REVIEW ESSAY: Probability in Artificial Intelligence

J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, San Mateo, California, 1988, xix + 552 pp. (Revised Second Printing 1991)

F. Bacchus, *Representing and Reasoning with Probabilistic Knowledge: A Logical Approach to Probabilities*, MIT Press, Cambridge, Massachusetts, 1990, xi + 233 pp.

P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, Springer Lecture Notes in Statistics No. 81, Springer, New York, 1993, xx + 522 pp.

Artificial intelligence has its roots in symbolic logic, and for many years it showed little interest in probability. But during the past decade, disinterest has been replaced by engagement. The flowering of expert systems during the 1980s strengthened ties between AI and areas of engineering and business that had long used probability and led to hybrid rule-based and probabilistic expert systems for a plethora of engineering and business problems, including speech recognition, vision, site selection, and process control. At the same time, probabilistic and statistical thinking has penetrated many areas of AI theory, including learning (Vapnik 1983, Valiant 1991), planning (Dean and Wellman 1991), and the evaluation of

artificial agents (Cohen 1990), to the point that AI has emerged as a contributor to the theory of probability and statistics.

What can the new role for probability in AI teach us about the philosophy of probability? Do the old interpretations of probability do justice to the new applications? Do the new applications justify old claims for the breadth of applicability of probability, or do they clarify limitations on probability?

The books under review provide an excellent starting point for assessing the philosophical implications of the growth of probability in AI. Judea Pearl is the single most influential advocate of probability in AI, and *Probabilistic Reasoning in Intelligent Systems* has been a major vehicle for this influence. Fahiem Bacchus is newer to the field, but his *Representing and Reasoning with Probabilistic Knowledge* represents the first real effort to deal with the distinction between subjective and objective probabilities within the logic-based tradition of AI. The third book, *Causation, Prediction, and Search*, represents one of the first fruits of the AI developments for the philosophy of statistics: three Carnegie-Mellon philosophers, Peter Spirtes, Clark Glymour, and Richard Scheines, inspired by can-do AI attitudes, challenge the conventional skepticism of statisticians about proving causation from correlation.

Because of its scope and influence, Pearl's book deserves special attention. Moreover, it is difficult to appreciate Pearl's philosophy of probability without understanding the context and practical importance of his work. So I will begin by describing the technical accomplishments of Pearl's book and by assessing his vision in the light of recent developments.

Then I will turn to the other two books. I will conclude by returning to my general questions about the scope and meaning of probability.

Pearl's Contributions. *Probabilistic Reasoning in Intelligent Systems* is a vigorous, sometimes confusing, always engaging mixture of philosophical, practical, and mathematical ideas. As Pearl says in his preface, there are "pointers to human-style reasoning in the midst of technical discussions, and references to computational issues in the midst of philosophical discussions." He explains that he adopted this style because he wanted to convey his own sense of excitement to his readers. Having talked with many of those readers, I can testify to his success.[1]

Pearl turns to probability to pursue a number of AI's goals. He wants to provide a model of human reasoning and use it in a way that is transparent to humans. In particular, he wants to generate verbal explanations even when the reasoning is numerical. He wants the computational efficiency needed for complex reasoning tasks. And he wants his methods to look familiar to those trained in AI—he wants to relate what he is doing to established ideas about knowledge representation, distributed processing, object-oriented programming, and constraint propagation. He is able to achieve these disparate goals by emphasizing graphical representations of probability distributions.

Pearl uses both undirected and directed graphs to represent conditional independence structures for multivariate probability distributions. The nodes of these graphs represent variables and the links (in the undirected case) or arrows (in the directed case) represent dependencies. More precisely, missing links or arrows represent

conditional independence relations. In the undirected graphs, separation means conditional independence; if two sets of variables are separated by a third, then they are conditionally independent given the third. In the directed graphs, the representation is more complex. The directed graph is assumed to have no directed cycles, so that the variables can be numbered (say $X_1,...,X_n$) in such a way that all arrows are from lower to higher numbers ($i<j$ whenever $X_i \rightarrow X_j$); we then assume that each X_i is independent of $\{X_1,...X_{i-1}\}$ given the subset of $\{X_1,...X_{i-1}\}$ with arrows to X_i. This is equivalent, as it turns out, to requiring that each variable be independent, given its 'parents', of its 'non-descendants.' Pearl is most interested in directed graphs, because he believes they can be used to represent the most powerful kind of human reasoning—causal reasoning.

