Statistical Science
1987, Vol. 2, No. 1, 3-44

Probability Judgment in Artificial Intelligence and Expert Systems

Glenn Shafer

Abstract. Historically, the study of artificial intelligence has emphasized symbolic rather than numerical computation. In recent years, however, the practical needs of expert systems have led to an interest in the use of numbers to encode partial confidence. There has been some effort to square the use of these numbers with Bayesian probability ideas, but in most applications not all the inputs required by Bayesian probability analyses are available. This difficulty has led to widespread interest in belief functions, which use probability in a looser way. It must be recognized, however, that even belief functions require more structure than is provided by pure production systems. The need for such structure is inherent in the nature of probability argument and cannot be evaded. Probability argument requires design as well as numerical inputs. The real challenge probability poses to artificial intelligence is to build systems that can design probability arguments. The real challenge artificial intelligence poses to statistics is to explain how statisticians design probability arguments.

Key words and phrases: Artificial intelligence, associative memory, Bayesian networks, belief functions, certainty factors, conditional independence, constructive probability, diagnostic trees, expert systems, production systems.

I have been asked to speak on the use of belief functions in artificial intelligence and expert systems. For the sake of perspective, I propose to address the broader topic indicated by my title. The theory of belief functions is part of the theory of probability judgment, and a general understanding of the role of probability judgment in artificial intelligence can help us understand the particular role of belief functions.

I will not attempt to evaluate all the ways in which probability has been used in artificial intelligence, nor even all the ways in which belief functions have been used. Instead, I will aim for some general insights into the interaction between probability ideas and artificial intelligence ideas. Many of my comments will be historical. I hope readers will forgive me for those cases where I belabor the obvious or repeat the well known; my excuse is that I hope to reach a dual audience—students of probability who may not know very much about artificial intelligence, and students of artificial intelligence who may not know very much about probability.

The first two sections of the paper are introductory in nature. Section 1 considers the reasons for the artificial intelligence community's initial disinterest in probability and its recent change of heart and outlines the paper's conclusions about how current expert systems fall short of putting probability judgment into artificial intelligence. Section 2 deals with probability judgment without reference to artificial intelligence; here I discuss the split between Bayesian and non-Bayesian methods and place the theory of belief functions in this historical context.

Section 3 reviews some strands of the development within artificial intelligence of ideas about using probability judgment in expert systems. Here we see how the general issues that separate the Bayesian and belief-function theories appear in the context of expert systems, and we gain some insight into why flexibility is harder to achieve with probability judgment than with other kinds of reasoning. Section 4 discusses the problem of giving an artificial intelligence a genuine capacity for probability judgment.

1. THE EMERGENCE OF PROBABILITY IN ARTIFICIAL INTELLIGENCE

Until recently, the artificial intelligence community showed relatively little interest in probability. There is little probability, for example, in the three-volume *Handbook of Artificial Intelligence* (Barr and

Glenn Shafer is a Professor in the School of Business, University of Kansas, Lawrence, Kansas 66045.

Feigenbaum, 1981, 1982; Cohen and Feigenbaum, 1982). During the past 4 or 5 years, however, probability and the management of uncertainty in intelligent systems has become a widely discussed topic. Why the initial disinterest, and why the change?

The reasons for the initial disinterest are clear. Probabilities are numbers, and number crunching is just what artificial intelligence was supposed not to be. When the artificial intelligence community was founded, computers were used mainly for number crunching. They were impressively good at this, but they were not intelligent. Intelligence seems to require more general kinds of symbol manipulation.

Moreover, when we begin to think about computer programs that will match the achievements of human intelligence, we find that we are thinking about programs with non-numerical inputs and outputs. What place is there for talk about numbers in the case of these programs? They are merely sets of rules for going from the inputs to the outputs, and while it might be possible to identify some intermediate steps that are analogous to operations on numerical probabilities, it seems pointless to do so. It seems better to tell what is really going on.

The prejudice against numbers in general and probabilities in particular has not entirely disappeared from artificial intelligence, and the argument sketched in the preceding paragraph is still made. This argument is part of the motivation for the continuing development within artificial intelligence of non-numerical methods for handling uncertainty. These include nonmonotonic logic (McCarthy, 1980; McDermott and Doyle, 1980; Reiter, 1980) and Paul Cohen's theory of endorsements (Cohen, 1985).

But the factors that caused this prejudice have substantially changed. The vague idea that artificial intelligence can be defined largely through the contrast with number crunching has been replaced by the equally vague but equally powerful idea that intelligence is produced by complexity and by access to large amounts of knowledge. Two specific openings have appeared for probability.

1. The ban on non-numerical inputs has been dropped in some cases. In addition to programs that try to match aspects of human intelligence, artificial intelligence is now also concerned with expert systems and other intelligent systems that interact with human users and can use numerical inputs supplied by these users.

2. The artificial intelligence community has absorbed David Marr's views on levels of explanation. In his work on vision, Marr convincingly made the point that full understanding of an intelligent system involves explanation at various levels. In addition to explanation at the level of implementation (what is really going on) we also need explanation at more abstract levels. "It's no use, for example, trying to understand the fast Fourier transform in terms of resistors as it runs on an IBM 370" (Marr, 1982, page 337). Understanding of this point takes the rhetorical force out of the argument that there is no place for probability ideas when inputs and outputs are non-numerical.

Most of the current interest in probability in artificial intelligence is the result of (1). In many cases it is impossible to build expert systems without the use of probability. But in the long run, (2) may be more important. Because of (2), we can now recognize the value to an artificial intelligence of an ability to design probability arguments and generate the numerical judgments they require.

The ban on numerical inputs in artificial intelligence was dropped because the artificial intelligence community became interested in expert systems. Why did this happen? The answer is that the community discovered ways of building expert systems that incorporated ideas that seemed to reflect important aspects of human intelligence. As I explain in Section 3, most of the expert systems developed within artificial intelligence have been production systems, relatively unstructured programs that have some of the flexibility in acquiring and using knowledge that is characteristic of intelligence.

I argue in this paper that the expert systems we can now build to use probability judgments do not have this kind of flexibility and hence fit awkwardly under the heading of artificial intelligence. The problem is that probability judgment requires an overall design and hence cannot be achieved by relatively unstructured methods of programming applied to individual numerical probabilities. I will argue in Section 4 that both the overall design of probability judgment and the determination of individual numerical probabilities can be achieved by an artificial intelligence only if it is equipped with a genuine associative memory.

As a result of the explosion of interest in expert systems, the field of artificial intelligence is now struggling to maintain its sense of identity. The idea of an expert system began in artificial intelligence, but any system with expert capabilities can justifiably claim the name, whether it is written in LISP or FORTRAN, and many systems developed outside of artificial intelligence have more impressive expert capabilities than those developed inside it. It is clear, therefore, that artificial intelligence must withdraw from its embrace of the whole field of expert systems in order to maintain intellectual coherence. But it is unclear just what parts of the field of expert systems will remain in the embrace. My suggestion here is that artificial intelligence will retain its newfound interest in probability but will look beyond the current expert systems to deeper uses of probability ideas.

2. BAYESIAN AND BELIEF-FUNCTION ARGUMENTS

In this section I review some general ideas about probability judgment, without reference to the particular problems of artificial intelligence. I begin by sketching a way of looking at the frequentist vs. Bayesian controversy, a controversy that has dominated discussions of probability judgment for more than a century. After developing a constructive understanding of the Bayesian theory, I introduce another constructive theory, the theory of belief functions. I argue that both theories should be thought of as languages for expressing probability judgments and constructing probability arguments.

2.1 Two Strategies for Probability Judgment

What we now call the mathematical theory of probability was originally called the theory of games of chance. Probability was an entirely different topic; something was probable when there was a good argument or good authority for it. When James Bernoulli and others began to use the word probability in connection with the theory of games of chance, they were expressing the ambition that this theory might provide a general framework for evaluating evidence and weighing arguments. But just how might this work? How can the theory of games of chance help us evaluate evidence?

In the nineteenth century, it became clear that there are two distinct strategies for relating evidence to the picture of chance. Today, these two strategies might be called the frequentist and Bayesian strategies, but in order to avoid some of the connotations of these names, let me call them, for the moment, the *direct probability* and *conditional probability* strategies.

The direct probability strategy relies on direct application of the idea that in life, as in games of chance, what happens most often is most likely to happen in a particular case under consideration. The ideal kind of evidence for this strategy is knowledge of the frequency of outcomes in similar cases. I assign a 98% probability to the prediction that a student who first appears 3 weeks after the beginning of my elementary statistics course will not be able to pass the course, because it has almost always turned out that way in the past.

The conditional probability strategy uses the picture of chance in a deeper way. It observes that games of chance unfold step by step, with the probabilities for different possible final outcomes changing at each step, and it suggests that the accumulation of evidence should change probabilities in a similar step by step way. Thus, my probability for whether the late-appearing student will pass my course should change when I learn more about his history and circumstances, just as my probability for whether two successive rolls of a die will add up to nine will change when I learn the result of the first roll. The conditional probability strategy usually leads to a more complicated argument than the direct probability strategy, since it involves construction of a probability measure over a more complicated frame and then the reduction of this measure and frame by conditioning.

In general, there is not, I believe, any *a priori* reason to prefer one of these two strategies to the other. We cannot say that it is normative to use one and irrational to use the other. They are both strategies for producing arguments, and it is the cogency of the arguments that must be evaluated. It may be most cogent to lump my new late-appearing student with all my past late-appearing students, on the grounds that particulars have not made much difference in the past. Or I may have had enough experience with late-appearing students like this one on some particulars that I can make a better direct probability argument by looking at the past frequency of success just for these late-appearing students. Or I may have the experience and insight needed to construct a probability measure that I can condition on the particulars. The issue cannot be settled in the abstract, without reference to the experience I bring to bear on the problem.

Moreover, neither of the two strategies is inherently more objective or subjective than the other. It is true that the direct probability strategy, since it tends to consider broader classes, is more likely to result in probability judgments based on actual frequency counts. But the objectivity of these frequencies must always be coupled with a subjective judgment of their relevance. And even with broad classes we most often have hunches and impressions rather than actual counts.

Historically, however, the direct probability strategy has come to be associated with claims to objectivity, whereas the conditional probability approach has come to be associated with claims to rationality. This fact seems to be a result of efforts to square the interpretation of probability with the empiricist and positivist philosophical trends of the late nineteenth and early twentieth centuries.

2.2 The Frequentist vs. Bayesian Deadlock

Laplace, writing at the beginning of the nineteenth century, was able to define numerical probability as the measure of the "reason we have to believe." But by the middle of the nineteenth century, many students of probability were looking for a more empirical definition. They found this definition in the idea of frequency, and they proceeded to reject those applications of probability theory that could not be based

on observed frequencies. In particular, they rejected Laplace's method of calculating the probability of causes, which is a special case of the conditional probability strategy.

The frequentist philosophy severely restricted the domain of application of numerical probability, and those who wanted to use numerical probability more generally were forced to search for a philosophical foundation for the conditional probability strategy that would fit the positivist mind-set. Such a philosophical foundation was finally established in the twentieth century by Ramsey, de Finetti, and especially Savage. These authors conceived the idea that subjective probability should be given a behavioral and hence positivist interpretation—a person's probabilities should be derivable from his choices. They formulated postulates for what they called rational behavior, postulates that assure that a person's choices do determine numerical probabilities. And they argued that it is normative to follow these postulates and hence normative to have subjective probabilities.

During the past two decades, the philosophical foundation provided by Savage's postulates has led to a remarkable resurgence, both mathematical and practical, of the conditional probability strategy. The resulting body of theory has been called "Bayesian," because the conditional probability strategy often uses Bayes' theorem.

Although the new Bayesian philosophy has played a historically valuable role in rescuing the conditional probability strategy from its frequentist opponents, it has its own obvious shortcomings. Most important, perhaps, is its inability to explain how the quality of a probability analysis depends on the availability and quality of relevant evidence. Whereas the frequentist philosophy tries to limit applications of probability to models for which we have clearly relevant and objective frequency counts, there is nothing in the Bayesian philosophy to make our choice of a model depend in any way on the availability of relevant evidence. The postulates apply equally to any model.

We have, then, a deadlock between two inadequate philosophies of probability. On the one side, the frequentist philosophy, which recognizes the relevance of evidence but tries to justify claims to objectivity by limiting numerical probability judgment to cases where the evidence is of an ideal form; on the other side, the Bayesian philosophy, which recognizes the subjectivity of all probability judgment but ignores the quality of evidence and claims it is normative to force all probability judgment into one particular mold.

We have been caught in this deadlock for three decades. We have tired of it, and we are inclined to ask the two sides to compromise (see, e.g., Box, 1980). But we have not been able to find a philosophical foundation for probability judgment that can resolve the deadlock.

I believe that the way out of the deadlock is to back up and recognize that a positivist philosophical account of probability is no longer needed. Our intellectual culture has moved away from positivism and toward various sorts of pragmatism, and once we recognize this we will be free to discard both the frequentists' claims to objectivity and the Bayesians' claims to normativeness.

2.3 Constructive Probability

In several recent papers (especially Shafer, 1981; Shafer and Tversky, 1985) I have proposed the name "constructive probability" for the pragmatic, postpositivist foundation that I think we need for probability judgment. The idea is that numerical probability judgment involves fitting an actual problem to a scale of canonical examples. The canonical examples usually involve the picture of chance in some way, but different choices of canonical examples are possible, and these different choices provide different theories of subjective probability, or, if you will, different languages in which to express probability judgments. No matter what language is used, the judgments expressed are subjective; the subjectivity enters when we judge that the evidence in our actual problem matches in strength and significance the evidence in the canonical example.

Within a given language of probability judgment, there can be different strategies for fitting the actual problem to the scale of canonical examples. The direct and conditional probability strategies described above live, I think, in the same probability language, the language in which evidence about actual questions is fit to canonical examples where answers are determined by known chances. We may call this language the Bayesian language. (For a more detailed account of different strategies that are available within the Bayesian language, see Shafer and Tversky (1985). The distinction between the direct and conditional probability strategies corresponds to the distinction that is made there between total-evidence and conditioning designs.)

The constructive viewpoint tells us that when we work within the Bayesian language we must make a judgment about how far to take the conditional probability strategy in each particular problem. We make this judgment on the basis of the availability of evidence to support the conditional and unconditional probability judgments that are required.

It may be useful to elaborate on this point. Suppose we want to make probability judgments about a frame of discernment S. (A *frame of discernment* is a list of possible answers to a question; we want to make

probability judgments about which answer is correct.) We reflect on our evidence, and we produce a list $E_1, \ldots, E_n$ of facts that seem to summarize this evidence adequately. The conditional probability strategy amounts to standing back from our knowledge of these n facts, pretending that we did not yet know them, and constructing a probability measure over a frame that considers not only the question considered by S but also the question whether $E_1, \ldots, E_n$ are or are not true; typically we construct this measure by making probability judgments $P(s)$ and $P(E_1 \,\&\, \ldots \,\&\, E_n \mid s)$ for each s in S. The problem with this strategy is that we now need to look for further evidence on which to base all these probability judgments. We have used our best evidence up, as it were, but now we have an even larger judgmental task than before. According to the behaviorist Bayesian theory, there is no problem—it is normative to have the requisite probabilities, whether we can identify relevant evidence or not. But according to the constructive viewpoint, there is a problem, a problem that limits how far we want to go. We may want to apply the conditional probability strategy to some of the E_i, but we may want to reserve the others to help us make the probability judgments (see Shafer and Tversky, 1985).

