return to discuss concerns expressed by some Bayesians that causality is an undefined concept and that, although the *do*-calculus can be an effective mathematical tool in certain tasks, it does not bring us closer to the deep and ultimate understanding of causality, one that is based solely on classical probability theory.

4.1 Is causality reducible to probabilities?

Unfortunately, aspirations for reducing causality to probability are both untenable and unwarranted. Philosophers have given up such aspirations twenty years ago, and were forced to admit extra-probabilistic primitives (such as "counterfactuals" or "causal relevance") into the analysis of causation (see Eells [1991] and Pearl [2000a, Section 7.5]). The basic reason was alluded to in Section 2: probability theory deals with beliefs about an uncertain, yet static world, while causality deals with changes that occur in the world itself, (or in one's theory of such changes). More specifically, causality deals with how probability functions change in response to influences (e.g., new conditions or interventions) that originate from outside the probability space, while probability theory, even when given a fully specified joint density function on all (temporally-indexed) variables in the space, cannot tell us how that function would change under such external influences. Thus, "doing" is not reducible to "seeing", and there is no point trying to fuse the two together.

Many philosophers have aspired to show that the calculus of probabilities, endowed with a time dynamic, would be sufficient for causation [Suppes, 1970]. A well known demonstration of the impossibility of such reduction (following Otte [1981]) goes as follows. Consider a switch X that turns on two lights, Y and Z, and assume that, due to differences in location, Z turns on a split second before Y. Consider now a variant of this example where the switch X activates Z, and Z, in turns, activates Y. This case is probabilistically identical to the previous one, because all functional and temporal relationships are identical. Yet few people would perceive the causal relationships to be the same in the two situations; the latter represents cascaded process, $X \longrightarrow Z \longrightarrow Y$, while the former represents a branching process, $Y \longleftarrow X \longrightarrow Z$. The difference shows, of course, when we consider interventions; intervening on Z would affect Y in the cascaded case, but not in the branching case.

The preceding example illustrates the essential role of *mechanisms* in defining causation. In the branching case, although all three variables are symmetrically constrained by the functional relationships: $X = Y$, $X = Z$, $Z = Y$, these relationships in themselves do not reveal the information that the three equalities are sustained by only two mechanisms, $Y = X$ and $Z = X$, and that the first equality would still be sustained when the second is violated. A set of mechanisms, each represented by an equation, is not equivalent to the set of algebraic equations that are implied by those mechanisms. Mathematically, the latter is defined as *one* set of n equations, whereas the former is defined as n separate sets, each containing one equation. These are two distinct mathematical objects that admit two distinct types of solution-preserving operations. The calculus of causality deals with the dynamics of such modular systems of equations, where the addition and deletion of equations represent interventions (see Appendix).

4.2 Is causality well-defined?

From a mathematical perspective, it is a mistake to say that causality is undefined. The *do*-calculus, for example, is based on two well-defined mathematical objects: a probability function P and a directed acyclic graph (DAG) D; the first is standard in statistical analysis while the second is a newcomer that tells us (in a qualitative, yet formal language) which mechanisms would remain invariant to a given intervention. Given these two mathematical objects, the definition of "cause" is clear and crisp; variable X is a *probabilistic-cause* of variable Y if $P(y|do(x)) \neq P(y)$ for some values x and y. Since each of $P(y|do(x))$ and $P(y)$ is well-defined in terms of the pair (P, D), the relation "probabilistic cause" is, likewise, well-defined. Similar definitions can be constructed for other nuances of causal discourse, for example, "causal effect", "direct cause", "indirect cause", "event-to-event cause", "scenario-specific cause", "necessary cause", "sufficient cause", "likely cause" and "actual cause" (see [Pearl, 2000a, pp. 222–3, 286–7, 319]; some of these definitions invoke functional models).

Not all statisticians/philosophers are satisfied with these mathematical definitions. Some suspect definitions that are based on unfamiliar non-algebraic objects (i.e., the DAG) and some mistrust abstract definitions that are based on unverifiable models. Indeed, no mathematical machinery can ever verify whether a given DAG really represents the causal mechanisms that generate the data — such verification is left either to human judgment or to experimental studies that invoke interventions. I submit, however, that neither suspicion nor mistrust are justified in the case at hand; DAGs are no less formal than mathematical equations, and questions of model verification need be kept apart from those of conceptual definition.