The idea of using separation in undirected graphs to mean conditional independence was well established in probability before Pearl; it is the basis of the theory of Markov fields. Pearl's use of directed graphs is more original. In retrospect, we can see such graphs embedded in Sewall Wright's method of path analysis, which dates back to the 1930s, and in the tradition of 'linear structural models' that it helped establish. But Wright and his successors did not thoroughly analyze their directed graphs in terms of conditional independence, and Pearl's analysis has greatly clarified one interpretation of these graphs and widened their scope of application.

Pearl brings a novel viewpoint to mathematical probability, raising and sometimes answering whole new classes of questions. The standard framework for mathematical probability, inspired by statistical problems, begins with a fixed probability space and defines variables in terms of that

space. Pearl explores frameworks more natural for computer scientists. For example, he considers variables names (what are called 'attributes' in some branches of computer science) and asks about the possibilities for constructing a probability space for variables with these names. He asks, for example, what constellations of conditional independence relations among the variables are possible. This is a constructive approach: we begin with a sketch and then fill in the details. And it indicates a direction in which mathematical probability needs to grow in order to be more useful in computer science.

A substantial portion of *Probabilistic Reasoning in Intelligent Systems* is devoted to a message-passing scheme for carrying out various probabilistic computations within a directed probability graph. This scheme is perhaps the most original and influential idea in the book. In the simplest form of the scheme, the directed probability graph is thought of as a computer architecture, and each variable is thought of as a processor. Each variable communicates its own prior probabilities to its daughters (X_j is a daughter of X_i when $X_i \varnothing X_j$), and gets likelihoods back. These local and relatively simple messages enable the variables to compute their own posterior probabilities, and similar messages are used to find the likeliest joint configurations of the variables. This scheme achieves a number of Pearl's goals. Since the prior probabilities and likelihoods have verbal interpretations, a qualitative trace of the computation can often, in simple cases, be translated into a verbal explanation. At the same time, the computations for complex cases (involving many variables) are made manageable (reduced to computations involving only a few variables at a

time). And the scheme incorporates or extends ideas that had been developed within AI without reference to probability. It is similar to message-passing schemes for inheritance hierarchies and other semantic nets, and it can easily be programmed in an object-oriented language.

Though Pearl emphasizes the subjective interpretation of probability, he is also interested in the case where a person bases his or her beliefs on experience or statistical data, and hence he is interested in understanding the extent to which and the computational ease with which the structures represented by his directed graphs can be discovered or verified in data. Suppose, for example, that we use data to make a list of the conditional independence statements satisfied by a collection of variables. If these statements are consistent with some directed graph, is there a computationally efficient (say polynomial in the number of variables and the number of independence statements) algorithm for identifying this graph? Pearl made limited progress on this and similar questions in *Probabilistic Reasoning in Intelligent Systems*, but as I shall emphasize later in this review, much more has been accomplished in recent years, both by Pearl in collaboration with his student T.S. Verma, and by others, especially Spirtes, Glymour, and Scheines.

Another important contribution of *Probabilistic Reasoning in Intelligent Systems* is its discussion of non-monotonic and default logic. These terms refer to formal reasoning systems, developed in AI in the 1980s, which allow conclusions to be retracted in the light of later information. The proponents of these systems have sought to avoid probabilistic interpretations, but Pearl shows that such interpretations can be

helpful in two circumstances. The first is where reasoning is based on the idea of 'almost all'; in this case a formalization in terms of relative frequency or probability close to one leads to simple inference rules that were first formulated by Adams (1975). The second is where a causal model is involved; Pearl shows that in this case defaults can be handled more sensibly if they are labelled as causal (when they reason from cause to effect) or evidential (when they reason from from effect to cause), just as they often are in a directed probability graph.

I should mention yet one more contribution in *Probabilistic Reasoning in Intelligent Systems*: Pearl's interpretation and critique of the Dempster-Shafer theory, an alternative to Bayesian probability that I have promoted over the course of two decades. I will leave the issues raised in this critique aside in this review; they have been debated in detail by Pearl, myself, and others in the September 1990 (Vol. 4, no. 5/6) and May 1992 (Vol. 6, no. 3) issues of the *International Journal of Approximate Reasoning*.