2.4 The Language of Belief Functions

Whereas the Bayesian probability language uses canonical examples in which known chances are attached directly to the possible answers to the question asked, the language of belief functions uses canonical examples in which known chances may be attached only to the possible answers to a related question.

Suppose S and T denote the sets of possible answers to two distinct but related questions. When we say that these questions are related, we mean that a given answer to one of the questions may fail to be compatible with some of the possible answers to the other. Let us write "$s\mathrm{C}t$" when s is an element of S, t is an element of T, and s and t are compatible. Given a probability measure P over S (assume for simplicity that P is defined for all subsets of S), we may define a function Bel on subsets of T by setting

$$\text{(1)} \qquad \mathrm{Bel}(B) = P\{s \mid \text{if } s\mathrm{C}t, \text{ then } t \text{ is in } B\}$$

for each subset B of T. The right-hand side of (1) is the total probability that P gives to those answers to the question considered by S that require the answer to the question considered by T to be in B; the idea behind (1) is that this probability should be counted as reason to believe that the latter answer is in B. We might, of course, have more direct evidence about the question considered by T, but if we do not, or if we want to leave other evidence aside for the moment, then we may call $\mathrm{Bel}(B)$ a measure of the reason we have to believe B based just on P.

The function Bel given by (1) is the *belief function* obtained by extending P from S to T. A probability measure P is a special kind of belief function; this is just the case where (i) $S = T$ and (ii) $s\mathrm{C}t$ if and only if $s = t$. Thus the language of belief functions is a generalization of the Bayesian language.

All the usual devices of probability are available to the language of belief functions, but in general we use them in the background, at the level of S, before we move to degrees of belief on T, the frame of interest.

Like other non-Bayesian approaches to probability judgment, the language of belief functions countenances the use of probability models that are less complete than Bayesian models. In order to obtain a belief function over T, we begin with a probability measure over S alone, and we use observed facts to create a compatibility relation C between S and T. A Bayesian conditional probability argument that used the frames S and T would extend the probability measure over S to a complete probability measure over $S \times T$, and it would then use the compatibility relation to condition this measure.

I have studied the language of belief functions in detail in earlier work—see especially Shafer (1976, 1986a). Here I will use some examples of (1) to illustrate the language and to contrast it with the Bayesian language.

Example 1. Is Fred, who is about to speak to me, going to speak truthfully, or is he, as he sometimes does, going to speak carelessly, saying whatever comes into his mind? Let S denote the possible answers to this question; $S = \{\text{truthful, careless}\}$. Suppose I know from experience that Fred's announcements are truthful reports on what he knows 80% of the time and are careless statements the other 20% of the time. Then I have a probability measure P over S: $P\{\text{truthful}\} = .8$, $P\{\text{careless}\} = .2$.

Are the streets outside slippery? Let T denote the possible answers to this question; $T = \{\text{yes, no}\}$. And suppose Fred's announcement turns out to be, "The streets outside are slippery." Taking account of this, I have a compatibility relation between S and T: truthful is compatible with yes but not with no, while careless is compatible with both yes and no. Applying (1), I find

$$\text{(2)} \qquad \mathrm{Bel}(\{\text{yes}\}) = .8 \text{ and } \mathrm{Bel}(\{\text{no}\}) = 0;$$

Fred's announcement gives me an 80% reason to believe the streets are slippery, and no reason to believe they are not.

How might a Bayesian argument using this evidence go? A Bayesian direct probability argument would use all my evidence, Fred's announcement included, to make a direct probability judgment about whether the

streets are slippery. The judgment that Fred is 80% reliable need not appear explicitly in such an argument. On the other hand, I can construct a Bayesian conditional probability argument using this judgment as one ingredient. I need two other judgments as well: (i) A prior probability, say p, for the proposition that the streets are slippery; this will be a judgment based on evidence other than Fred's announcement. (ii) A conditional probability, say q, that Fred's announcement will be accurate even though it is careless. I can construct a probability measure from these judgments, and I can condition this measure on the content of Fred's announcement.

The probability measure constructed in this conditional argument is formally a measure over $S \times T$, where T is still the set of answers to the question whether the streets are slippery,

$$T = \{\text{yes, no}\},$$

but where S now tells us not only whether Fred is truthful or careless but also whether he is accidentally telling the truth in case he is careless,

$$S = \{\text{truthful, careless but accurate, careless and inaccurate}\}.$$

My probabilities for T are p for yes and $1 - p$ for no. My probabilities for S are .8 for truthful, $.2q$ for careless but accurate, and $.2(1 - q)$ for careless and inaccurate. Assuming probabilistic independence between the state of the streets and Fred's behavior, I multiply these numbers to obtain the product probability measure on $S \times T$, given in the second column of Table 1. Conditioning this measure on the content of Fred's announcement means eliminating the three possibilities marked with an × in the table; since Fred said the streets are slippery, he cannot be truthful or accurate if the answer to T is no, and he cannot be inaccurate if the answer to T is yes. Having eliminated these three possibilities, I renormalize the probabilities for the other three so that they add to one; this means multiplying each probability by K, where $K = 1/(.8p + .2qp + .2(1 - q)(1 - p))$. This results in the posterior probabilities given in the third column in Table 1. Adding the first two nonzero probabilities in this column, I obtain my total posterior probability that Fred's announcement that the streets are slippery is true:

$$(3) \qquad \frac{.8p + .2qp}{.8p + .2qp + .2(1 - q)(1 - p)}.$$

Is the Bayesian argument (3) better than the belief-function argument (2)? This depends on whether I have the evidence required. If I do have evidence to support the judgments p and q—if, that is to say, my situation really is quite like a situation where the streets and Fred are governed by these known chances, then (3) is a cogent argument, and it is better than (2) because it takes more evidence into account. But if the evidence on which I base p and q is of much lower quality than the evidence on which I base the number 80%, then (2) will be the better argument.

The traditional debate between the frequentist and Bayesian views has centered on the quality of the evidence for prior probabilities. It is worth remarking, therefore, that q, rather than p, may well be the weak point in the argument (3). I probably will have some other evidence about whether it is slippery outside, but I may have no idea about how likely it is that Fred's careless remarks will accidentally be true.

A critic of the belief-function argument (2) might be tempted to claim that the Bayesian argument (3) shows (2) to be wrong even if I do lack the evidence needed to supply p and q. Formula (3) gives the correct probability for whether the street is slippery, the critic might contend, even if I cannot say what this probability is, and it is almost certain to differ from (2). This criticism is fundamentally misguided. In order to say that (3) gives the "correct" probability, I must be able to convincingly compare my situation to the picture of chance. And my inability to model Fred when he is being careless is not just a matter of not knowing the chances—it is a matter of not being able to fit him into a chance picture at all.

Example 2. Suppose I do have some other evidence about whether the streets are slippery: my trusty indoor–outdoor thermometer says that the

TABLE 1

	(s, t)	Probability of (s, t)	
		Initial	Posterior
	(Truthful, yes)	$.8p$	$.8pK$
×	(Truthful, no)	$.8(1 - p)$	0
	(Careless but accurate, yes)	$.2qp$	$.2qpK$
×	(Careless but accurate, no)	$.2q(1 - p)$	0
×	(Careless and inaccurate, yes)	$.2(1 - q)p$	0
	(Careless and inaccurate, no)	$.2(1 - q)(1 - p)$	$.2(1 - q)(1 - p)K$

TABLE 2

s	Probability of s		Elements of T compatible with s
	Initial	Posterior	
(Truthful, working)	.792	0	
(Truthful, not)	.008	.04	Yes
(Careless, working)	.198	.95	No
(Careless, not)	.002	.01	Yes, no

temperature is 31° Fahrenheit, and I know that because of the traffic ice could not form on the streets at this temperature.

My thermometer could be wrong. It has been very accurate in the past, but such devices do not last forever. Suppose I judge that there is a 99% chance that the thermometer is working properly, and I also judge that Fred's behavior is independent of whether it is working properly or not. (For one thing, he has not been close enough to my desk this morning to see it.) Then I have determined probabilities for the four possible answers to the question, "Is Fred being truthful or careless, and is the thermometer working properly or not?" For example, I have determined the probability $.8 \times .99 = .792$ for the answer "Fred is being truthful, and the thermometer is working properly." All four possible answers, together with their probabilities, are shown in the first two columns of Table 2. I will now construct a belief function over T by using these four answers as my frame S.

Taking into account what Fred and the thermometer have said, I obtain the compatibility relation between S and T given in the last column of the Table 2. (Recall that T considers whether the streets are slippery; $T = \{\text{yes, no}\}$.) The element (truthful, working) of S is ruled out by this compatibility relation (since Fred and the thermometer are contradicting each other, they cannot both be on the level); hence, I condition the initial probabilities by eliminating the probability for (truthful, working) and renormalizing the three others. The resulting posterior probabilities on S are given in the third column of the Table 2.

Finally, applying (1) with these posterior probabilities on S, I obtain the degrees of belief

(4) $\quad \text{Bel}(\{\text{yes}\}) = .04 \text{ and } \text{Bel}(\{\text{no}\}) = .95.$

This result reflects that fact that I put much more trust in the thermometer than in Fred.

The preceding calculation is an example of *Dempster's rule of combination* for belief functions. Dempster's rule combines two or more belief functions defined on the same frame but based on independent arguments or items of evidence; the result is a belief function based on the pooled evidence. In this case the belief function given by (2), which is based on Fred's testimony alone, is being combined with the belief function given by

(5) $\quad \text{Bel}(\{\text{yes}\}) = 0 \text{ and } \text{Bel}(\{\text{no}\}) = .99,$

which is based on the evidence of the thermometer alone. In general, as in this example, Dempster's rule corresponds to the formation and subsequent conditioning of a product measure in the background. See Shafer (1986a) for a precise account of the independence conditions needed for Dempster's rule.

Example 3. Dempster's rule applies only when two items of evidence are independent, but belief functions can also be derived from models for dependent evidence.

Suppose, for example, that I do not judge Fred's testimony to be independent of the evidence provided by the thermometer. I exclude the possibility that Fred has tampered with the thermometer and also the possibility that there are common factors affecting both Fred's truthfulness and the thermometer's accuracy. But suppose now that Fred does have regular access to the thermometer, and I think that he would likely know if it were not working. And I know from experience that it is in situations where something is awry that Fred tends to let his fancy run free.

In this case, I would not assign the elements of S the probabilities given in the second column of Table 2. Instead, I might assign the probabilities given in the second column of Table 3. These probabilities follow from my judgment that Fred is truthful 80% of the time and that the thermometer has a 99% chance of working, together with the further judgment that Fred has a 90% chance of being careless if the thermometer is not working.

When I apply (1) with the posterior probabilities given in Table 3, I obtain the degrees of belief

$$\text{Bel}(\{\text{yes}\}) = .005 \text{ and } \text{Bel}(\{\text{no}\}) = .95.$$

These differ from (4), even though the belief functions based on the separate items of evidence will still be given by (2) and (5).

In this example, the combination of two belief functions (2) and (5) departed from Dempster's rule in that the probability measure constructed over the joint probability space in the background was not a product measure. This is just one of the ways the language of belief functions can take dependence into account.

TABLE 3

s	Probability of s		Elements of T compatible with s
	Initial	Posterior	
(Truthful, working)	.799	0	
(Truthful, not)	.001	.005	Yes
(Careless, working)	.191	.950	No
(Careless, not)	.009	.045	Yes, no

Another way is to modify the compatibility relation between the joint probability space and the frame T (Shafer, 1986a). Another is to rework the way the evidence is broken up, so that different items of evidence better correspond to independent uncertainties (Shafer, 1984).

2.5 Conclusion

I would like to emphasize that nothing in the philosophy of constructive probability or the language of belief functions requires us to deny the fact that Bayesian arguments are often valuable and convincing. The examples I have just discussed were designed to convince the reader that belief-function arguments are sometimes more convincing than Bayesian arguments, but I am not claiming that this is always or even usually the case. What the language of belief functions does require us to reject is the philosophy according to which use of the Bayesian language is normative.

From a technical point of view, the language of belief functions is a generalization of the Bayesian language. But as our examples illustrate, the spirit of the language of belief functions can be distinguished from the spirit of the Bayesian language by saying that a belief-function argument involves a probability model for the evidence bearing on a question, whereas a Bayesian argument involves a probability model for the answer to the question.

Of course, the Bayesian language can also model evidence. The probability judgments made in a belief-function argument can usually be extended to a Bayesian argument that models both the answer to the question and the evidence for it by assessing prior probabilities for the answer and conditional probabilities for the evidence given the answer. The only problem is that we may lack the evidence needed to make all the judgments required by this Bayesian argument convincing. The advantage gained by the belief-function generalization of the Bayesian language is the ability to use certain kinds of incomplete probability models.

3. THE ATTEMPT TO USE PROBABILITY IN PRODUCTION SYSTEMS

The field of expert systems developed within artificial intelligence from efforts to apply systems of production rules to practical problems. The current interest in probability judgment in artificial intelligence began with efforts to incorporate probability judgments into production rules. In this section I review these efforts and relate them to what we learned in the preceding section about the Bayesian and belief-function languages.

A production rule is simply an if-then statement, interpreted as an instruction for modifying the contents of a data base. When the rule is applied, the action specified by its right-hand side is taken if the condition on its left-hand side is found in the data base. A production system is a collection of production rules, which are repeatedly applied to the data base either in the same predetermined order or else in an order determined by some relatively simple principle. Production systems were used in programming languages in the early 1960s, and they were advanced as cognitive models by Newell and Simon in the late 1960s and early 1970s (Newell and Simon, 1965; Newell, 1973). These systems are attractive models for intelligence because their knowledge is represented in a modular way and is readily available for use. Each rule represents a discrete chunk of knowledge that can be added to or removed from the system without disrupting its ability to use the other chunks, and the system regularly checks all the chunks for their relevance to the problem at hand (Davis and King, 1984).

When artificial intelligence workers undertook, in the 1970s, to cast various bodies of practical knowledge in the form of production rules, they found that in many fields knowledge cannot be encoded in the form of unqualified if-then statements. Instead, probability statements seem to be required: "If $E_1, E_2, \ldots E_n$, then probably (or usually or almost certainly) H." So these workers found themselves trying to use production systems to manipulate probability judgments.

Many tacks were taken in the effort to use probability in production systems, but I would like to emphasize two lines of development. One of these begins with PROSPECTOR and leads to Pearl and Kim's elegant work on the propagation of Bayesian probability judgments in causal trees, while the other begins with the certainty factors of MYCIN and leads to the use of belief functions in diagnostic trees. I will review these two lines of development in turn.

As it turns out, the results of both lines of development can be unified in a general scheme for propagating belief functions in trees (Shenoy and Shafer, 1986). I will briefly describe this general scheme.