Consider, for example, the concept of a distribution *mean*. Even non-Bayesians perceive this notion to be well-defined, for it can be computed from any given (non-pathological) distribution function, even before ensuring that we can estimate that distribution from the data. We would certainly not declare the mean "ill-defined" if, for any reason, we find it hard to estimate the distribution from the available data. Quite the contrary; by defining the mean in the abstract, as a functional of *any* hypothetical distribution, we can often prove that the defining distribution need not be estimated at all, and that the mean can be estimated (consistently) directly from the data. Remarkably, by taking seriously the abstract (and untestable) notion of a distribution, we obtain a license to ignore it. An analogous logic applies to causation. Causal quantities are first defined in the abstract, using the pair (P, D), and this abstract definition then provides a theoretical framework for deciding, given the type of data available, which of the assumptions embodied in the DAG are ignorable, and which are absolutely necessary for establishing the target causal quantity from the data.[9]

The separation between concept definition and model verification is even more pronounced in the Bayesian framework, where purely judgmental concepts, such as the prior distribution of the mean, are perfectly acceptable, as long as they can be assessed reliably from one's experience or knowledge. Dennis Lindley has remarked recently (personal communication) that "causal mechanisms may be easier

[9]I have used a similar logic in defense of counterfactuals [Pearl, 2000a], which Dawid [2000] deemed dangerous on account of being untestable. (See, also Dawid [2001], this volume.) Had Bernoulli been constrained by Dawid's precautions, the notion of a "distribution" would have had to wait for another "dangerous" scientist, of Bernoulli's equal, to be created.

to come by than one might initially think". Indeed, from a Bayesian perspective, the newcomer concept of a DAG is not an alien at all — it is at least as legitimate as the probability assessments that a Bayesian decision-maker pronounces in constructing a decision tree. In such construction, the probabilities that are assigned to branches emanating from a decision variable X correspond to assessments of $P(y|do(x))$ and those assigned to branches emanating from a chance variable X correspond to assessments of $P(y|x)$. If a Bayesian decision-maker is free to assess $P(y|x)$ and $P(y|do(x))$ in any way, as separate evaluations, the Bayesian should also be permitted to express his/her conception of the mechanisms that entail those evaluations. It is only by envisioning these mechanisms that a decision maker can generate a coherent list of such a vast number of $P(y|do(x))$ type assessments.[10] The structure of the DAG can certainly be recovered from judgments of the form $P(y|do(x))$ and, conversely, the DAG combined with a probability function P dictates all judgments of the form $P(y|do(x))$. Accordingly the structure of the DAG can be viewed as a qualitative parsimonious scheme of encoding and maintaining coherence among those assessments. And there is no need to translate the DAG into the language of probabilities to render the analysis legitimate. Adding probabilistic veneer to the mechanisms portrayed in the DAG may make the do calculus appear more traditional, but would not change the fact that the objects of assessment are still causal mechanisms, and that these objects have their own special grammar of generating predictions about the effect of actions. In summary, recalling the ultimate Bayesian mission of fusing judgment with data, it is not the language in which we cast judgments that legitimizes the analysis, but whether those judgments can reliably be assessed from our store of knowledge and from the peculiar form in which this knowledge is organized.

If it were not for this concern to maintain reliability (of judgment), one could easily translate the information conveyed in a DAG into purely probabilistic formulae, using hypothetical variables. (Translation rules are provided in [Pearl, 2000a, p. 232]). Indeed, this is how the potential-outcome approach of Neyman [1923] and Rubin [1974] has achieved statistical legitimacy: judgments about causal relationships among observables are expressed as statements about probability functions that involve mixtures of observable and counterfactual variables. The difficulty with this approach, and the main reason for its slow acceptance in statistics, is that judgments about counterfactuals are much harder to assess than judgments about causal mechanisms. For instance, to communicate the simple assumption that symptoms do not cause diseases, we would have to use a rather roundabout expression and say that the probability of the counterfactual event "disease had symptoms been absent" is equal to the probability of "disease had symptoms been present". Judgments of conditional independencies among such counterfactual events are even harder for researchers to comprehend or to evaluate.

[10] Coherence requires, for example, that for any x, y, and z, the inequality $P(y|do(x), do(z)) \geq P(y, x|do(z))$ be satisfied. This follows from the property of composition (see Appendix, Eq. (6), or [Pearl, 2000a; pp. 229]

5 SUMMARY

This paper calls attention to a basic conflict between mission and practice in Bayesian methodology. The mission is to express prior knowledge mathematically and reliably so as to assist the interpretation of data, hence the acquisition of new knowledge. The practice has been to express prior knowledge as prior probabilities — too crude a vocabulary, given the grand mission. Considerations of reliability (of judgment) call for enriching the language of probabilities with causal vocabulary and for admitting causal judgments into the Bayesian repertoire. The mathematics for interpreting causal judgments has matured, and tools for using such judgments in the acquisition of new knowledge have been developed. The grounds are now ready for mission-oriented Bayesianism.