Assessment. In the seven years since *Probabilistic Reasoning in Intelligent Systems* appeared, Pearl and others have developed many of its ideas further, and their connections to established or developing ideas in other areas have become clearer. With the hindsight this makes possible, I would like to give my own current assessment of the book—my own outline, as it were, for a revised edition.[2]

In many ways, my revision would make the book less exciting. In place of Pearl's attempt to show how one unified approach to probabilistic reasoning can meet all our goals, I would ask for a sorting out. Directed

graphs seem to be more successful for some of Pearl's purposes than for others, and more successful in some problems than in others. Where do these relative advantages lie, and what are the alternatives when directed graphs are not so successful? Similarly, Pearl's conditional independence mathematics seems to be more relevant to some applications than others. What exactly is its role? Pearl's technical contributions would survive this sorting out, but they would emerge looking more like a sharpening of some of the items in our probabilistic and statistical toolboxes than like a general way of looking at reasoning.

Pearl's conditional-independence mathematics actually play little role in the current practice of probabilistic expert systems. Typically, the construction of such systems does not start with a random assortment of conditional independence relations, whose implications and representation we then investigate using Pearl's algorithms. Instead, as Pearl himself points out, we look at the variables in some order $X_1,...,X_n$, chosen in accordance with our a *priori* judgments of possible causal influence, and we subjectively assess probabilities for each X_i conditional on a subset of $\{X_1,...,X_{i-1}\}$. By multiplying these conditional probabilities, we construct a probability distribution for the whole set of variables. This multiplication expresses implicit judgments of conditional independence, for in the constructed distribution, X_i is conditionally independent of $\{X_1,...,X_{i-1}\}$ given the subset, but we need not emphasize this fact or investigate what other conditional independence relations hold. Instead, we can get on with the practical tasks: using observations to monitor and update probabilities

in the model and computing probabilities and expectations for particular individuals.

The problem of monitoring and correcting subjectively assessed probabilities in a directed graphical model is a complicated one, and has been receiving increasing attention from statisticians (for a good review, see Spiegelhalter et al. 1993). This has produced some new ideas for statistics; most importantly, it has brought to the fore new ideas about 'prequential' monitoring that significantly generalize the traditional framework for statistical testing (see Dawid 1984 and Vovk 1993). But it has also brought Pearl's ideas into much closer contact with the statistical literature. After we have grappled at length with problems of monitoring and estimation in directed graphical models, we are apt to think of these graphs as one more kind of statistical model.

Another important area of progress has been in computing prior and posterior probabilities for directed graphical models. Pearl's message-passing scheme did this for simple cases in a way that allows us to interpret many of the auxiliary quantities used in the course of the computation as probabilities or likelihoods, and this permitted him to translate the computation into a verbal explanation. But this works well only if the graph forms a tree. It turns out that in more densely connected graphs, the desire for verbal interpretation conflicts with the goal of computational efficiency, which seems to be better served by methods most easily understood in terms of undirected graphs (see, e.g., Jensen et al. 1990). These methods still use message-passing, but the messages can no longer be interpreted in terms of prior probabilities and likelihoods. This decoupling of verbal explanation

from computation deprives us of some of the excitement we found in Pearl's text, but it is consistent with progress in experimental pschology. Human beings apparently do not always use the same system for verbal reasoning and recognition that they use for more computationally intensive tasks—the system we use to recognize objects, for example, seems to be distinct from the system we use to navigate their three-dimensional geometry.

The passage of time has also made more salient a number of other connections between Pearl's modelling and computational ideas and similar ideas in other domains. Most striking in this regard is the mushrooming use of Gibbs sampling in Bayesian statistics. An old idea from statistical mechanics, Gibbs sampling has long been used in physics and operations research, but it was brought to the attention of statisticians primarily by Pearl's advocacy of its use in directed graphical models and by Stuart and Don Geman's work (Geman and Geman 1984) on undirected graphical models for image recognition. As it turns out, Gibbs sampling has been of limited use in expert sytems, because its requirement that probabilities always be positive rules out the categorical or logical relationships that are often expressed in these models. But the method is enjoying great success in image recognition, where each pixel of the picture is a variable, and in Bayesian statistics, where posteriors may be difficult to compute by integration even when there are relatively few variables (Gelfand and Smith 1990, Gelman and Rubin 1993, Geyer 1993).