3.1 Bayesian Networks

The artificial intelligence workers at SRI who developed the PROSPECTOR system for geological exploration in the middle 1970s thought of production rules as a means for propagating probabilities through a network going from evidence to hypotheses. Figure 1, taken from Duda, Hart and Nilsson (1976), gives an example of such a network; here, E_i denotes an item of evidence, and H_j denotes a hypothesis. The idea is that the user of the system should specify that

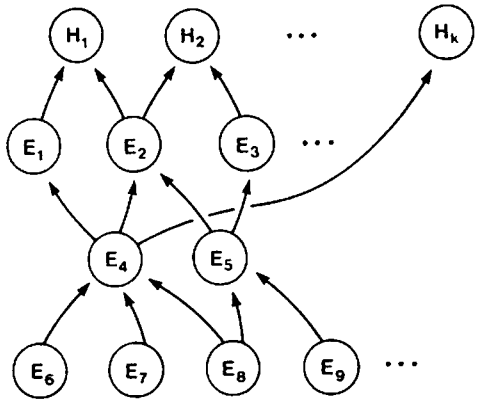

FIG. 1. *PROSPECTOR's inference network.*

some of the E_i at the bottom of the network are true and some are false, or should make probability judgments about them, and the production rules, corresponding to conditional probabilities for the links in the network, should propagate these probability judgments through the network to produce judgments of the probabilities of the hypotheses.

Unfortunately, the introduction of probabilities into production rules does not square well with the modularity we want these rules to have. The PROSPECTOR workers wanted to be able to elicit from a geologist statements of the form, "If E_i and E_j and . . . , then E_r, with probability p," and they wanted to allow the geologist to make each of these statements independently. But this led to problems in putting the statements together into a calculation of the probabilities of the hypotheses. For example: (1) The conditional probabilities elicited may not be sufficient to determine a joint probability measure over all the E's and H's. The geologist might give rules corresponding to $P(E_5 \mid E_8)$ and $P(E_5 \mid E_9)$ in Figure 1 but neglect or feel unable to give a rule corresponding to $P(E_5 \mid E_8 \,\&\, E_9)$. (2) The conditional probabilities that are given may be inconsistent. (3) The network may have cycles, which will cause trouble when propagation is attempted.

These problems were handled in PROSPECTOR in relatively ad hoc ways. Problem (1) was handled partly by independence assumptions and partly by maximum-minimum rules reminiscent of the theory of fuzzy sets. Problem (2) was handled by formulating rules of propagation which did not always accord with the rules of probability but which were insensitive to some kinds of inconsistencies. Problem (3) was handled by arbitrarily rejecting new production rules when they would introduce cycles into the network already constructed.

PROSPECTOR was only modestly successful, but it was very influential in the questions it raised. The PROSPECTOR workers subscribed to Bayesian principles, and they were conscious of their failure to follow those principles completely. Is it possible to do better? Can probability judgments be treated modularly within the Bayesian language? To what extent is the propagation of probabilities possible within this language?

The best work that has been done in response to these questions is that of Judea Pearl and his students at UCLA (Pearl, 1982, 1986; Kim, 1983; Kim and Pearl, 1983). Pearl has shown that we can make sense of the independence assumptions needed to construct a probability measure over a network from simple conditional probabilities and we can propagate updated probabilities through the network in a simple and elegant way provided that the network has a causal interpretation and a relatively simple form: it must be a simple directed tree or else a more general type of directed tree that we may call a *Kim tree*.

Recall that a tree is a graph in which there are no cycles. A simple directed tree is a tree in which the links are assigned directions that all run outward (or downward, if we want) from a single initial node, as in Figure 2a. A Kim tree is a tree in which the links are assigned arbitrary directions. Such a tree can always be laid out so that the directions are downward, as in Figure 2b. In Pearl's work, the nodes of a tree correspond to random variables, and the directions of the links are interpreted as directions of causation. Thus each variable is influenced by the variables above it in the graph and influences the variables below it. An observation of the value of one variable is diagnostic evidence about the value of a higher variable and causal evidence about the value of a lower variable.

Once a Kim tree is constructed for a problem, the construction of a probability measure over it and the updating of the measure are straightforward. Given the independence conditions of Pearl and Kim, which are reasonable in the causal context, a measure over the tree can be constructed from prior probabilities for the topmost nodes and conditional probabilities for all the links. Moreover, this construction is straightforward; there are no complicated consistency conditions that the conditional probabilities must meet. Once construction is completed, the measure can be stored and updated locally. At each node we store information about the conditional probabilities corresponding to incoming and outgoing links, the current probability measures for the variable at the node and the variables at neighboring higher nodes, and likelihood-type information from neighboring lower nodes. When the value of a variable is then observed, this information can be propagated through the network to update the entire probability measure in one pass. All computations are made locally, with

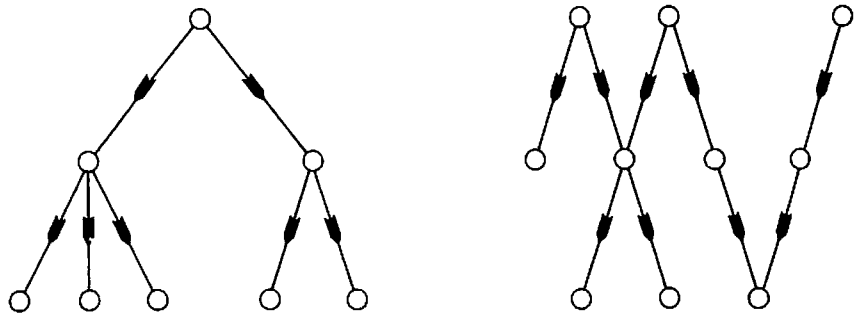

FIG. 2. *Pearl's causal tree.*

each node communicating only updated local information to its neighbors.

An obvious shortcoming of this elegant scheme is its restriction to Kim trees. In few problems will the causal relations that we think important take so simple a form. Kim (1983) and Pearl (1986) have shown how such trees might be used to approximate more realistic models; they propose first using a more general graph to elicit a probability measure from an expert, and then approximating this measure with a Kim tree. This solution does not seem very satisfactory, however. It is not clear that the approximation will be satisfactory, and more importantly, the constructive nature of the initial probability measure is put into question. In a Kim tree the initial probability measure can be constructed from probability and conditional probability judgments without concerns about consistency, but in a more general graph consistency conditions will be so complicated that it will be impossible for us to hope they will be met unless we pretend that we are indeed eliciting a measure instead of constructing one.

Another obvious shortcoming is the restriction to thoroughly causal models. In a sense, of course, all evidence is causal. With sufficient complication, we can always construct a model that relates the facts we observe to deeper causes and also relates these causes to the questions that interest us. But we may lack the evidence needed to make good probability judgments relative to such a model.

3.2 Certainty Factors and Belief Functions

The work on the MYCIN system for medical diagnosis began earlier and has been more extensive than the work on PROSPECTOR. It has also had more effect on subsequent expert systems; various versions of EMYCIN, the expert system shell that was abstracted from MYCIN, are now being widely used. The story of the MYCIN effort has been told in a recent book (Buchanan and Shortliffe, 1984), which includes extensive discussion of the certainty factors that were used by MYCIN and the similarities of these certainty factors to the values of belief functions.

MYCIN departed from the pure production system picture by using a backward-chaining strategy to select production rules to apply. This means that it selected rules by comparing their right-hand sides to goals instead of comparing their left-hand sides to statements already accepted. If the right-hand side of a rule matched a goal, its left-hand side was then established as a goal, so that there was a step by step process backward from conclusions to the knowledge needed to establish them.

MYCIN also differed from PROSPECTOR in that the MYCIN workers rejected at the outset the idea that the numerical probability judgments associated with the rules could or should be understood in Bayesian terms. They emphasized this point by calling these numbers "certainty factors" rather than probabilities. And they formulated their own rules for combining these certainty factors.

In spirit, and to a considerable extent in form, these rules agree with Dempster's rule for combining independent belief functions. I would explain this coincidence by saying that in developing their calculus for certainty factors, Shortliffe and Buchanan were trying to model the probabilistic nature of evidence while avoiding the complete probability models needed for Bayesian arguments.

In recent work (Gordon and Shortliffe, 1984, 1985), some of the MYCIN workers have taken a close look at the similarity between the calculus of certainty factors and the language of belief functions and have asked how belief functions can contribute further to the MYCIN project. They have drawn two main conclusions. First, it is sensible to modify some of the rules for certainty factors to put these rules into more exact agreement with the rules for belief functions. Second, the diagnosis problem that was central to MYCIN can be understood more clearly in terms of belief functions if it is explicitly expressed as a problem involving hierarchical hypotheses.

The term "hierarchical hypotheses" refers to the fact that the items of evidence in a diagnostic problem tend to support directly only certain subsets of the frame of discernment, subsets which can be arranged in a tree. Figure 3, taken from Gordon and Shortliffe (1984), illustrates the point. The four nodes at the bottom of this tree represent four distinct causes of cholestatic jaundice; they form the frame of discernment for the diagnostic problem. Some items of evidence may directly support (or directly refute) one of these causes for a particular patient's jaundice. Other evidence may be less specific. There may, for example, be evidence that the jaundice is due to an intrinsic liver problem, either hepatitis or cirrhosis. On the other hand, it is hard to imagine a single item of medical evidence supporting the subset {cirrhosis, gallstone} without supporting one of these more directly: this is reflected by the fact that this subset does not correspond to an intermediate node of the tree.

This picture suggests that a belief-function argument based on such medical evidence may involve combining many belief functions by Dempster's rule, where each belief function is a simple support function focused on a subset in the tree or its complement. (A simple support function is a belief function obtained from (1) when S has only two elements and one of these is compatible with all the elements of T.)

Two concerns can be raised about this use of Dempster's rule. First, there is the issue of computational complexity. Since the computational complexity of Dempster's rule increases exponentially with the size of the frame, it might not be feasible to implement the rule for a large diagnostic tree. Second, there is the issue of dependence. Will the items of evidence bearing on different nodes of the tree all be independent?

As it turns out, computational complexity is not a problem. By taking advantage of the tree structure, we can devise remarkably efficient algorithms for implementing Dempster's rule (Shafer and Logan, 1985).

Violations of the independence assumptions needed for Dempster's rule pose a more worrisome problem. It seems unlikely that the uncertainties involved in a very large number of items of medical evidence will all be independent. This does not mean that a belief-function analysis will be impossible or unsatisfactory, but it does mean that a satisfactory belief-function analysis may require modeling dependencies in the evidence.

3.3 Propagating Belief Functions in Trees

It turns out that Pearl's method of propagating Bayesian probabilities in causal trees and Shafer and Logan's method of combining simple support functions in diagnostic trees are both special cases of a general scheme for propagating belief functions in qualitative Markov trees. The following comments on this general scheme are relatively technical but may be of interest to some readers. For more detail, see Shenoy and Shafer (1986).

The idea of a qualitative Markov tree is based on the idea of qualitative conditional independence. We say that two partitions $\mathbf{P}_1$ and $\mathbf{P}_2$ of a frame S are *conditionally independent* given a third partition P if $P \cap P_1 \cap P_2 \neq \emptyset$ whenever $P_1 \in \mathbf{P}_1$, $P_2 \in \mathbf{P}_2$, $P \cap P_1 \neq \emptyset$, *and* $P \cap P_2 \neq \emptyset$. This means that once we know which element of $\mathbf{P}$ contains the truth, knowledge of which element of $\mathbf{P}_1$ contains the truth tells us nothing more about which element of $\mathbf{P}_2$ contains the truth. Qualitative conditional independence is important for belief functions, because it is legitimate, when $\mathbf{P}_1$ and $\mathbf{P}_2$ are conditionally independent given $\mathbf{P}$, and we want to combine a belief function on $\mathbf{P}_1$ with a belief function on $\mathbf{P}_2$, to first simplify both to belief functions on $\mathbf{P}$. This can be helpful if $\mathbf{P}$ is a relatively coarse partition, for then the combination is easier to think about and computationally more feasible.

A qualitative Markov tree is a tree of partitions with the property that the disconnected branches that

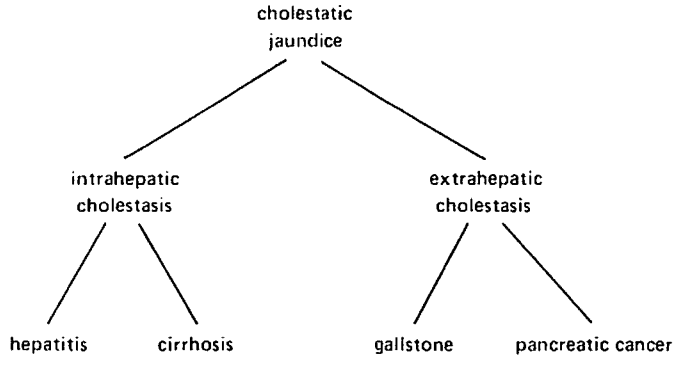

FIG. 3. *A diagnostic tree.*

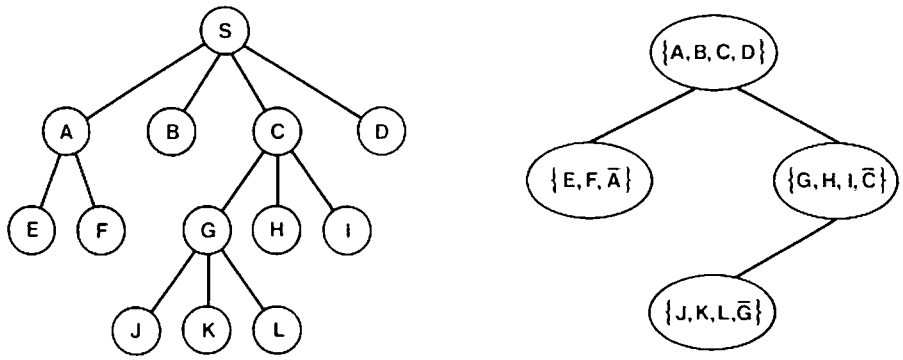

FIG. 4. *The tree of partitions. (b) derived from a diagnostic tree (a).*

result from the removal of a partition **P** are always conditionally independent given **P**. We obtain a qualitative Markov tree if we replace each random variable in a Bayesian causal tree with the partition of the sample space it induces. We can also construct a qualitative Markov tree from a diagnostic tree; for each mother node in the diagnostic tree, we form a partition whose elements are the daughters of the mother and the complement of the mother. Figure 4b shows the qualitative Markov tree obtained in this way from the diagnostic tree of Figure 4a.

Suppose we wish to combine belief functions defined on various partitions in a qualitative Markov tree. It is legitimate to do so in a stepwise way, simplifying the belief function on one partition to a belief function on its neighbor, combining all the belief functions projected to the neighbor in this way, and then projecting to the next neighbor. The schemes of Pearl and Shafer and Logan both turn out to be special cases of this simple general idea.

In addition to generalizing Pearl and Shafer and Logan, this scheme for propagating belief functions in trees promises to be useful as a general framework for designing probability arguments. Independent items of evidence often bear on different but related partitions (or questions, or variables), and a qualitative Markov tree provides a way of keeping track of the relations.

3.4 Conclusion

The preceding look at attempts to use probability judgment in expert systems justifies at least one general conclusion: probability judgment in expert systems is very much like probability judgment everywhere else. The general issues about probability judgment that we identified in Section 2 all reappear in the expert systems work. In expert systems, as elsewhere, probability judgment is constructive and requires an overall design. It is sometimes possible to provide such a design within the Bayesian language, but Bayesian designs often demand judgments for which we do not have adequate evidence. And belief-function analyses often require models for dependent evidence.