APPENDIX

CAUSAL MODELS, ACTIONS AND COUNTERFACTUALS

This appendix presents a brief summary of the structural-equation semantics of causation and counterfactuals as defined in Balke and Pearl [1995], Galles and Pearl [1997, 1998], and Halpern [1998]. For detailed exposition of the structural account and its applications see [Pearl, 2000a].

Causal models are generalizations of the structural equations used in engineering, biology, economics and social science.[11] World knowledge is represented as a modular collection of stable and autonomous relationships called "mechanisms", each represented as a function, and changes due to interventions or unmodelled eventualities are treated as local modifications of these functions.

A causal model is a mathematical object that assigns truth values to sentences involving causal relationships, actions, and counterfactuals. We will first define causal models, then discuss how causal sentences are evaluated in such models. We will restrict our discussion to recursive (or feedback-free) models; extensions to non-recursive models can be found in Galles and Pearl [1997, 1998] and Halpern [1998].

DEFINITION 1 (Causal model).
A *causal model* is a triple

$$M = \langle U, V, F \rangle$$

where

(i) U is a set of variables, called *exogenous*. (These variables will represent background conditions, that is, variables whose values are determined outside the model.)

[11] Similar models, called "neuron diagrams" [Lewis, 1986, p. 200; Hall, 1998] are used informally by philosophers to illustrate chains of causal processes.

(ii) V is an ordered set $\{V_1, V_2, \ldots, V_n\}$ of variables, called *endogenous*. (These represent variables that are determined in the model, namely, by variables in $U \cup V$.)

(iii) F is a set of functions $\{f_1, f_2, \ldots, f_n\}$ where each f_i is a mapping from $U \times (V_1 \times \ldots \times V_{i-1})$ to V_i. In other words, each f_i tells us the value of V_i given the values of U and all predecessors of V_i. Symbolically, the set of equations F can be represented by writing[12]

$$v_i = f_i(pa_i, u_i) \;\; i = 1, \ldots, n$$

where pa_i is any realization of the unique minimal set of variables PA_i in V (connoting *parents*) sufficient for representing f_i.[13] Likewise, $U_i \subseteq U$ stands for the unique minimal set of variables in U that is sufficient for representing f_i.

Every causal model M can be associated with a directed graph, $G(M)$, in which each node corresponds to a variable in V and the directed edges point from members of PA_i toward V_i (by convention, the exogenous variables are usually not shown explicitly in the graph). We call such a graph the *causal graph* associated with M. This graph merely identifies the endogenous variables PA_i that have direct influence on each V_i but it does not specify the functional form of f_i.

For any causal model, we can define an *action* operator, $do(x)$, which, from a conceptual viewpoint, simulates the effect of external action that sets the value of X to x and, from a formal viewpoint, transforms the model into a *submodel*, that is, a causal model containing fewer functions.

DEFINITION 2 (Submodel).
Let M be a causal model, X be a set of variables in V, and x be a particular assignment of values to the variables in X. A submodel M_x of M is the causal model

$$M_x = \langle U, V, F_x \rangle$$

where

(4) $\quad F_x = \{f_i : V_i \notin X\} \cup \{X = x\}$

In words, F_x is formed by deleting from F all functions f_i corresponding to members of set X and replacing them with the set of constant functions $X = x$.

If we interpret each function f_i in F as an independent physical mechanism and define the action $do(X = x)$ as the minimal change in M required to make

[12] We use capital letters (e.g., X, Y) as names of variables and sets of variables, and lower-case letters (e.g., x, y) for specific values (called realizations) of the corresponding variables.

[13] A set of variables X is *sufficient* for representing a given function $y = f(x, z)$ if f is trivial in Z—that is, if for every x, z, z' we have $f(x, z) = f(x, z')$.

$X = x$ hold true under any u, then M_x represents the model that results from such a minimal change, since it differs from M by only those mechanisms that directly determine the variables in X. The transformation from M to M_x modifies the algebraic content of F, which is the reason for the name *modifiable structural equations* used in [Galles and Pearl, 1998].[14]

DEFINITION 3 (Effect of action).
Let M be a causal model, X be a set of variables in V, and x be a particular realization of X. The *effect of action* $do(X = x)$ on M is given by the submodel M_x.