Time has also revealed the close similarity between Pearl's message-passing methods and the computational methods used for hidden Markov

models in speech recognition (Rabiner 1989) and other engineering problems and for the Kalman filter in control problems (Dempster 1990).

In view of all these connections, an up-to-date treatment of Pearl's topic would need to take a broader view its relation to other widely used probabilistic methods in engineering and science. This, in turn, would broaden the philosophical basis of the discussion, for while Pearl emphasizes the subjective interpretation of his models, related models are widely used in non-Bayesian as well as Bayesian contexts. Hidden Markov models for speech recognition, for example, are usually treated as objective models and tested and compared in the traditional sense of non-Bayesian statistics.

This review of recent progress has been limited to probability, but there is also vigorous continuing work on non-probabilistic and quasi-probabilistic methods for handling uncertainty in expert systems. In particular, it has been shown that many of the computational methods developed for probabilitistic systems, especially the message-passing schemes, can also be used by non-probabilistic systems. Shenoy and Shafer (1990) have analyzed axiomatically the abstract structure that is common to systems that can use such schemes.

Bacchus on Statistical and Propositional Probability. Symbolic logic has probably played as important a role in philosophy as it has in AI, but AI researchers often approach logic with an optimism beyond the wildest dreams of philosophical logicians; they imagine that logical deduction can serve as a tool for commonsense reasoning, and they treat the

completeness results of first order logic as reassurance about this possibility.

Fahiem Bacchus, in *Representing and Reasoning with Probabilistic Knowledge*, has constructed a synthesis of logic and probability in this spirit. And he has done so in a way that makes room for both frequencies, which he calls statistical probabilities, and degrees of belief, which he calls propositional probabilities.

Bacchus treats statistical probabilities by adding to first-order logic the facility to form terms that refer to the frequency with which a formula is satisfied. These terms can then be combined in other formulas, using arithmetic relations and logical connectives. Conditional frequencies (the frequency with which one formula is satisfied when another is satisfied) are also allowed. This logic has both a semantics—a model is a domain together with a measure—and an axiomatization, which is complete with respect to denumerable domains.

Inspired by earlier work by Halpern (1989), Bacchus adds to his statistical logic 'propositional' probability and expectation operators, which can again be used to form new terms that enter into new formulas. The resulting logic can be used to discuss probabilities for arbitrary propositions, including propositions that involve statistical assertions. Bacchus puts a rather Carnapian interpretation on his 'propositional' probabilities. They are hypothetical rather than actual degrees of belief. An actual subjective probability measure enters the story only as part of a model for interpreting the logic; subjectivity is thus relegated to the semantics of the logic. We are supposed to use the logic to reason about our

probabilities before committing ourselves to values for them, just as we might use unadorned first-order logic to reason about sentences before committing ourselves to an interpretation of the predicates in the sentences.

Bacchus provides an axiomatization for his combined logic, but his real interest lies not in reasoning within the logic but in using it in the spirit in which AI writers on default and non-monotonic reasoning have previously used ordinary first-order logic. That is to say, he is interested in principles for jumping to conclusions. The principles that he advances are, of course, the familiar ones: the principle of direct inference (which philosopher's are accustomed to calling Miller's principle) and various principles for narrowing reference classes.

The originality and importance of Bacchus's work lies, I think, in the simple fact that he has insisted on taking statistical knowledge seriously within the framework of symbolic logic. This provides a clear challenge to the many logicians, in both philosphy and AI, who have sought to avoid statistical knowledge or even to substitute logic for it. Bacchus's logic is a logic in the traditional sense, and it is comprehensive as a treatment of probabilty; it encompasses both statistical knowledge and belief in a straightforward way. So it is natural to see formal logics that deal with frequency or belief in a more restricted way as subsets of Bacchus's logic. We can apply this attitude, for example, to the logics of probability quantifiers, developed in the philosophical literature by Keisler (1977), Hoover (1978) and Vickers (1988). These authors aim to capture a logical conception of probability distinct from any frequency conception, but it is hard to see how the intutitions that justify their logics can be restrained from

leading us to Bacchus's more comprehensive logic, at which point we clearly have a statistical semantics. Similarly, as Bacchus himself emphasizes, the Hintikka-style logic for fallible and recursive belief (KD45 or weak S5 with consistency) is a subset of Bacchus's logic, and its belief operator can be derived from Bacchus's propositional probability operator. Thus Bacchus's logic provides a framework within which to evaluate debates between those who treat default reasoning probabilistically and those who would treat it in terms of self-referential belief.