Production systems were attractive to the artificial intelligence community because these systems seemed to have the flexibility in acquiring and using knowledge that seems characteristic of intelligence. But it seems fair to say that the attempt to incorporate probability judgment into production systems has failed. The most successful production systems are still those, like R1 and DART, that do not attempt to use numerical measures of uncertainty. Many expert systems have recently been built using the EMYCIN shells, but more often than not the builders of these systems ignore the "certainty factor" capacities of the shells.

It appears that probability judgment simply does not have the extremely modular character that made production systems so attractive. Almost always, probability judgment involves not only individual numerical judgments but also judgments about how these can be put together. This is because probability judgment consists, in the final analysis, of a comparison of an actual problem to a scale of canonical examples.

I believe that progress will be made over the next few years in using probability in expert systems. But these systems will be intensely interactive. They will depend on the human user to design the probability argument for the particular evidence at hand: they will be able at most to help the user construct his or her causal, diagnostic, or qualitative Markov tree. And they will also depend on the human user to supply individual numerical probability judgments.

4. THE CONSTRUCTION OF ARGUMENTS

A genuine capacity for probability judgment in an artificial intelligence would involve both the ability to generate numerical probability judgments and the ability to design probability arguments. How might

these abilities be programmed? We do not have an answer, but we should start thinking about the question.

As the result of the work by psychologists during the past decade, especially the work of Kahneman and Tversky (see Kahneman, Slovic, and Tversky, 1982), we do have some ideas about how people generate numerical probability judgments. They conduct internal sampling experiments, they make similarity judgments, they construct causal models and perform mental simulations with these models, they consider typical values and discount or adjust these, and so on. An obvious and appropriate strategy for artificial intelligence is to try to implement these heuristics.

The heuristics sometimes lead to systematic mistakes or biases, and it is by demonstrating these biases that the psychologists have convinced us that people use them. There is a tendency, therefore, to think that people are doing something suboptimal or unnormative when they use them. Indeed, proponents of the Bayesian philosophy frequently assert that the psychological work only demonstrates what people do and is irrelevant to what people should do. When we face up to the artificial intelligence problem, however, we see that the heuristics are really all we have. People have to use such heuristics if they are to make quick probability judgments about questions they have not previously considered, and our programs will also have to use them if they are going to be equally flexible. The challenge is to figure out how to use the heuristics well enough that using them will not usually cause mistakes.

It is more difficult to say anything about how we might build the ability to design probability arguments. The lesson from Section 3 is clear, though: the chunks that we try to fit together when we search for a convincing argument must be larger than the chunks represented by production rules. It is also clear that the ability to construct cogent probability arguments must include an ability to evaluate whether a probability argument is cogent.

I believe that our ability to build systems with human-like capabilities in designing probability arguments and generating numerical probability judgments will ultimately depend on our ability to build associative memories. With a genuine associative memory, we could retrieve stored experiences that approximately match any arbitrary new situation, not just those that match a relatively few situations we might specify in advance. The retrieval of such stored experiences on a fine scale would permit us to calculate frequencies that could serve as numerical probability judgments, and the comparison to other problems on a coarser scale could give hints for the design of a probability argument. Associative memory is currently an active and exciting field of research in artificial intelligence (Hinton and Anderson, 1981; Hopfield, 1982: Kohonen, 1984). It is a field where statisticians should be making a greater contribution than they are.

The entire field of artificial intelligence poses a challenge to students of probability. I believe that probability judgment will turn out to be possible and important in artificial intelligence, but the extent of its ultimate usefulness cannot be taken for granted; it must be demonstrated.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation under Grants IST-8405210 and IST-8610293.

REFERENCES

BARR, A. and FEIGENBAUM, E. A., eds. (1981, 1982). *The Handbook of Artificial Intelligence*, **1**, **2**. William Kaufmann, Los Altos, Calif.

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. Roy. Statist. Soc. Ser. A* **143** 383–430.

BUCHANAN, B. G. and SHORTLIFFE, E. H., eds. (1984). *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, Mass.

COHEN, P. R. (1985). *Heuristic Reasoning about Uncertainty: An Artificial Intelligence Approach*. Pitman, Boston.

COHEN, P. R. and FEIGENBAUM, E. A., eds. (1982). *The Handbook of Artificial Intelligence*, **3**. William Kaufmann, Los Altos, Calif.

DAVIS, R. and KING, J. J. (1984). The origin of rule-based systems in AI. In Buchanan and Shortliffe (1984), 20–52.

DUDA, R. O., HART, P. E. and NILSSON, N. J. (1976). Subjective Bayesian methods for rule-based inference systems. *AFIPS Conf. Proc.* **45** 1075–1082. (Reprinted in *Readings in Artificial Intelligence*, edited by B. L. Webber and N. J. Nilsson. Tioga Publishing, Palo Alto, Calif., 1981, pages 192–199.)

GORDON, J. and SHORTLIFFE, E. H. (1984). The Dempster-Shafer theory of evidence. In Buchanan and Shortliffe (1984), 272–292.

GORDON, J. and SHORTLIFFE, E. H. (1985). A method for managing evidential reasoning in hierarchical hypothesis spaces. *Artificial Intelligence* **26** 323–358.

HINTON, G. E. and ANDERSON, J. A., eds. (1981). *Parallel Models of Associative Memory*. Erlbaum, Hillsdale, N. J.

HOPFIELD, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Nat. Acad. Sci. U.S.A.* **79** 2554–2558.

KAHNEMAN, D., SLOVIC, P. and TVERSKY, A. (1979). *Judgments under Uncertainty: Heuristics and Biases*. Cambridge Univ. Press, Cambridge.

KIM, J. H. (1983). CONVINCE: A conversational inference consolidation engine. Ph.D. dissertation, Dept. of Computer Science, Univ. California, Los Angeles.

KIM, J. H. and PEARL, J. (1983). A computational model for causal and diagnostic reasoning in inference systems. *Proc. of the Eighth International Joint Conference on Artificial Intelligence* 190–193. William Kaufmann, Los Altos, Calif.

KOHONEN, T. (1984). *Self-Organization and Associative Memory*. Springer, New York.

MARR, D. (1982). *Vision*. Freeman, San Francisco.

McCarthy, J. (1980). Circumspection—a form of non-monotonic reasoning. *Artificial Intelligence* **13** 27–39.

McDermott, D. and Doyle, J. (1980). Non-monotonic logic I. *Artificial Intelligence* **13** 41–72.

Newell, A. (1973). Production systems: models of control structures. In *Visual Information Processing* (W. G. Chase, ed.) 463–526. Academic, New York.

Newell, A. and Simon, H. A. (1965). An example of human chess play in the light of chess playing programs. In *Progress in Biocybernetics* (N. Wiener and J. P. Schade, eds.). Elsevier, Amsterdam.

Pearl, J. (1982). Reverend Bayes on inference engines: a distributed hierarchical approach. *Proc. of the Second National Conference on Artificial Intelligence, American Association for Artificial Intelligence* 133–136. William Kaufmann, Los Altos, Calif.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence* **29** 241–288.

Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence* **13** 81–132.

Shafer, G. (1976). *A Mathematical Theory of Evidence.* Princeton Univ. Press, Princeton, N. J.

Shafer, G. (1981). Constructive probability. *Synthese* **48** 1–60.

Shafer, G. (1984). The problem of dependent evidence. Working paper no. 164, School of Business, Univ. Kansas.

Shafer, G. (1986a). Belief functions and possibility measures. To appear in *The Analysis of Fuzzy Information* (J. C. Bezdek, ed.) 1. CRC Press.

Shafer, G. and Logan, R. (1985). Implementing Dempster's rule for hierarchical evidence. Working paper no. 174, School of Business, Univ. Kansas. To appear in *Artificial Intelligence.*

Shafer, G. and Tversky, A. (1985). Languages and designs for probability judgment. *Cognitive Sci.* **9** 309–339.

Shenoy, P. and Shafer, G. (1986). Propagating belief functions with local computations. *IEEE Expert.* **1** 43–52.

Statistical Science
1987, Vol. 2, No. 1, 3–44

The Probability Approach to the Treatment of Uncertainty in Artificial Intelligence and Expert Systems

Dennis V. Lindley

Abstract. Arguments are adduced to support the claim that the only satisfactory description of uncertainty is probability. Probability is described both mathematically and interpretatively as a degree of belief. The axiomatic basis and the use of scoring rules in developing coherence are discussed. A challenge is made that anything that can be done by alternative methods for handling uncertainty can be done better by probability. This is demonstrated by some examples using fuzzy logic and belief functions. The paper concludes with a forensic example illustrating the power of probability ideas.

Key words and phrases: Artificial intelligence, expert systems, probability, scoring rules, coherence, decision-making, Bayes theorem, fuzzy logic, belief functions, forensic evidence.

1. INTRODUCTION

Our concern in this paper is not with a general discussion of artificial intelligence (AI) and expert systems (ES) but with one particular aspect of them, namely the occurrence of uncertainty statements within AI or ES. We discuss how they should be made, what they mean, and how they combine together.

Uncertainty is obviously present in most ES algorithms because experts can rarely be totally sure of the statements they make. Thus, in medical ES, the presence of a symptom array does not invariably imply the presence of one disease, so that diagnosis is inherently uncertain. Even the symptom may exhibit uncertainty for doctors may differ in their interpretations (see Section 10). Prognosis is clearly even more uncertain. When discussing purely deterministic procedures there may be some merit in introducing uncertainty. For example, chess is a game with perfect information yet AI programs sometimes incorporate uncertainty as a way of avoiding the terrible complexity of the game. So uncertainty, while perhaps not ubiquitous, frequently occurs. Our task is to study approaches to dealing with it within AI and ES.

Dennis V. Lindley was, until his retirement in 1977, Head of the Department of Statistics at University College, London. He states that "The present paper was written because of a strong conviction that probability is the only satisfactory description of uncertainty." His mailing address is 2 Periton Lane, Minehead, Somerset TA24 8AQ, England.

2. THE INEVITABILITY OF PROBABILITY

Our thesis is simply stated: *the only satisfactory description of uncertainty is probability.* By this is meant that every uncertainty statement must be in the form of a probability; that several uncertainties must be combined using the rules of probability; and that the calculus of probabilities is adequate to handle *all* situations involving uncertainty. In particular, alternative descriptions of uncertainty are unnecessary. These include the procedures of classical statistics; rules of combination such as Jeffrey's (1965); possibility statements in fuzzy logic, Zadeh (1983); use of upper and lower probabilities, Smith (1961), Fine (1973); and belief functions, Shafer (1976). We speak of "the inevitability of probability."

3. MATHEMATICAL AND PHYSICAL MEANINGS FOR PROBABILITY

Before defending the thesis, it had better be made clear what we mean by probability. Most emphatically, we do not just mean numbers lying between 0 and 1: it is more interesting than that. There are two ways of responding to a question about the meaning of probability. One is to describe the concept mathematically. The other is to consider its interpretation in

the physical world. We consider both of these responses.

Mathematically, probability is a function of two arguments: the event A about which you are uncertain, and your knowledge H when you make the uncertainty statement. We write $p(A \mid H)$; read, the probability of A, given H. The function obeys the three rules:

Convexity $0 \leq p(A \mid H) \leq 1$ and $p(A \mid H) = 1$ if H is known by you logically to imply A.

Addition $p(A_1 \cup A_2 \mid H) = p(A_1 \mid H) + p(A_2 \mid H) - p(A_1 \cap A_2 \mid H)$.

Multiplication $p(A_1 \cap A_2 \mid H) = p(A_1 \mid H) \cdot p(A_2 \mid A_1 \cap H)$.

We could elaborate on these rules, for example, by discussing whether the events have to form a σ-field, whether the addition law holds for an enumerable infinity of events, whether $p(A \mid H) = 1$ *only* if H is known by you logically to imply A, and in other ways. But these would merely add mathematical glosses to the key ideas that probability lies between 0 and 1 and obeys two distinct rules of combination. From these three rules, perhaps modified slightly, all of the many, rich and wonderful results of the probability calculus follow. They may be described as the axioms of probability. We prefer not to describe them this way because, as will be seen below, they can be derived from other, more basic, axioms and consequently appear as theorems.

The interpretation of $p(A \mid H)$ is that it is your subjective belief in the truth of A were you to know that H was true. It is often referred to as subjective probability because it is ascribable to a subject, you; and also to distinguish it from another use of probability called frequentist or objective. This latter we shall call *chance*, thus avoiding the adjective for probability. It is convenient to think of $p(A \mid H)$ as a measurement: like a measurement of length or temperature. It measures belief, not temperature. Like all measurements it has a standard. We may take the simple example of balls in an urn. For you, $p(A \mid H) = a$ if you are indifferent between receiving a prize contingent on A, knowing H, and receiving the same prize contingent on a black ball being drawn at random from an urn containing a proportion a of black balls. Of course, other ways are possible. It is a defect of many other approaches to the measurement of uncertainty that they do not have a standard by which to judge their statements.

4. THE USE OF SCORING RULES

Having interpreted probability in two, important ways, let us turn to the defense of the thesis of the inevitability of probability. The task is to study uncertainty, particularly in the context of AI and ES. As scientists and engineers we would expect to measure our object of study, to describe the uncertainty numerically. If we agree to do this, we have to decide what rules the numbers obey: for example, can we add them, like lengths? One way is to think of possible rules and choose some that seem reasonable. This is the method of classical statistics, fuzzy logic, and belief functions. There is another method.

Suppose that in expressing your belief in A, given H, you provide a numerical value a. In what sense is a a "good" measurement of your belief? De Finetti (1974/5) had the idea of introducing a score function, which scores your measurement or, as we usually prefer, your assessment of your uncertainty of A, given H. For the two functions, f_0 and f_1, the score, when a is announced as the assessment, is defined to be:

$f_1(a)$ if both A and H are true.

$f_0(a)$ if H is true, but A false, and

zero if H is false.

De Finetti used the quadratic or Brier score: $f_0(a) = a^2$, $f_1(a) = (1 - a)^2$. With the quadratic, a near 1(0) will give a low score when A is true (false) and H true. If H is false the statement about A is irrelevant since it was made on the supposition of H.

Suppose now that you, or the expert in ES, does this with several event pairs; (A_i, H_i) is scored on each and the scores added. Then de Finetti showed for the quadratic rule, that the values a_i must obey the rules of probability. Lindley (1982) generalized the result and showed that virtually any score leads to probability: some scores are eccentric and result in only two possible values for a whatever be A and H. A consequence of de Finetti's result is that someone using rules for the combination of the a_i that are not probabilistic—for example, those of belief functions—will have a worse score, whatever be the truth or falsity of the A's and H's, than the probabilist. Notice how eminently practical this approach is. The "expertise" of an expert could be assessed by keeping a check on his scores. Of two probabilists, either one may do better than the other, but both will do better than someone not using the probability calculus.