DEFINITION 4 (Potential response).
Let Y be a variable in V, let X be a subset of V, and let u be a particular value of U. The *potential response* of Y to action $do(X = x)$ in situation u, denoted $Y_x(u)$, is the (unique) solution for Y of the set of equations F_x.

We will confine our attention to actions in the form of $do(X = x)$. Conditional actions, of the form "$do(X = x)$ if $Z = z$" can be formalized using the replacement of equations by functions of Z, rather than by constants [Pearl, 1994]. We will not consider disjunctive actions, of the form "$do(X = x \text{ or } X = x')$", since these complicate the probabilistic treatment of counterfactuals.

DEFINITION 5 (Counterfactual).
Let Y be a variable in V, and let X be a subset of V. The counterfactual expression "The value that Y would have obtained, had X been x" is interpreted as denoting the potential response $Y_x(u)$.

Definition 5 thus interprets the counterfactual phrase "had X been x" in terms of a hypothetical external action that modifies the actual course of history and imposes the condition "$X = x$" with minimal change of mechanisms. This is a crucial step in the semantics of counterfactuals [Balke and Pearl, 1994], as it permits x to differ from the actual value $X(u)$ of X without creating logical contradiction; it also suppresses abductive inferences (or backtracking) from the counterfactual antecedent $X = x$.[15]

It can be shown [Galles and Pearl, 1997] that the counterfactual relationship just defined, $Y_x(u)$, satisfies the following two properties:
Effectiveness:
For any two disjoint sets of variables, Y and W, we have

(5) $Y_{yw}(u) = y$.

[14] Structural modifications date back to Marschak [1950] and Simon [1953]. An explicit translation of interventions into "wiping out" equations from the model was first proposed by Strotz and Wold [1960] and later used in Fisher [1970], Sobel [1990], Spirtes et al. [1993], and Pearl [1995]. A similar notion of sub-model is introduced in Fine [1985], though not specifically for representing actions and counterfactuals.

[15] Simon and Rescher [1966, p. 339] did not include this step in their account of counterfactuals and noted that backward inferences triggered by the antecedents can lead to ambiguous interpretations.

In words, setting the variables in W to w has no effect on Y, once we set the value of Y to y.

Composition:

For any two disjoint sets of variables X and W, and any set of variables Y,

(6) $W_x(u) = w \Longrightarrow Y_{xw}(u) = Y_x(u).$

In words, once we set X to x, setting the variables in W to the same values, w, that they would attain (under x) should have no effect on Y. Furthermore, effectiveness and composition are $complete$ whenever M is recursive (i.e., $G(M)$ is acyclic) [Galles and Pearl, 1998; Halpern, 1998], that is, every property of counterfactuals that follows from the structural model semantics can be derived by repeated application of effectiveness and composition.

A corollary of composition is a property called $consistency$ by [Robins, 1987]:

(7) $(X(u) = x) \Longrightarrow (Y_x(u) = Y(u))$

Consistency states that, if in a certain context u we find variable X at value x, and we intervene and set X to that same value, x, we should not expect any change in the response variable Y. Composition and consistency are used in several derivations of Section 3.

The structural formulation generalizes naturally to probabilistic systems, as is seen below.

DEFINITION 6 (Probabilistic causal model).
A probabilistic causal model is a pair

$$\langle M, P(u)\rangle$$

where M is a causal model and $P(u)$ is a probability function defined over the domain of U.

$P(u)$, together with the fact that each endogenous variable is a function of U, defines a probability distribution over the endogenous variables. That is, for every set of variables $Y \subseteq V$, we have

(8) $P(y) \stackrel{\Delta}{=} P(Y = y) = \sum_{\{u \mid Y(u)=y\}} P(u)$

The probability of counterfactual statements is defined in the same manner, through the function $Y_x(u)$ induced by the submodel M_x. For example, the *causal effect* of X on Y is defined as:

(9) $P(Y_x = y) = \sum_{\{u \mid Y_x(u)=y\}} P(u)$

Likewise, a probabilistic causal model defines a joint distribution on counterfactual statements, i.e., $P(Y_x = y, Z_w = z)$ is defined for any sets of variables Y, X, Z, W, not necessarily disjoint. In particular, $P(Y_x = y, X = x')$ and $P(Y_x = y, Y_{x'} = y')$ are well defined for $x \neq x'$, and are given by

$$(10)\quad P(Y_x = y, X = x') = \sum_{\{u | Y_x(u)=y \ \& \ X(u)=x'\}} P(u)$$

and

$$(11)\quad P(Y_x = y, Y_{x'} = y') = \sum_{\{u \ | \ Y_x(u)=y \ \& \ Y_{x'}(u)=y'\}} P(u).$$