Does *Representing and Reasoning with Probabilistic Knowledge* advance the project of implementing probability in artificial intelligence? This is doubtful; being an extension of first-order logic, Bacchus's logic is no more implementable than first-order logic is. More than sensitive to this point, Bacchus argues at length that formal logics are essential tools for analyzing knowledge, even if they do not give us practical ways to manipulate it. Like many other logicians working in AI, Bacchus claims that he is showing us what we should do in principle. We cannot do it in practice, but it provides a standard of coherence to our practice.

This last argument does not make sense to me; I do not see how the internal coherence of any logic can suffice to make it a standard for practice or a useful tool for analyzing our knowledge. A logic can be relevant to our knowledge only if our knowledge takes the particular form the logic treats. So Bacchus's logic does nothing to get us past the traditional objections to the generality of probability. Before it makes sense to use his logic we must have both a well-defined reference class that makes it meaningful to talk about frequencies and well-defined betting rates that make it meaningful to

talk about subjective probabilities. There must be reference in application as well as in semantics. In my judgment, real progress in integrating probability and logic will require the use of a constructive approach to logic, such as Per Martin Löf's type theory, which integrates syntax and semantics (Ranta 1994).

Causation from Correlation? Spirtes, Glymour, and Scheines are philosophers, but they are Carnegie-Mellon philosophers. In their treatment of probability and causation they have largely abandoned the philosophical tradition associated with Reichenbach and Salmon in favor of ideas coming from statistics and artificial intelligence. Their earlier book (Glymour, Scheines, Spirtes, and Kelly 1987) showed how to use partial correlations from data to search for structural or causal models. The book was strongly influenced by the AI tradition associated with Herbert Simon, with its emphasis on search, by the early psychometricians, especially Charles Spearman, and by the current use of linear structural modelling in psychology and sociology. It was a very literate book, and it attracted considerable attention, but many would-be readers, myself included, were unable to make sense of its mathematical foundations. In retrospect, what was missing was Pearl's conditional-independence ideas, which provide a clear mathematical definition, at least, of what we are searching for. Their new book, *Causation, Prediction, and Search*, is founded on Pearl's conditional-independence mathematics and thus is able to take a large step towards a deeper and more coherent understanding of causal modelling.

Causation, Prediction, and Search is aptly placed in Springer's Lecture Notes in Statistics series. It is as unpolished as one would expect

lecture notes to be, and though its authors are philosophers, it is weak in its philosophical explication of causation. Its greatest contributions are its algorithms for finding directed graphical models that fit the conditional-independence structure observed in data and its application of these algorithms to real examples.

The algorithms that form the backbone of *Causation, Prediction, and Search* are the product of interaction between its authors and Verma and Pearl, whose work is reported in Pearl and Verma (1991) and Verma and Pearl (1992). I have not sorted out the relative contributions of the two research groups, but I can report the basic ideas of the algorithms.

We start with the assumption that the variables with which we are working can be put into a directed graph (initially unknown to us) in which arrows have a subtle causal interpretation. The subtlety lies in the fact that we must interpret groups of arrows rather than individual arrows. The arrows pointing to a given variable X indicate that the variables from which they point—X's parents—interact to influence X, while earlier ancestors have only an indirect influence on X. This leads to the conditional independence relations we discussed earlier: since the influence of ancestors is through the parents; X is conditionally independent of ancestors (and other non-descendants as well) given the parents. We further assume that these causal conditional independence relations and the further conditional independence relations that they imply (which can be listed using Pearl's graphical criterion of 'd-separation') are the only conditional independence relations obeyed by the joint distribution of the variables;

there are no other causal relations, and conditional independence does not happen by accident, with no causal explanation.