5. AXIOMATIC APPROACH

In an alternative approach we think about the concept of uncertainty and try to latch onto simple, basic principles that ought to be present in any study of uncertainty; such that any violation of a principle would, when exposed, make the argument look ridiculous. The principles, self-evident truths, are called axioms and from these we would hope to deduce, by mathematical reasoning, the rules that the numbers

obey. Euclidean geometry is the famous example of this procedure when applied to the measurement of length. This program was first carried out for beliefs in 1926 by Ramsey (1931). The best of the known examples is Savage (1954). DeGroot (1970) presents what is perhaps the most readable version. All of these approaches lead to the result that the numbers must obey exactly the three rules of probability above. In other words, the "axioms" of probability have been deduced from other, simpler ideas that more legitimately can, because of their self-evidentiary nature, be called axioms.

Let the converse be emphasized: any violation of the rules must correspond to some violation of the basic axioms, of those rules whose violation would look ridiculous. We really have no choice about the rules governing our measurement of uncertainty: they are dictated to us by the inexorable laws of logic. Of course, they are entirely dependent on the chosen axioms and the history of mathematics warns us not to be too complacent about the "sacred" rightness of axioms. But at the moment, the axioms are unassailed and all variants produce minor variants in probability.

6. COHERENCE

At this point we should perhaps digress to discuss an important aspect of the Ramsey/Savage/de Finetti approaches that is often over-looked. The discussion will also help to explain why nonprobabilistic views have had some success in AI or ES even though the ideas are unsound. The rules of probability show how different uncertainty statements have to fit together. Thus, the multiplication rule above refers to three assessments and says that one of them must be the product of the other two. Instead of "fitting together" we talk of coherence. The results just described can be stated as showing that coherence can only be achieved by means of probability. We may say belief functions are incoherent (they do not obey the addition rule).

Coherence is not peculiar to the measurement of belief. It applies to all measurement: for example, of length. If ABC is a triangle with a right angle at B, it makes perfectly good sense to say $AB = 2$ or $AC = 4$ or $BC = 3$, or even to make two of these statements together. But make all three together and you are incoherent, for Pythagoras demands that $AC^2 = AB^2 + BC^2$, which is not true of the numbers given. Similarly one can say that $p(A_1 \mid H) = \frac{1}{2}$ or $p(A_2 \mid A_1 \cap H) = \frac{2}{3}$ or $p(A_1 \cap A_2 \mid H) = \frac{1}{4}$, but one cannot make all three statements simultaneously. The multiplication law replaces Pythagoras. It is curious that coherence is strictly adhered to with lengths but often ignored with beliefs, reflecting the immaturity of belief measurement.

And that explains why nonprobabilistic procedures can sometimes appear sensible. The adherents never make enough statements for coherence to be tested. They only tell us the equivalent of $AB = 2$ and $AC = 4$, never discussing BC, for to do so would reveal the unsound nature of the argument.

7. BAYES THEOREM

One example of coherence is so important in AI and ES that we should perhaps consider it now. Interchanging A_1 and A_2 in the above statement of the multiplication law and recognizing that $A_1 \cap A_2 = A_2 \cap A_1$, we immediately have that

$$p(A_1 \mid H)p(A_2 \mid A_1 \cap H) = p(A_2 \mid H)p(A_1 \mid A_2 \cap H).$$

By using the equivalent result but with $\bar{A}_2$, replacing A_2, we have

$$\frac{p(A_2 \mid A_1 \cap H)}{p(\bar{A}_2 \mid A_1 \cap H)} = \frac{p(A_1 \mid A_2 \cap H)}{p(A_1 \mid \bar{A}_2 \cap H)} \frac{p(A_2 \mid H)}{p(\bar{A}_2 \mid H)}.$$

This is Bayes theorem in odds form. (The odds (on) A are simply the ratio t of $p(A)$ to $p(\bar{A})$: the odds against are the inverse of this. In practice they are usually quoted as t to 1 on or t to 1 against with $t \geq 1$.) To appreciate what it says, temporarily omit H from the notation and language, recognizing that it is present in every conditioning event in the statement of the theorem. Then the result is that the odds, $p(A_2)/p(\bar{A}_2)$, of A_2 are changed, due to the additional knowledge of A_1, into $p(A_2 \mid A_1)/p(\bar{A}_2 \mid A_1)$ by multiplying by $p(A_1 \mid A_2)/p(A_1 \mid \bar{A}_2)$. The multiplier is called the likelihood ratio. It is the ratio of the probabilities of the additional knowledge A_1, given A_2 and then given $\bar{A}_2$. Thus an AI system faced with uncertainty about A_2 and experiencing A_1 has to update its uncertainty by considering how probable what it has experienced is, both on the supposition that A_2 is true, and that A_2 is false. Any other procedure is incoherent. Most intelligent behavior is simply obeying Bayes theorem. A high level of intelligence consists in recognizing a new pattern. This is not allowed for in Bayes theorem, nor in any other paradigm known to me. The simple AI systems that we have at the moment must be Bayesian.

8. A CHALLENGE

Let us summarize where we have got to in the argument. On the basis of simple, intuitive rules (or using a technique of scoring statements of uncertainty), it follows that probability is the only way of handling uncertainty. In particular other ways are unsound and essentially ad hoc in that they lack an axiomatic basis.

There is however more than just the inevitability of probability. There is the consideration that probability is totally adequate for all uncertain situations encountered so far. This is often denied. The following statements are taken from Zadeh (1983):

> "A serious shortcoming of [probability-based] methods is that they are not capable of coming to grips with the pervasive fuzziness of information in the knowledge base, and, as a result, are mostly ad hoc in nature."
>
> "The validity of [Bayes rule] is open to question since most of the information in the knowledge base of a typical expert system consists of a collection of fuzzy rather than nonfuzzy propositions."

Shafer (1982) says, in comparing belief functions and Bayesian methods, "The theory of belief functions offers an approach that better respects the realities and limitations of our knowledge and evidence."

I offer a challenge to these writers and to all who espouse nonprobabilistic methods for the study of uncertainty. The challenge is that anything that can be done by these methods can be better done with probability. I think this is a fair challenge. It is a requirement that the method has been used and is not just a topic for theorizing, which rules out some speculations in the alternative paradigms. If the challenge fails then we shall really have advanced: for an inadequacy in probability will have been exposed and the need for an alternative justified. The challenge is in the spirit of Popper who partly judges the merit of a theory on its capability of being destroyed; for the rich calculus of probability leads to many testable conclusions. It is also relevant to Popperian ideas because he has discussed certain inadequacies in probability. These have been disposed of by Jeffreys (1961).

As these words are being written it is impossible to know what challenges might arise. All that can be done is to take material already in the literature and examine that. I begin with fuzzy ideas.

9. PROBABILITY IN PLACE OF FUZZINESS

As an example of a fuzzy proposition Zadeh (1983) cites "Berkeley's population is over 100,000." He says it is fuzzy because "of an implicit understanding that *over 100,000* means *over 100,000 but not much over 100,000*" (his italics). (He might also have added that Berkeley is fuzzy. Does it refer to the town in Gloucestershire or that in California? And population: does it merely refer to permanent residents or are students included? These are not jibes: my point is that nearly all statements are imprecise.)

The probabilistic approach would be to give a probabilistic statement about a quantity *that can be evaluated*. The qualification is important, de Finetti has emphasized. As far as possible all probabilities should refer to propositions or events that can realistically be tested for truth or falsity. This is because we want to *use* them. It may be necessary to introduce other propositions but only as aids to the calculation of testable ones. (In statistics parameters are used for this purpose. An example in Section 14 will use guilt of a suspect.) A possible quantity to discuss in the fuzzy statement is the answer the relevant city official in Berkeley would give when asked for the population of Berkeley. If this is denoted X, then the probabilistic statement corresponding to that quoted is $p(X \mid H)$, where H is the knowledge possessed by the maker of the statement. It would have a mode a little over 100,000 if the statement is in H.

It is important to notice that in applications it may not be necessary to specify the full probability distribution $p(X \mid H)$. For example, it may be enough to quote its mean, the expectation of X given H; what de Finetti calls the *prevision* of X given H. More sophistication may require the variance of X, or equivalently, the prevision of X^2 given H. Fractiles of X are another possibility.

All fuzzy propositions of this type can be interpreted probabilistically in a manner similar to our treatment of Berkeley. "Henry is young" needs a little care. It clearly refers to Henry (whom I take to be a well defined person) and an uncertain quantity X, his age. But the description is very vague. Made on campus, Henry might be only 19; made at a faculty dinner Henry might be 30; made in a home for senior citizens, he might be 65. Consequently, H is very relevant to this result. Without context $p(X \mid H)$ will need to be appreciable even for $X = 65$.

10. NUMERICAL EXPRESSION OF FUZZINESS

Another example is both more serious and more elaborate. "John has duodenal ulcer (CF = 0.3)" (CF is an abbreviation for certainty factor). It is a well known feature of medical studies that many concepts are imprecisely defined and that a difficulty in using medical records resides in the varied use different doctors make of the same term. Nevertheless doctors find it useful to identify features like "duodenal ulcer." The situation can be described probabilistically by introducing Δ, an ill-defined but supposedly real ailment, duodenal ulcer, and also D_i the appreciation of duodenal ulcer by doctor i. The fuzziness of the concept can be captured by considering $p(D_i \mid \Delta)$ and $p(D_i \mid \bar{\Delta})$, the probability that doctor i will say John has duodenal ulcer both when John has, and does not have, true duodenal ulcer. (Useful comparison can be made with Bayes theorem above: Δ replaces A_2, D_i replaces A_1, and H is omitted from the present notation.) Notice that Δ may not be a testable quantity. It is introduced as a parameter to facilitate the calculation of quantities that are testable. For example, if the

above statement is made by a first doctor, what is the probability that a second will agree? $p(D_2 \mid D_1)$ can be evaluated by extending the conversation to include Δ. For example, the D_i might be independent, given Δ.

This second fuzzy statement introduces a numerical measure in the form of a certainty factor, here 0.3. This contrasts with the apparently similar numerical assertion that the probability (on an undefined H) that John has a duodenal ulcer is 0.3 in at least two ways. First, certainty factors combine by rules that are different from those of the probability calculus, so that they would inevitability produce worse scores in an adequate test than would probabilities. Furthermore, these rules have no axiomatic basis and are merely inventions of fertile, unconstrained minds. The second difference between certainty factors and probabilities is that the operational meaning of the latter is clear whereas that of the former is not. We may say that probabilities have standards, possibilities do not. One standard for probability was mentioned above: balls in an urn. But expectation of benefit or a uniform distribution may replace these. All measurement requires a standard and certainty factors are dubious because they do not have them. What does CF = 0.3 mean?

The literature on fuzzy logic is vast, complicated, and somewhat obscure. I have surely missed some examples that it would be useful to test against the challenge which remains: anything fuzzy logic can do, probability can do better.

11. INCOHERENCE AND BELIEF FUNCTIONS

We next turn from fuzzy logic to belief functions. I have already considered a good example of Shafer's (1982) in the discussion to that paper. It is repeated here partly because to do so is simpler for me than to take another one; and also because it is then possible to respond to Shafer's reaction to my probabilistic argument. Before giving this it might be useful to exhibit incoherence in the use of belief functions. (The argument also applies to fuzzy methods.)

We follow Shafer and write Bel(A) for the belief in A, omitting reference to the conditioning event. Now it is possible that

$$\text{Bel}(A) + \text{Bel}(\bar{A}) < 1$$

(similarly for certainty factors). Write Bel(A) = a, Bel($\bar{A}$) = b so that $a + b < 1$. (Necessarily $a, b \geq 0$.) Let us score such a belief using the quadratic rule. The possible scores are:

$$A \quad \text{true} \quad (a-1)^2 + b^2,$$

$$\bar{A} \quad \text{true} \quad a^2 + (b-1)^2.$$

Now replace a by a', b by b' where $a' = a + \varepsilon$, $b' = b + \varepsilon$, and $\varepsilon = \frac{1}{2}(1 - a - b)$. It easily follows that $a' + b' = 1$ and that both

$$(a'-1)^2 + b'^2 < (a-1)^2 + b^2$$

and

$$a'^2 + (b'-1)^2 < a^2 + (b-1)^2.$$

Consequently it is certain (irrespective of whether A or $\bar{A}$ is true) that beliefs a and b will score worse than probabilities a' and b', adding to one. The result generalizes with any score.

12. PROBABILITY IN PLACE OF BELIEF FUNCTIONS

Now for Shafer's example. Imagine a disorder called "ploxoma," which comprises two distinct "diseases": θ_1 = "virulent ploxoma," which is invariably fatal, and θ_2 = "ordinary ploxoma," which varies in severity and can be treated. Virulent ploxoma can be identified unequivocally at the time of a victim's death, but the only way to distinguish between the two diseases in their early stages seems to be a blood test with three possible outcomes, labeled x_1, x_2, and x_3. The following evidence is available: (i) Blood tests of a large number of patients dying of virulent ploxoma showed the outcomes x_1, x_2, and x_3 occurring 20, 20, and 60% of the time, respectively. (ii) A study of patients whose ploxoma had continued so long as to be almost certainly ordinary ploxoma showed outcome x_1 to occur 85% of the time and outcomes x_2 and x_3 to occur 15% of the time. (The study was made before methods for distinguishing between x_2 and x_3 were perfected.) There is some question whether the patients in the study represent a fair sample of the population of ordinary ploxoma victims, but experts feel fairly confident (say 75%) that the criteria by which patients were selected for the study should not affect the distribution of test outcomes. (iii) It seems that most people who seek medical help for ploxoma are suffering from ordinary ploxoma. There have been no careful statistical studies, but physicians are convinced that only 5–15% of ploxoma patients suffer from virulent ploxoma.

My reply was as follows. The first piece of evidence (i) establishes in the usual way that the chances for a person with virulent ploxoma to have blood test results of types x_1, x_2, and x_3 are 0.2, 0.2, and 0.6. The second (ii) is subtler for two reasons: x_2 and x_3 are not distinguished in the data, and the patients in the study are not judged exchangeable with other patients so that the chances β in the study and γ for the new patients are not necessarily equal. The first presents no difficulty since the likelihood for the data is $\beta_1^r(\beta_2 + \beta_3)^{n-r}$, where $r = 0.85n$ and n is the number of patients in the study. The distribution of β given the data can therefore be found. Let $p(\gamma \mid \beta)$ be the

conditional distribution of γ, given β. This concept replaces the single figure of 75% quoted by Shafer and which yields a discount rate of $\alpha = 0.25$. It would be possible to suppose $\gamma = \beta$ with probability 0.75 and is otherwise uniform in the unit interval in imitation of belief functions: but this may be an unrealistic description of the situation. The third piece of evidence (iii) says the distribution of the chance θ that a patient has virulent ploxoma, $p(\theta)$, is essentially confined to the range (0.05 to 0.15). We are now ready to perform the requisite probability calculations.

Let G be the event that a new patient, George, has virulent ploxoma and let g_i be the result of his blood test. We require $p(G \mid g_i, E)$ where E is the evidence. From (iii) $p(G) = \int \theta p(\theta)\, d\theta$. From (i) $p(g_i \mid G, E) = 0.2$ for $i = 1, 2$ and 0.6 for $i = 3$. From (ii)

$$p(g_i \mid \bar{G}, E) = \int\int \gamma_i p(\gamma \mid \beta) p(\beta \mid E)\, d\beta\, d\gamma$$

$$= \int E(\gamma_i \mid \beta) p(\beta \mid E)\, d\beta$$

and the calculations can be completed in the usual way using Bayes' theorem. If $E(\theta) = 0.10$, $E(\gamma_i \mid \beta) = \beta_i$, and $E(\beta_2 \mid \beta_1) = \frac{1}{2}(1 - \beta_1)$ then the probabilities of G given g_i are, respectively, 0.025, 0.229, and 0.471.