When x and x' are incompatible, Y_x and $Y_{x'}$ cannot be measured simultaneously, and it may seem meaningless to attribute probability to the joint statement "Y would be y if $X = x$ *and* Y would be y' if $X = x'$." Such concerns have been a source of recent objections to treating counterfactuals as jointly distributed random variables [Dawid, 2000]. The definition of Y_x and $Y_{x'}$ in terms of two distinct submodels, driven by a standard probability space over U, demonstrates that joint probabilities of counterfactuals have solid mathematical and conceptual underpinning and, moreover, these probabilities can be encoded rather parsimoniously using $P(u)$ and F.

Computer Science Department, University of California, USA.

BIBLIOGRAPHY

[Balke and Pearl, 1994] A. Balke and J. Pearl. Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, volume I, pages 230–237. MIT Press, Menlo Park, 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, 1995.

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

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

[Dawid, 2000] A.P. Dawid. Causal inference without counterfactuals (with comments and rejoinder). *Journal of the American Statistical Association*, 95(450):407–448, June 2000.

[DeFinetti, 1974] B. DeFinetti. *Theory of Probability: A Critical Introductory Treatment,* 2 volumes (Translated by A. Machi and A. Smith). Wiley, London, 1974.

[Duncan, 1975] O.D. Duncan. *Introduction to Structural Equation Models*. Academic Press, New York, 1975.

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

[Fine, 1985] K. Fine. *Reasoning with Arbitrary Objects*. B. Blackwell, New York, 1985.

[Fisher, 1970] F.M. Fisher. A correspondence principle for simultaneous equations models. *Econometrica*, 38(1):73–92, January 1970.

[Galles and Pearl, 1997] D. Galles and J. Pearl. Axioms of causal relevance. *Artificial Intelligence*, 97(1-2):9–43, 1997.

[Galles and Pearl, 1998] D. Galles and J. Pearl. An axiomatic characterization of causal counterfactuals. *Foundation of Science*, 3(1):151–182, 1998.

[Goldberger, 1972] A.S. Goldberger. Structural equation models in the social sciences. *Econometrica: Journal of the Econometric Society*, 40:979–1001, 1972.

[Hall, 1998] N. Hall. Two concepts of causation, 1998. In press.

[Halpern, 1998] J.Y. Halpern. Axiomatizing causal reasoning. In G.F. Cooper and S. Moral, editors, *Uncertainty in Artificial Intelligence*, pages 202–210. Morgan Kaufmann, San Francisco, CA, 1998.

[Heckman, 2001] J.J. Heckman. Econometrics and empirical economics. *Journal of Econometrics*, 100(1):1–5, 2001.

[Koopmans, 1953] T.C. Koopmans. Identification problems in econometric model construction. In W.C. Hood and T.C. Koopmans, editors, *Studies in Econometric Method*, pages 27–48. Wiley, New York, 1953.

[Lewis, 1986] D. Lewis. *Philosophical Papers*. Oxford University Press, New York, 1986.

[Lindley and Novick, 1981] D.V. Lindley and M.R. Novick. The role of exchangeability in inference. *The Annals of Statistics*, 9(1):45–58, 1981.

[Marschak, 1950] J. Marschak. Statistical inference in economics. In T. Koopmans, editor, *Statistical Inference in Dynamic Economic Models*, pages 1–50. Wiley, New York, 1950. Cowles Commission for Research in Economics, Monograph 10.

[Neyman, 1923] J. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–480, 1990. [Translation]

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

[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, 1995a] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–710, December 1995.

[Pearl, 1995b] J. Pearl. Causal inference from indirect experiments. *Artificial Intelligence in Medicine*, 7(6):561–582, 1995.

[Pearl, 2000a] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.

[Pearl, 2000b] J. Pearl. Comment on A.P. Dawid's, Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):428–431, June 2000.

[Robins, 1987] J.M. Robins. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Diseases*, 40(Suppl 2):139S–161S, 1987.

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

[Savage, 1962] L. J. Savage. *The Foundations of Statistical Inference*. Methuen and Co. Ltd., London, 1962.

[Simon and Rescher, 1966] H.A. Simon and N. Rescher. Cause and counterfactual. *Philosophy and Science*, 33:323–340, 1966.

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

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

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

[Strotz and Wold, 1960] R.H. Strotz and H.O.A. Wold. Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica*, 28:417–427, 1960.

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

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