Given these assumptions, can we identify the directed graph from data—i.e., from observations of all the variables for a number of individuals? In principle, we sometimes can, if we observe enough individuals. With enough observations, we could identify all the conditional independence relations that hold among the variables, and we could then identify the graph or graphs that imply exactly these conditional independence relations—all of them and no others. It seems doubtful, however, that this project could really be carried out. There are so many conditional independence relations involved and so many potential graphs that merely counting them all may be computationally impossible, and the statistical problem of interpreting so many simultaneous dependent tests of independence appears insoluble. Much to the surprise of statisticians such as myself, however, Pearl, Verma, and the Carnegie-Mellon philosophers have shown that the problem often is soluble at a practical level. It is true that it is impractical to give simultaneous significance levels for the tests of independence, and it is also true that we never have enough data to test conditional independencies that involve conditioning on many variables, but if the graph we are trying to identify is relatively sparsely connected, we take a relaxed attitude towards the significance level, and we are shrewd in our choice of which conditional indpendencies to test first, we can often idenify the graph. And our ability to identify it improves dramatically if we have a priori knowledge that constrains the possible causal relations.

The algorithms reported in *Causation, Prediction, and Search* can give several kinds of results. We may find a unique directed graph that fits the data. We may find several directed graphs that fit the data—several directed graphs that imply exactly the same conditional independence relations that we find in the data. We may be told that there is no directed graph that implies exactly the conditional independence relations found in the data, but that the addition of more variables to the graph—unobserved or latent variables—could remedy the situation. Or we may be told that it is impossible to reproduce observed conditional independencies with a directed graph even if we allow latent variables.

It should be emphasized that the progress reported in *Causation, Prediction, and Search* has depended on sidestepping a number of problems, especially the problem of providing a probabilistic interpretation for the simultaneous tests of significance. Statisiticians have been so impressed by these problems, and so chastened by experience with the nonsense that can result from cavalier use of simultaneous tests, that they have counseled against ever expecting to prove causation from correlation. Computer science has produced new attitudes, which may be more approrpriate when the observations number in the thousands rather than the tens or hundreds. For computer scientists, problems related to the fallibility of individual tests seem secondary; the pressing question is whether we have time to perform enough tests to draw a conclusion. In addition to the progess in causal inference that we are considering here, the new computer-science attitudes have also led to several other new branches of statistical theory, including Vapnik's work on identifying functions, Valiant's work on

learning, and the whole research community that has recently grown up under the rubric of "computational learning theory."

The mathematics of *Causation, Prediction, and Search* is unpolished, especially when it delves into the possibilities for latent variables. And it is a rather unfamiliar mathematics for the applied statisticians who will be most interested in the results. Consequently, it make take considerable time for the new methods to be integrated into statistical practice. In the long run, however, I think that this approach to causal search will contribute to greater understanding and discipline in the use of factor analysis.

From the philosophical viewpoint, there remains a huge gap in the argument. This is the causal intepretation of the unknown directed graph for which we are searching. What do we mean when we say that X is directly caused by its parents and only indirectly caused by earlier variables, and how do we get from an answer to this question to the idea of conditional independence? Chapter 3 of *Causation, Prediction, and Search* is devoted to this question, but it is the weakest chapter of the book, and in the end it takes refuge in the contention that we should go aheand and talk about causation even if we do not understand it. After all, we have done a lot with probability, and no one understands it either.

In recent work (Shafer 1995), I have made some suggestions of my own for closing the gap between causation and conditional independence. These suggestions begin with the contention that it is more natural to talk about causation in the context of an event tree—a tree that lays out the possible ways that a sequence of experiments might come out. (Perhaps we spin a fair coin, then roll a fair die or draw a card depending on how the flip

comes out, etc. Or perhaps there is an experiment determining the occupations of a person's parents, then another experiment, the probabilities for which depend on how the first experiment comes out, determining the person's education, etc.) It is natural and unmysterious to say that a certain event, or the value of a certain variable, can be caused by certain experiments in an event tree—or more precisely, by certain outcomes of these experiments. If we do not see the the experiments or the tree—if instead we observe only certain events or variables—then we cannot talk so directly about causation. But we can make less direct statements that lead to conditional independence for variables. One key concept is 'tracking.' We say that a set of variables A tracks a variable X if at all the points in the tree where A is resolved in a given way (i.e., points where the variables in A come to have certain definite values), the probabilities for X are always the same. It turns out that if A tracks X and if after A's determination the probabilities for X do not change at the same time as the probabilities for the variables in another set B, then X and B are conditionally independent given A. Thus the conditional independence relations found in directed graphs can be interpreted in terms of qualitative conditions (when probabilities change) on an unseen event tree. These qualitiative conditions, in turn, can be interpreted as statements about causation; they say something indirectly about the experiments which influence the happening of events and the values of variables.