It may be objected that this analysis virtually ignores the uncertainty about the study and about θ. It does so because they are irrelevant. The interested reader may like to consider the case of George and Henry and their blood tests. Then the uncertainties will matter: for example, $E(\gamma_i^2 \mid \beta)$, involving the conditional variance of γ_i, will arise.

Shafer in response says that "Lindley insists that the uncertainties affecting this study are irrelevant and should be ignored. Is this reasonable? Suppose that instead of having only 75% confidence in the study we have much less confidence. Is there not some point where even Lindley would chuck out the study and revert to the prior 5–15%?" My reply is that Shafer is correct and that the uncertainty does matter a little, for it affects $E(\gamma \mid \beta)$. Were we to have no confidence at all in the study then $E(\gamma \mid \beta)$ would not depend on β, and $p(g_i \mid \bar{G}, E)$ would be simply $E(\gamma_i)$ about which no information is given. (The prior on θ seems irrelevant.)

Consequently I feel that the challenge has been well met with the example and, by a Popperian argument, the credibility of probability theory is increased.

13. COMPLEXITY, COVERAGE, DECISIONS AND RICHNESS

Here are four miscellaneous remarks.

1. It should be noted that fuzzy logic and belief functions are considerably more complicated concepts than those of probability. With belief functions we start effectively with probabilities over the power set of the original events, itself much more complicated than the original set, and then have to elaborate on that. Dempster's rule of combination is vastly more involved than Bayes and then only applies in certain cases. Fuzzy logic leads to nonlinear programming and contains great complexities of language and ideas. Yet probability is extremely simple, using only three rules and containing rich concepts like independence and expectation.

Certainly if my challenge fails it will be necessary to introduce some change into probability ideas, which will almost surely increase the complexity, yet be necessary and rewarding. But until that happens is it not best to accept the advice of William of Ockham and not multiply entities beyond necessity?

2. It is not implied in the challenge that probability can handle every problem involving uncertainty: the claim is merely that probability can do better than the alternatives. I believe that it has the potentiality to solve every uncertain situation but there are some for which the available techniques are inadequate. It is absurd to think that any paradigm can quickly resolve every relevant puzzle: some may resist solution for decades. For example, the medical problem of handling large numbers of indicants in diagnosis is currently unresolved because we do not have adequate techniques for handling the complicated dependencies that exist. (And certainly belief functions do not.) We need more research into applied probability and less into fancy alternatives.

3. Why do we want to study uncertainty? Aside from the intellectual pleasure it can provide, there is only one answer: to be able to make decisions in the face of uncertainty. Studies that do not have the potentiality for practical use in decision making are seriously inadequate. An axiomatic treatment of decision making shows (Savage, 1954; DeGroot, 1970) that maximization of expected utility is the only satisfactory procedure. This uses, in the expectation calculation, the probabilities and these, and only these, are the quantities needed for coherent decision making by a single decision maker. Only the utilities, dependent on the consequences, not on the uncertainties, need to be added to make a rational choice of action. How can one use fuzzy logic or belief functions to decide? Indeed, consider a case where $\text{Bel}(A) + \text{Bel}(\bar{A}) < 1$. Because you have so little belief in either outcome do you, like Buradin's ass, starve to death in your indecision between A and its negation? Reality demands probability.

4. It is sometimes said, as in the quotes from Zadeh above, that probability is inadequate. This sense of inadequacy sometimes arises because people only think of probability as a value between 0 and 1,

forgetting the whole concept of coherence and, in particular, ignoring the addition and multiplication laws. In fact probability is a rich and subtle concept capable of dealing with beautifully delicate and important problems. This richness is hard to convey without deep immersion in the topic. In order to display this, and also to try to avoid the impression that this paper is entirely concerned with bashing other ideas, I conclude by discussing a situation that arises in forensic science or criminalistics. It has been much discussed in the literature; a convenient reference is Eggleston (1983). An almost identical problem has been considered by Diaconis and Zabell (1982) using Jeffrey's rule. For reasons given below, I think their treatment is unsatisfactory.

14. A PROBABILITY EXAMPLE

A crime has been committed by a person who is to be found among a population of $(n + 1)$ people. One of these is referred to as the suspect, the others are labeled in a noninformative way from 1 to n. Let G_s be the event that the suspect is guilty, G_i that person i is $(1 \leq i \leq n)$. Initially, $p(G_s) = \pi$, $p(G_i) = (1 - \pi)/n$ for all i. (Some forms of the problem have $\pi = (n + 1)^{-1}$, which probabilistically does not distinguish the suspect from the other n.)

An investigator studying the crime says "the evidence suggests the criminal is left-handed." This is a fuzzy statement and its probabilistic interpretation requires care. After discussion the investigator says that the probability that the criminal is left-handed is P. This is still ambiguous. Diaconis and Zabell appear to interpret it to mean that the probability that the criminal will be found among the left-handers in the group of $(n + 1)$ is P. I think a British forensic scientist would mean that if he had the criminal in front of him, the probability that he would be found to be left-handed is P. The former is the chance of guilt among left-handers; the latter of left-handedness among the guilty. Also the former requires reference to the population, and the latter does not. Typical forensic evidence makes no mention of a population, only of the criminal, and so the latter interpretation is appropriate. There is a confusion between $p(A \mid B)$ and $p(B \mid A)$.

Working with the forensic interpretation, the formal statement is $p(l_i \mid G_i) = P$, where l_i denotes the event that person i is left-handed $(1 \leq i \leq n$ and $i = S)$. It was emphasized in the discussion of Bayes theorem that it is essential to consider the evidence A_1 both on A_2 and on $\bar{A}_2$. So here we need, in addition to $p(l_i \mid G_i)$, $p(l_i \mid \bar{G}_i)$. The latter is the chance that anyone is left-handed and may ordinarily be equated to the frequency of left-handedness in the population, p say. So $p(l_i \mid \bar{G}_i) = p$ for all i, including S. Presumably $P > p$. (In some forms of the problem $P = 1$ and the forensic evidence is firm. This can realistically arise when dealing with blood types that can be identified without error.) Diaconis and Zabell do not consider p. This seems strange because the presence of an unusual trait intuitively carries more weight than a common one. The formal analysis below will confirm this.

15. THE ROLE OF ADDITIONAL EVIDENCE

Now consider various forms of additional evidence.

Evidence E_1. The suspect is found to be left-handed. In the notation this is the event l_s. Simple use of Bayes theorem

$$p(G_s \mid l_s) = p(l_s \mid G_s)p(G_s)/p(l_s)$$

yields

$$p(G_s \mid l_s) = P\pi/\{P\pi + p(1 - \pi)\} \tag{1}$$

which clearly exceeds π. E_1 is indicative of the suspect's guilt.

Evidence E_2. Person no. 1 is left-handed. This is l_1. Now with both E_1 and E_2

$$p(G_s \mid l_s l_1) \propto p(l_s l_1 \mid G)p(G) = Pp\pi.$$

Similarly,

$$p(G_1 \mid l_s l_1) \propto Pp(1 - \pi)/n$$

and

$$p(G_i \mid l_s l_1) \propto p^2(1 - \pi)/n \quad \text{for} \quad 2 \leq i \leq n.$$

Thus,

$$p(G_s \mid l_s l_1) = P\pi/\{P\pi + P(1 - \pi)/n + p(1 - \pi)(n - 1)/n\}. \tag{2}$$

Rearranging the denominator as $P\pi + p(1 - \pi) + (P - p)(1 - \pi)/n$ we see that (2) is less than (1): the knowledge of another left-handed person in the population has slightly decreased the probability that S is guilty. Notice that when $n = 1$, $p(G_s \mid l_s l_1) = \pi$: the evidence that *all* the population is left-handed has not changed the suspect's probability for guilt at all.

Evidence E_3. There are no left-handers among the n people.

Combined with E_1 this means that the suspect is the only left-hander. Denoting E_3 by l_0, a use of Bayes theorem similar to that employed with E_1 and E_2 gives

$$p(G_s \mid l_s l_0) \propto p(l_s l_0 \mid G_s)p(G_s) = P(1 - p)^n\pi$$

and

$$p(G_i \mid l_s l_0) \propto p(l_s l_0 \mid G_i)p(G_i) = p(1 - p)^{n-1}(1 - P)(1 - \pi)/n.$$

Hence,

$$p(G_s \mid l_s l_0) = P\pi/\{P\pi + p(1-\pi)(1-P)/(1-p)\}. \tag{3}$$

This clearly exceeds $p(G_s \mid l_s)$, equation (1), if $P > p$, showing that E_3 increases the probability that the suspect is guilty. Indeed, if $P = 1$, (3) gives 1 as it should.

Evidence E_4. There is at least one left-hander among the n people.

E_4 is the negation of E_3 and may be written $\bar{l}_0$. It differs from E_2 in that the latter names a specific left-hander, person no. 1. We have

$$p(l_s \bar{l}_0 \mid G_s) = p(l_s \mid G_s) - p(l_s l_0 \mid G_s) = P - P(1-p)^n$$

and

$$p(l_s \bar{l}_0 \mid G_i) = p(l_s \mid G_i) - p(l_s l_0 \mid G_i) = p - p(1-p)^{n-1}(1-P).$$

A further use of Bayes theorem gives

$$p(G_s \mid l_s \bar{l}_0) = [P\pi - P(1-p)^n\pi]/C \tag{4}$$

where

$$C = P\pi + p(1-\pi) - (1-p)^n\{P\pi + p(1-\pi)(1-P)/(1-p)\}.$$

If $n = 1$ this gives π in agreement with $p(G_s \mid l_s l_1)$, equation (2). It is easy to see that $p(G_s \mid l_s \bar{l}_0) < p(G_s \mid l_s)$, equation (1), so that E_4 slightly decreases the probability of the suspect's guilt.

Now for a subtlety: compare (2) and (4), that is the probability that the suspect is guilty given, in (2), the name of a left-hander and in (4) the mere presence of a left-hander. These are different. It is not too hard to verify by induction on n that

$$p(G_s \mid l_s l_1) < p(G_s \mid l_s \bar{l}_0)$$

for $n > 1$, so that the definitive knowledge of the left-handedness of person no. 1 reduces the suspect's guilt probability by more than does the mere evidence of someone's left-handedness.

I leave the reader to think out whether the following argument is correct. Knowing there is a left-hander in the n (E_4), no information about the suspect's guilt can possibly be provided by telling me the number of one of them. Accepting this, you are told it is person no. 1. Since (2) and (4) differ (and calling person no. 1 Smith for dramatic effect) the evidence "Smith is left-handed" and "There are left-handers, one of whom is called Smith" have different evidential value.

16. CONCLUSION

Our argument may be summarized by saying that probability is the only sensible description of uncertainty and is adequate for all problems involving uncertainty. All other methods are inadequate. The justification for the position rests on the formal, axiomatic argument that leads to the inevitability of probability as a theorem and also on the pragmatic verification that probability does work. My challenge that anything that can be done with fuzzy logic, belief functions, upper and lower probabilities, or any other alternative to probability, can better be done with probability, remains.

ACKNOWLEDGMENTS

This research was supported by Grant DAAG29-84-K0160 from the United States Army Research Office and Contract N00014-77-C-0263 (Project NR042-372) from the Office of Naval Research with The George Washington University.

REFERENCES

DE FINETTI, B. (1974/5). *Theory of Probability* 1, 2. Wiley, New York.

DEGROOT, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

DIACONIS, P. and ZABELL, S. L. (1982). Updating subjective probability. *J. Amer. Statist. Assoc.* 77 822–830.

EGGLESTON, R. (1983). *Evidence, Proof and Probability*. Weidenfeld and Nicolson, London.

FINE, T. L. (1973). *Theories of Probability: An Examination of Foundations*. Academic, New York.

JEFFREY, R. (1965). *The Logic of Decision*. McGraw-Hill, New York.

JEFFREYS, H. (1961). *Theory of Probability*. Clarendon Press, Oxford.

LINDLEY, D. V. (1982). Scoring rules and the inevitability of probability (with discussion). *Internat. Statist. Rev.* 50 1–26.

RAMSEY, F. P. (1931). Truth and probability. In *The Foundations of Mathematics and Other Logical Essays* 156–198. Kegan Paul, Trench, Trubner, London.

SAVAGE, L. J. (1954). *The Foundations of Statistics*. Wiley, New York.

SHAFER, G. (1976). *A Mathematical Theory of Evidence*. Princeton Univ. Press, Princeton, N. J.

SHAFER, G. (1982). Belief functions and parametric models (with discussion). *J. Roy. Statist. Soc. Ser. B* 44 322–352.

SMITH, C. A. B. (1961). Consistency in statistical inference and decision (with discussion). *J. Roy. Statist. Soc. Ser. B* 23 1–37.

ZADEH, L. A. (1983). The role of fuzzy logic in the management of uncertainty in expert systems. *Fuzzy Sets and Systems* 11 199–227.

Statistical Science
1987, Vol. 2, No. 1, 3-44

Probabilistic Expert Systems in Medicine: Practical Issues in Handling Uncertainty

David J. Spiegelhalter

Abstract. The development of expert systems in medicine has generally been accompanied by a rejection of formal probabilistic methods for handling uncertainty. We argue that a coherent probabilistic approach can, if carefully applied, meet many of the practical demands being made, and briefly illustrate our claim with three current projects.

Key words and phrases: Evidence propagation, knowledge representation, graphical models, imprecise probabilities.

1. INTRODUCTION

The first problem in discussing "uncertainty in expert systems" comes in defining our terms. We shall view "expert systems" as being programs intended to provide judgments or advice to users in a reasonably convincing manner, in which *knowledge*, whether represented as rules, networks, or frames, is generally characterized by local relationships between propositions of varying generality; *data* is obtained on a new case, upon which the "knowledge" is brought to bear by means of a controlling mechanism. Characteristics which are often said to distinguish such programs from standard statistical or mathematical models include the frequent use of subjective judgments for both the qualitative structure and any accompanying quantification, the emphasis on explanation, and the incorporation of both knowledge and data that is fragmentary. Systems are also often intended to enable "learning," in which knowledge is adjusted in the light of data on past cases.

Within this context the term "uncertainty" is used in a very wide sense and this has led to considerable argument about the role appropriate to formal probabilistic reasoning (Cheeseman, 1985; Spiegelhalter, 1986a, 1986b; and papers in Kanal and Lemmer, 1986). Some misunderstanding may have arisen from the common use of expressions of the form

IF conditions X hold, THEN Y with certainty P.

If Y is a random event which is currently unknown, the statistical view is that P represents a kind of "predictive" uncertainty expressed as a probability (see Lindley, page 18). However, in many expert system applications, precisely the same representation is used when Y signifies some *action* or choice, and P essentially corresponds to "procedural" uncertainty, in that doubt is being expressed about the reasonableness of, or the support for, performing an act or making an assumption. Thus, Cohen (1985, page 52) states that "one's certainty in a result should depend on what the result is wanted for," and Van Melle et al. (1981, page 5) say that "certainty factors" in EMYCIN combine subjective probabilities and utilities to measure "importance."

Thus, there is clearly great potential for confusion between the fairly restricted, statistical sense of uncertainty as applied to *facts* and the use of the term in a broader, linguistic sense in describing uncertainty about *acts*. To try to avoid this confusion in this short paper, we shall explicitly restrict attention to uncertainty concerning potentially verifiable, but currently unknown, events.

We shall concentrate on practical, rather than philosophical, issues concerning the way uncertainty is handled in existing programs. We shall not consider in detail either the representation of knowledge or the control of the program. Published examples motivate the search for a methodology that satisfies a number of demands, and three current projects will then be used to illustrate some specific aspects of the attempt to use probabilistic methods in as effective a way as possible. Finally, an attempt is made to bring the argument together into a prospect for future developments.

2. DEMANDS MADE OF A CALCULUS

The particular complexity of many medical problems has challenged the notion of a rigorous unified treatment of uncertainty and, in general, ad hoc quantifications have been used to measure evidence

David J. Spiegelhalter is a statistician at the Medical Research Council, MRC Biostatistics Unit, 5 Shaftesbury Road, Cambridge CB2 2BW, England.

for various possible underlying hypotheses (Szolovits and Pauker, 1978). The complex interrelationships between disease processes and manifestations have led to various systems for propagating degrees of certainty arising from fragmentary data and combining evidence from different sources. PIP (Pauker et al., 1976) and INTERNIST/CADUCEUS (Miller, Pople, and Myers, 1982) both essentially score hypotheses using evidence from current symptoms that support a hypothesis, which is discounted by a score expressing absent symptoms that would be expected, and a score expressing present symptoms that would *not* be expected. MYCIN/EMYCIN use a more modular structure in which certainty factors are attached to propositions, although CASNET/EXPERT (Kulikowski and Weiss, 1982) propagates weights through a causal network. A statistical system such as that of de Dombal et al. (1972) begins with knowledge derived from a data base, but the simplistic independence assumptions made in combining evidence (although effective in discrimination) ensure that the certainty propagated is not expected to be interpretable as a probability—the same holds for the Bayesian updating technique in PROSPECTOR (Duda, Hart, and Nilsson, 1976). Fuzzy reasoning (Adlassnig, 1980; Fieschi et al., 1983) has also been used as a means of capturing the ill-defined nature of many clinical terms.

We can identify a number of considerations that have led to the procedures that have been adopted and that are currently being researched. The strongest has been the claim that a single probability of a hypothesis, even if it were based on extensive data, is not sufficient to convince a clinician: the *evidence* on which to base a conclusion must be retrievable, to enable conflicts and doubtful contributions to be identified. A particular case of this demand for justification is the situation where little relevant data is available and there is essentially *ignorance* concerning the possibility of a hypothesis. This arises particularly in medicine due to the hierarchical, taxonomic structure of disease descriptions in which evidence may be available which supports a general disease category but gives no indication of the relative plausibility of the subcategories of disease. Thus, the hierarchical hypothesis structure is viewed as a natural justification for *ranges* of uncertainty, for which a number of schemes exist (see, for example, Quinlan, 1983), although as we shall emphasize later it is not generally made clear whether such ranges are due to inadequate knowledge or inadequate data. The demand that individual contributions of pieces of evidence should be identified, and that evidence should be able to focus on groups of diseases without distinguishing within that group, has led naturally to the study of the possible role of belief functions in medicine (Gordon and Shortliffe, 1984). Much attention is now being paid to solving the accompanying computational problems and making some allowance for dependencies between sources of evidence. The concept of discounting in belief functions could also be seen as a means of allowing for doubt about the precise numbers to be placed on evidential statements.

To summarize: current interest is focussed on schemes that can propagate measures of uncertainty through complex relationships often defined on a hierarchical structure, that can identify conflicting evidence and lack of evidence, and can cope with incoming data that do not follow a predefined order. The reasoning process should be justifiable and fairly intuitive, and allowance for imprecise specification of numerical relationships would be an advantage.

Although the above desiderata appear admirable, we feel there is an important item that has been largely ignored in practice. This concerns the *operational meaning* of the quantities which express uncertainty and which allows a reasonable basis for both assessment of inputs and criticism of outputs of a system. In the following examples we describe attempts to retain meaning while responding to demands and constraints made by the real practical problems of interest. Refer to Pearl (1986a, b) for further discussion on how probabilistic reasoning can be adapted to meet the demands of expert systems.

3. EXAMPLES OF PROBABILISTIC ANALYSIS

GLADYS—the GLAsgow DYSpepsia System

GLADYS is a program designed to interview patients presenting to a clinic with dyspepsia, and provide a reasoned probabilistic diagnosis based on the symptoms alone. It was developed at the Diagnostic Methodology Research Unit at Glasgow, and runs on a microcomputer with a special keyboard to record patient responses. The control of the interview is strictly algorithmic, in that branches to more detailed interrogation are taken depending on the results to trigger questions. The interview has been found to be accurate and acceptable (Lucas et al., 1976). The responses are analyzed according to a scoring system derived from a modified logistic regression technique, described in detail in Spiegelhalter and Knill-Jones (1984), of which certain aspects are relevant to the issues raised in the previous section.

Firstly, there is a real need to deal with *hierarchical* disease structures, in which for example, certain features may discriminate the generic class peptic ulcer (PU) from other diseases, although other items of information are relevant to discriminating duodenal from gastric ulcer (GU) within the peptic ulcer class. This is accomplished by calculating probabilities

conditional on the branch in the hierarchy and then multiplying downward to obtain the overall probability: for example, we calculate $p(\mathrm{GU} \mid \mathrm{PU})$ and $p(\mathrm{PU})$ from which $p(\mathrm{GU}) = p(\mathrm{GU} \mid \mathrm{PU})p(\mathrm{PU})$.

Secondly, the scoring system allows *explanation* of the final probability in terms of the contributing pieces of evidence. For example, a patient described in Spiegelhalter and Knill-Jones (1984) provided the following evidence relevant to a diagnosis of gallstones:

Evidence FOR gallstones		Evidence AGAINST gallstones	
History less than 6 months	77	Pain not severe enough to warrant emergency call to doctor	−43
Pain comes in "attacks"	177		
Can enumerate attacks	63	Pain does not radiate	−38
Attacks produce restlessness	31		
Pain in right hypochondrium	77		
Total	425		−81
Balance of evidence	+344	(Total evidence = 425 + 81 = 506; conflict ratio = 506/344 = 1.5)	
Initial score	−300	(Corresponding to prevalence of 4.7%)	
Final score	44 = 61% chance of gallstones		

Some explanation of the above "explanation" is necessary. The scores given to findings are 100 $\log_e$(likelihood ratios) adjusted, roughly speaking, for correlations between items of information. Thus, the initial score of $S = -300$ is transformed to a prior probability $p = 1/\{1 + \exp(-S/100)\} = .047$, which is simply the inverse of $S = 100 \log_e\{p/(1 - p)\}$. The "conflict ratio" (= total evidence/ | balance of evidence |) is a rough measure of how much the total evidence obtained contradicts itself: a high ratio, say above around 2.5, suggests the clinician should check some of the important questions. The initial score is based on a prevalence in an urban clinic and could be altered depending on circumstances. The scores come from analysis of a data base of 1200 cases and the statistical modelling means the final probabilities are reasonably well calibrated, in that of patients presenting as above, around 60% should turn out to have gallstones as a major cause of their symptoms. This is a very popular characteristic of the system. There is, however, no reason why the scores should not be subjectively assessed provided one monitors whether the predictions have similar properties of calibration.

Thirdly, *imprecision* of the quantification could be incorporated by placing standard errors on the predictions. The above example has a standard error of 42 on the final score corresponding to a rough 95% interval of (.40, .78) on the predictive probability. Finally, *ignorance* may be viewed retrospectively in terms of the total evidence received either for or against a proposition. However, as suggested in Spiegelhalter and Knill-Jones (1984), we may also quantify prospective ignorance in terms of the results that may occur when the data of which we are currently ignorant becomes available. This concept translates into calculating the predictive distribution of the possible final probabilities that may be ascribed to a disease. Tukey (1984) recommended that such distributions should be included as part of the explanation facilities. Thus before an interview, a patient has a fairly *precise* probability of gallstones (95% interval .03, .07), but one based on an ignorance reflected in the wide distribution of feasible probabilities that could be taken on when data become available; whereas at the end of the interview, there is a relatively *imprecise* probability with a 95% interval of (.40, .78), but no remaining ignorance within the bounded context of the system.

We would not normally consider GLADYS as an expert system since it does not use knowledge representation techniques derived from AI, it is not based on expert opinion and it does not operate interactively. However, many of our aims match those of classic expert systems, except that we are determined to remain, as far as possible, within a probabilistic framework.

A Diagnostic System for Chest Diseases

A group at the Chest Clinic at Westminster Hospital are developing a system for probabilistic diagnosis of patients presenting with a normal chest x-ray. The system uses simple independent Bayes updating assuming mutually exclusive disease categories, and our only concern here is with the subjective probability assessments on which the system is initially based. The consultant physician has been required to assess prior probabilities for each of the diseases conditional on the age group of the patient and the main presenting symptoms, as well as the probabilities of the secondary symptoms conditional on each of the diseases. Around each probability he was required to place an interval reflecting his confidence in the point probability. By viewing this range as an approximate 90% interval around a binomial

probability one can derive a rough implicit sample size on which his judgment of each probability has been based. These measures of imprecision are currently not propagated through the consultation, although Rauch (1984) suggests ad hoc methods of doing this while allowing for correlated judgments. However, the implicit sample sizes allow the probabilities to be stored as a fraction r/n, and where a confirmed case with the relevant symptom is found the probability may be updated to $(r + 1)/(n + 1)$. This emphasises that probabilistic systems may be based on subjective opinion, and yet a rational means of allowing that opinion to learn from experience is easily available.

IMMEDIATE—A System for General Practice

In contrast to GLADYS, IMMEDIATE is a rule-based AI system written in PROLOG which is being developed by a group centered at the Medical Computation Unit at the University of Manchester. It is designed to support certain activities of general practitioners and its control philosophy is described elsewhere (Dodson and Rector, 1985).

Two aspects of its development are of interest here. Firstly, although the knowledge structure and uncertainty propagation bears some resemblance to that of PROSPECTOR, a deliberate aim is that the probabilities should be made to cohere: thus initial probability judgments should form a valid joint distribution, and, as data arrives, uncertainty be propagated in a way that retains its interpretation as subjective probability. Secondly, part of the control mechanism is based on a range of ignorance or evidence availability that is an explicit calculation of the maximum and minimum probabilities of a proposition that could be achieved when further information becomes available. This may be seen as a summary measure of the predictive distributions of final probabilities described under GLADYS. Explicitly calculating the range of potential probabilities of a proposition helps toward an assessment of the importance of establishing relevant patient characteristics, which in turn ensures that the clinician is informed as to the most telling questions to ask.

4. DISCUSSION

The preceding section is an inadequate glimpse of some work currently being carried out in probabilistic systems, and we have only been able to mention aspects according to their capacity to illustrate the practical implementation of important issues in the handling of uncertainty. In this section, we attempt to summarize these issues with the aid of examples drawn from the systems introduced above.

Status of Propositions

It is clearly preferable that all propositions in a system are crisply defined and, at least theoretically, verifiable at some point in the future, as required by Smith (1961) or de Finetti (1974). Nevertheless, the inevitable imprecision of statements (e.g., "the pain is relieved by food") makes it tempting to allow degrees of truth of propositions and adapt a fuzzy calculus. It should, however, be emphasized that it is not the true state of the world to which the system has access, but the *assertion* of the state of the world (The patient has replied YES to the question "Is the pain relieved by food?"), and this is necessarily made crisp by the restricted means one has to put information into the system (e.g., just a YES/NO button). An expert system can therefore force the user to be categorical in his assertions, although we acknowledge that user demand for qualifications of degree may create the need for an alternate calculus to deal with partly true propositions.

A statistician may tend to view a knowledge base as a set of related nodes, each corresponding to a random variable which may take on a number of mutually exclusive and exhaustive values. The rules attempt to define a distribution on the variables. For control purposes, however, it may be necessary to have action nodes which correspond to conclusions on which further analysis is conditioned. These may well not be strictly verifiable propositions; for example, in a system designed for statistical analysis, there may be assertions of normal errors or linear relationship. Strictly speaking a decision-theoretic argument should be used for any interim decision made in the control of a consultation, but this is not usually practicable. As suggested in the "Introduction," the justification for probability is not so clear in these cases, instead it could be reasonable to adopt a calculus of compatibility or degree of support for a hypothesis or conclusion for which a probability is not well defined.

Knowledge Representation and Explanation

We feel that probabilistic methods can handle hierarchical taxonomic structures without extending into belief function methodology (Pearl, 1986b). There is, however, a great need for further work in coherent assessment and propagation of probabilities through the network structures arising from rule-based systems. The graphical representations of certain log linear models described by, for example, Wermuth and Lauritzen (1983) are crucial, with propagation schemes extended from those of Kim and Pearl (1983); Spiegelhalter (1986b) describes efficient propagation schemes allowing for imprecise probabilities and automatic tuning of the subjective assessments.

Subjective judgments may be deliberately *overspecified* to allow for identification of incoherence due to poor assessments or weak modelling, or *underspecified* and padded out using, for example, the maximum entropy methods of Cheeseman (1983). By using such a structure and explanation facilities similar to GLADYS, one should be able to fulfill the aim, described by Dempster (1985), of justifying quantified judgment explicitly in terms of the sources of evidence.

Intervals and Probabilities

As we emphasised in discussing GLADYS, two types of range of probability must be distinguished. The first, due to inadequacies in the knowledge base, concerns the *imprecision* in the quantifications. This may be represented by a standard error or even a fuzzy qualifier, but in either case the range represents a type of automatic sensitivity analysis conditional on the data already obtained. This interval will generally tend to widen as more data come in.

This should be contrasted with an interval based on *ignorance* concering the current case, and one way in which this can be defined is in terms of the probabilities that could be taken on when the unknown data, denoted X, becomes available. If D represents a disease with current probability $p(D)$, then the predictive distribution of the eventual probability $p(D \mid X)$ may either be fully calculated as in GLADYS or summarized by its range as in IMMEDIATE. We note that by conditional expectation, $E\{p(D \mid X)\} = p(D)$. Hence our current probability may simply be thought of as the mean of the distribution of possible final probabilities. This distribution narrows as the consultation proceeds.

In this way *ignorance* is explicitly defined in terms of the X that we do not yet know. In real life, X is unbounded and so such a calculation is unreasonable, but it is important to note that an expert system is *bounded* and so can always explicitly state what information is missing, provided a suitably efficient search routine is available.