The interpretation in terms of event trees can be used to make precise Reichenbach's principle of the common cause (when events or variables are correlated, there is always at least one experiment that affects both), and it

also permits an interpretation of Reichenbach's contention that the direction of time can sometimes be deduced from statistics. If we suppose that the conditional independencies within a family of variables result from the process of tracking that I have just described—then the directed graphs for the family produced by the Pearl-Verma-Spirtes-Glymour-Scheines algorithms imply constraints on the order in which the values of the variables are determined.

The Interpretation of Probability in AI Systems. In *Probabilistic Reasoning in Intelligent Systems*, Pearl argues for a synthesis of subjective and objective probability—what he calls in his preface a 'computation-minded' interpretation of probability. For Pearl, probability is initially subjective, but it follows the rules for frequencies for two reasons: because we want to match our beliefs with our experience, and because the rules for frequency are unique in the flexible way in which they handle dependency information.

Pearl does not give much more argument or explanation than this for his computation-minded interpretation, but I believe that the development of the theory of probabilistic expert systems has given it a deeper justification. As I mentioned above, probabilistic expert systems turn out to be a natural setting for the ideas of Dawid and Vovk on the evaluation of probabilistic predictions. The probabilities in a system may initially be subjective, but as we use the system, we have more and more opportunity to evaluate and improve its probabilistic predictions. These predictions usually do not involve the identical and independent trials that inspired the frequency interpretation, but they can nevertheless be scored, using a proper scoring

rule. Scores for successive predictions, even successive predictions involving different questions, can be added, and as Dawid and Vovk show, the cumulative score should, if the probability predictions are valid, follow a law of large numbers and even a central limit theorem. This implies a limited frequency interpretation of the probabilities: if we make successive predictions, each hedged with a probability, the average probability will approximate the proportion of correct predictions. But this limited frequency interpretation is embedded in a framework that is subjective inasmuch as it describes the epistemic situation of a particular system; the frequency is relative to the events on which the system ventures to bet and it expresses limits on the system's ability to predict these events; it is not relative to a reference class defined independently of the system.

When we turn to the task of inferring conditional independencies from data, the subjective aspect of probability recedes. Now we are in the world of statistics, trying to infer probabilistic structure from data, and we are not even doing so in a Bayesian way. Spirtes, Glymour, and Scheines accordingly adopt a propensity interpretation for the probabilities they are investigating. Interestingly, however, the event-tree story that I sketched above brings us back to the predictive framework. The event tree is not observed by us, but in order to make sense of the probabilities in this tree, we must consider an ideal observer whose knowledge unfolds as events move down the tree (Shafer 1993). Indeed, event trees provide a general framework for the Dawid-Vovk theory of probabilistic prediction.

The interaction of probability with AI obviously has not produced a new consensus about the meaning of probability. All the traditional

interpretations are now being espoused or used in various ways within AI. We see probabilistic expert systems touted as an application of purely subjective probability, we see the statistical methods emerging from AI used with a propensity interpretation, and we see logics intended to make room for both frequencies and beliefs. But for my own part, I see the ingredients for a new synthesis, a computation-minded and prediction-minded interpretation of probability that recognizes both the belief and frequency aspects of the same system of probabilities, and that sees in this dual role a necessary condition for the successful application of mathematical probability to reasoning tasks. Different applications may emphasize more the objective or subjective aspect of mathematical probability, but no application can do without both, and when there is no way to bring the two together, we are outside the domain where mathematical probability is useful.

NOTES

Research for this essay has been partially supported by the National Science Foundation through grant #SBE9213674. The essay has also benefited from the author's participation in a seminar on probabilistic causation in the Department of Philosophy at Princeton University. The author would like to acknowledge the contributions of the other participants: Dick DeVeaux, Adam Grove, Gil Harman, Paul Holland, Dick Jeffrey, and Bas van Fraassen.

[1] In the spirit of full disclosure, I should mention that I have enjoyed a direct professional relationship with Pearl. The most notable fruit of this

relationship is a jointly edited book of readings, *Readings in Uncertain Reasoning*, published by Morgan Kaufmann in 1990.

[2] The second revised printing, published in 1991, is not a major revision; it retains the page numbering of the 1988 edition. It does, however, give many references to more recent work.

REFERENCES

Adams, E.: 1975, *The Logic of Conditionals*, D. Reidel, Dordrecht.

Cohen, P. R.: 1990, 'A Survey of the Eighth National Conference on Artificial Intelligence: Pulling Together or Pulling Apart?' *AI Magazine*, Vol. 11, No. 4, pp. 17-41.

Dawid, A.P.: 1984, 'Statistical Theory—The Prequential Approach,' *Journal of the Royal Statistical Society, Series A*, **147** 277-305.

Dean, T., and M. Wellman: 1991, *Planning and Control*, Morgan Kaufmann, San Mateo, California.

Dempster, A.P.: 1990, 'Normal Belief Functions and the Kalman Filter,' Technical Report, Department of Statistics, Harvard University.

Gelfand, A.E., and A.F.M. Smith: 1990, 'Sampling-Based approaches to Calculating Marginal Densities,' *Journal of the American Statistical Association*, **85** 398-409.

Geman, S., and D. Geman, D.: 1984, 'Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6** 721-742.

Glymour, C., R. Scheines, P. Spirtes, and K. Kelly: 1987, *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modelling*, Academic Press, New York.

Jensen, F.V., K.G. Olesen, and S.K. Andersen: 1990, 'An Algebra of Bayesian Belief Universes for Knowledge-Based Systems,' *Networks*, **20** 637-59.

Pearl, J., and Verma, T.S.: 1991, 'A Theory of Inferred Causation,' in *Principles of Knowledge Representation and Reasoning; Proceedings of the Second International Conference*, Morgan Kaufmann, San Mateo, California, pp. 441-452.

Rabiner, L.R.: 1990, 'A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,' in A. Waible and K.-F. Lee (eds.), *Readings in Speech Recognition*, Morgan Kaufmann, San Mateo, California, pp. 267-296.

Ranta, Aarne: 1994, *Type-Theoretic Grammar*. Oxford, Oxford University Press.

Seillier-Moiseiwitsch, F., and A.P. Dawid: 1993, 'On Testing the Validity of Sequential Probability Forecasts, *Journal of the American Statistical Association*, **88** 355-359.

Shafer, Glenn: 1993, 'Can the Various Meanings of Probability Be Reconciled?' in Gideon Keren and Charles Lewis (eds.), *A Handbook for Data Analysis in the Behavorial Sciences: Methodological Issues*, Lawrence Erlbaum, Hillsdale, New Jersey, pp. 165-196.

Shafer, Glenn: 1995, *The Art of Causal Conjecture*. Cambridge, Massachusetts, MIT Press.

Shenoy, Prakash P., and Glenn Shafer: 1990, 'Axioms for Probability and Belief-Function Propagation,' in R.D. Shachter, T.S. Levitt, L.N. Kanal, and J.F. Lemmer (eds.),.*Uncertainty in Artificial Intelligence, Vol. 4*, North-Holland, Amsterdam, pp. 169-198.

Spiegelhalter, David J., A. Philip Dawid, Steffen L. Laurtizen, and Robert G. Cowell: 1993, 'Bayesian Analysis in Expert Systems (with discussion),' *Statistical Science*, **8** 219-283.

Valiant, L.: 1991, 'A View of Computational Learning Theory,' in C.W. Gear (ed.), *Computation and Cognition; Proceedings of the First NEC Research Symposium*, SIAM, Philadelphia, pp. 32-51.

Vapnik, V.: 1983, *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York.

Verma, T.S., and J. Pearl: 1992, 'An Algorithm for Deciding if a Set of Observed Independencies has a Causal Explanation,' in D. Dubois, M.P. Wellman, B. D'Ambrosio, and P. Smets (eds.), *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Mateo, California, pp. 323-330.

Vickers, John M.: 1988, *Chance and Structure: An Essay on the Logical Foundations of Probability*. Clarendon Press, Oxford.

Vovk, V.G.: 1993, 'A Logic of Probability, with application to the Foundations of Statistics (with discussion),' *Journal of the Royal Statistical Society, Series B*, **55** 317-351.

GLENN SHAFER

Department of Accounting and Information Systems

Graduate School of Management, Rutgers University
180 University Avenue, Newark, New Jersey 07102