Operational Meaning

Our practical experience has strongly influenced us toward establishing operational meaning to our expression of uncertainty. This has three stages: firstly, the *inputs*, based on either real or imaginary past data, must have sufficient interpretation to allow informed argument. Clinicians often disagree strongly about subjective probabilities, but we have found the resulting discussions illuminating and constructive. The problems of agreeing on numbers with no verifiable interpretation is vividly illustrated in the fascinating transcript of an argument concerning certainty factors contained in the book on the MYCIN projects (Buchanan and Shortliffe, 1984). Secondly, preserving operational meaning in the propagation of uncertainty requires attention to the coherence of the assessments when placed in a large, complex knowledge base. Finally, the *outputs* need to have an externally verifiable interpretation in terms of their calibration against experience. Such calibration is not part of the axioms of subjective probability, but we have found an enthusiastic response from clinical colleagues when they find the predictions provide reasonable betting odds. Of course, a system may process information solely with the aim of providing a, possibly ranked, set of alternatives with some attached measure of evidential support. However, if a system is to be used to guide the *choice* of an option, or its outputs are to be used as inputs to another system, this seems to be inadequate. In fact, a subjectivist statistician may view a diagnostic expert system as a coherence machine, which takes in relevant information, and throws out acceptable betting odds on future events.

Finally, perhaps the most important reason for interpretable quantification is the need for *learning*. As we have illustrated with the chest disease system, updating of subjective probabilities is feasible and should provide a convergence of opinion that may overcome local biases which may otherwise render a system unacceptable.

REFERENCES

ADLASSNIG, K. P. (1980). A fuzzy logical model of computer-assisted medical diagnosis. *Methods Inf. Med.* 9 141–148.

BUCHANAN, B. G. and SHORTLIFFE, E. H., eds. (1984). *Rule-based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.* Addison-Wesley, Reading, Mass.

CHEESEMAN, P. (1983). A method of computing generalised Bayesian probability values for expert systems. *Proc. of the Eighth International Joint Conference on Artificial Intelligence.* 198–202. William Kaufmann, Los Altos, Calif.

CHEESEMAN, P. (1985). In defense of probability. *Proc. of the Ninth International Joint Conference on Artificial Intelligence.* 1002–1009.

COHEN, P. R. (1985). *Heuristic Reasoning about Uncertainty: An Artificial Intelligence Approach.* Pitman, Boston.

DE DOMBAL, F. T., LEAPER, D. J., STANILAND, J. R., MCCANN, A. P. and HORROCKS, J. C. (1972). Computer-aided diagnosis of acute abdominal pain. *British Med. J.* 2 9–13.

DE FINETTI, B. (1974). *Theory of Probability* 1. Wiley, New York.

DEMPSTER, A. P. (1985). Probability, evidence and judgment. In *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds.) 2 119–131. North-Holland, Amsterdam.

DODSON, D. C. and RECTOR, A. L. (1985). Importance-driven distributed control of diagnostic inference. In *Research and Development in Expert Systems* (M. A. Bramer, ed). Cambridge Univ. Press, Cambridge.

DUDA, R. O., HART, P. E. and NILSSON, N. J. (1976). Subjective Bayesian methods for rule-based inference systems. *Proc. AFIPS Natl. Compt. Conf.* 47 1075–1082.

FIESCHI, M., JOUBERT, M., FIESCHI, D., BOTTI, G. and ROUX, M. (1983). A program for expert diagnosis and therapeutic decision. *Med. Informatics* 8 127–135.

GORDON, J. and SHORTLIFFE, E. H. (1984). The Dempster-Shafer theory of evidence. In Buchanan and Shortliffe (1984), 272–292.

KANAL, L. N. and LEMMER, J., eds. (1986). *Uncertainty in Artificial Intelligence*. North Holland, Amsterdam.

KIM, J. H. and PEARL, J. (1983). A computational model for causal and diagnostic reasoning in inference systems. *Proc. of the Eighth International Joint Conference on Artificial Intelligence*. 190–193. William Kaufmann, Los Altos, Calif.

KULIKOWSKI, C. A. and WEISS, A. M. (1982). Representation of expert knowledge for consultation: the CASNET and EXPERT projects. In *Artificial Intelligence in Medicine* (P. Szolovits, ed.) 21–55. Westview Press, Colorado.

LUCAS, R. W., CARD, W. I., KNILL-JONES, R. P., WATKINSON, G. and CREAN, G. P. (1976). Computer interrogation of patients. *British Med. J.* 2 623–625.

MILLER, R. A., POPLE, H. E., JR. and MYERS, J. D. (1982). INTERNIST-1: An experimental computer-based diagnostic consultant for general internal medicine. *N. Engl. J. Med.* 307 468–476.

PAUKER, S. G., GORRY, G. A., KASSIRER, J. P. and SCHWARTZ, W. B. (1976). Towards the simulation of clinical cognition: taking a present illness by computer. *Amer. J. Med.* 60 981–986.

PEARL, J. (1986a). On evidential reasoning in a hierarchy of hypotheses. *Artificial Intelligence* 28 9–15.

PEARL, J. (1986b). Fusion, propagation and structuring in belief networks. *Artificial Intelligence* 29 241–288.

QUINLAN, J. R. (1983). Inferno: a cautious approach to uncertain inference. *Comput. J.* 26 255–269.

RAUCH, H. E. (1984). Probability concepts for an expert system used for data fusion. *Artificial Intelligence Mag.* 55–60.

SMITH, C. A. B. (1961). Consistency in statistical inference and decision (with discussion). *J. Roy. Statist. Soc. Ser. B* 23 1–37.

SPIEGELHALTER, D. J. (1986a). A statistical view of uncertainty in expert systems. In *Artificial Intelligence and Statistics* (W. Gale, ed.) 17–56. Addison-Wesley, Reading, Mass.

SPIEGELHALTER, D. J. (1986b). Probabilistic reasoning in predictive expert systems. Cited in Kanal and Lemmer (1986).

SPIEGELHALTER, D. J. and KNILL-JONES, R. P. (1984). Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology (with discussion). *J. Roy. Statist. Soc. Ser. A* 147 35–77.

SZOLOVITS, P. and PAUKER, S. G. (1978). Categorical and probabilistic reasoning in medical diagnosis. *Artificial Intelligence* 11 115–144.

TUKEY, J. W. (1984). Discussion of Spiegelhalter and Knill-Jones. *J. Roy. Statist. Soc. Ser. A* 147 62–64.

VAN MELLE, W., SCOTT, A. C., BENNETT, J. S. and PEAIRS, M. A. S. (1981). The EMYCIN Manual. Report HPP-81-16, Computer Science Dept., Stanford Univ.

WERMUTH, N. and LAURITZEN, S. L. (1983). Graphical and recursive models for contingency tables. *Biometrika* 70 537–552.

Comment

Stephen R. Watson

1. COMMENTS ON SHAFER'S PAPER

One of the things that makes Shafer's theory interesting is that it can be seen as an alternative to the traditional probability theory. Is this really so, however? Firstly, note that one of the strengths of subjective probability theory is the clear cut nature of the axiomatic support for the theory. Indeed, as Lindley's contribution shows, it is possible to claim that probability theory is the only theory one could possibly use to represent uncertainty. Shafer's theory does not as yet have such a clear cut support. For example, although Shafer recognizes the importance of canonical examples, as yet belief function theory is not provided with as strong an axiomatic support as that which is available for probability theory.

It can be claimed, however, that belief functions are indeed rooted in probability theory. It is just that the probability is associated with a power set rather than a simple set. If this interpretation of belief function theory is accepted, then indeed there is no problem, since the philosophical support for probability theory clearly also will support belief function theory. However, Shafer seems in some of his writings not to be very happy with this interpretation of his theory. And if he rejects this interpretation then the problem of a philosophical foundation for belief function theory remains.

The second point I make here concerns concepts of independence. Shafer touches on this point in his paper, but it is worth saying again that concepts of independence in belief function theory are not yet clear. In the application of Dempster's rule to determine the support for a hypothesis on the basis of two pieces of evidence, there is a rather vague notion that the two pieces of evidence should be independent in

Stephen R. Watson is Peat Marwick Professor of Management Studies in the University of Cambridge. He is also Head of the Management Studies Group in the Department of Engineering and a Fellow of Emmanuel College. His address is Department of Engineering, University of Cambridge, Mill Lane, Cambridge CB2 1RX, England.

some way. The detailed meaning of this concept of independence is far from clear. Shafer recognizes this difficulty and in his discussion of frames is attempting to overcome it. It is sufficient to say at this point, however, that we do not yet know how to handle dependence concepts in belief function theory in a way which is intuitively understandable.

2. COMMENTS ON LINDLEY'S PAPER

The conviction with which Professor Lindley speaks, and the sheer power of his argument impel users of alternatives to probability theory to respond to his arguments. If we do not accept the inevitability of probability, why not?

Users of Shafer's theory or Zadeh's theory can, and in fact have in the past, respond that they do indeed accept the inevitability of probability. As Dempster has commented, belief function theory is founded on probability, and so there is no contradiction in using belief function theory at the same time as using probability theory. Moreover, one can think of fuzzy set theory as being a heuristic approach in situations where a full probabilistic analysis is far too complicated to be undertaken.

It is, however, also possible to take issue with Lindley's argument. In other words, it is possible to question some of the premises in his argument and thereby avoid the full power of his conclusions. If one investigates the development of subjective probability theory exemplified by Savage's approach, it is possible to ask whether we are prepared to accept the axioms. It is a commonplace now that people do not behave as though they accept Savage's axioms, reasonable as they undoubtedly are. Of course, these axioms are normative and it can be argued, as indeed Lindley does argue, that the fact that we fail to abide by the axioms does not mean that we should not attempt to do so. Indeed he would say that the first act of a rational man is to agree to the axioms, and then attempt to construct his behavior in accordance with these axioms. If, however, we are not prepared to do this, then what happens to us is a matter of practice. It could be argued that if we are consistent in our failure to abide by the axioms, then our opponents can turn us into a money pump or construct a Dutch Book of gambles against us. Of course, we do not do this in practice. We just recognize when we are about to get cornered in this way, and change one of our judgments, possibly in a yet more inconsistent way with our past judgments. There is, therefore, nothing mandatory about accepting Savage's axioms, and we can therefore escape Lindley's conclusions if we wish to.

In his contribution, Lindley gave a very clear account of an alternative way of showing the inevitability of the probability. This was based on the notion of scoring systems. It is indeed quite remarkable that no matter what kind of scoring system one adopts, the numbers that one employs to describe uncertainty must (after an appropriate transformation) satisfy the rules of probability theory. Compelling as this argument is, we have to point out that in practice no Great Scorer exists. There is nobody hovering about us being prepared to dock our pay should we use numbers which fail to conform to the rules of probability theory in our descriptions of uncertainty. Thus while the argument is elegant and powerful, there is nothing inherently irrational in not accepting it, because in practice scoring systems do not exist.

Of course the proof of the pudding is in the eating. If it can be shown that in the long run any person who fails in his assessment of uncertainty to combine his numbers as though they were probabilities will lose out inexorably, then indeed we have a problem in refusing to accept probability theory. But to my understanding practical proofs of this kind are not yet available.

Thus, it is possible to escape the inevitability of probability; it has to be admitted, however, that there is no alternative theory which has the strength of support, and elegant support at that, which is available for probability theory.

The chief drawback with using probability theory is the complexity that sometimes results, and the need to assess an often surprisingly large number of conditional probabilities. In legal work, for example, great difficulty can arise; some interesting work by Schum (1981) shows how problematic probabilistic inference can get. In one simple murder case, with five pieces of evidence, he needed to make 27 probability assessments. Lindley suggests the principle of Occam's razor should be applied to our topic: simplify where possible. Sometimes probabilistic analysis is far from simple.

3. COMMENTS ON SPIEGELHALTER'S PAPER

Spiegelhalter's paper is a most interesting account of the construction of an expert system for medical diagnosis. He gives us some important insights into the practical problems of constructing an expert system, which is both computable and also useful. This raises the general question of how one determines whether a particular expert system, as represented on some computer, is actually a good one or not. The issues involved are very similar to those involved in validating a model. Firstly, one needs the system to be faithful to some normative principle. In my view one should start with probability theory since it has the strongest theoretical base, but be prepared to

adopt other approaches as heuristics or as richer representations of the issues involved. It seems that Spiegelhalter's approach has been similar.

Secondly, one could validate an expert system by its comparison with expert performance. One can ask whether the diagnosis achieved by Spiegelhalter's system was better or worse than that achieved by competent diagnosticians. There is of course a debate over whether an expert system should be appraised in this way. Is the goal to reproduce the abilities of an expert, or to improve on the abilities of available human judges? If it is the former, then indeed it is sensible to compare performance with experts, but in this case one wonders why one should not use the experts themselves. This could be answered by observing that very often experts are in short supply. If, on the other hand, our goal is to improve on human inference behavior, then the criterion of conformity with some expert performance is not appropriate. A final measure of the appropriatness of an expert system is user satisfaction. To what extent do the people who interact with the expert system feel that the system is of use to them? In Spiegelhalter's case there are two kinds of people involved, namely the patients and the doctors. As Spiegelhalter observes, it is very important that the doctors are supportive of the endeavor and that they do not feel that their professional competence is in any way being threatened. It is perhaps more important, however, that the patients feel that they are being properly attended to. Spiegelhalter seems to have achieved success on both fronts.

4. SUMMARY

Although the purpose of the conference was to discuss the use of the different theories for the representation of uncertainty in expert systems, the principal speakers, perhaps wisely, devoted their discussion mainly to arguing the cases for the use of their different theories in general. On the basis of the discussions we had at this conference, it seems to me that one can summarize as follows. Probability theory has a strong intellectual support and in principle there is no reason why one should not be satisfied with this theory. Its use does, however, lead to enormous problems of complexity, and as a matter of practice it is necessary to seek for approximations. Fuzzy set theory can be viewed as a heuristic for handling those situations where imprecise inputs and imprecise inferences are required without the need to resort to the greater complexity of probability theory. Belief function theory can be thought of as a way of representing inferences from evidence within the probabilistic framework.

There are yet other alternative approaches to handling uncertain inferences which were not mentioned at the conference, and notable among these is the nonmonotonic logic of Doyle. Recently Cohen (Cohen, Watson and Barrett, 1985) has suggested a combination of Doyle's theory with both Shafer's and Zadeh's which he has referred to as the nonmonotonic probabilist. This seems an exciting possibility for approaching the problem at the heart of this conference.

ADDITIONAL REFERENCES

COHEN, M. S., WATSON, S. R. and BARRETT, E. (1985). Alternative theories of inference in expert systems for image analysis. Technical Report 85-1, Decision Science Consortium, Falls Church, Va.

SCHUM, D. A. (1981). Sorting out the effects of witness sensitivity and response-criterion placement upon the inferential value of testimonial evidence. *Organizational Behavior and Human Performance* 27 153–196.

Comment

A. P. Dempster and Augustine Kong

The papers by Shafer and Spiegelhalter are valuable summaries by acknowledged leaders in active research fields. There is much food for thought in both papers, and many of the techniques and issues raised by these authors will gradually become better understood as the field of uncertainty assessment in expert systems advances. Our research on models and techniques for belief function analysis (Kong, 1986; Dempster and Kong, 1986) is complementary to that of Shafer and Spiegelhalter. We all seek to provide tools for real applications, based on carefully constructed analyses expressed through mathematically well-articulated principles of uncertain reasoning.

Lindley is on a different track. He rehearses familiar normative arguments for the Bayesian paradigm, evidently seeking to persuade less committed colleagues to abandon their fallacious ways. Unfortunately, he shows no interest in understanding how his

A. P. Dempster is a Professor and Augustine Kong is an Instructor at the Department of Statistics, Harvard University, Cambridge, Massachusetts 02138.

competitors really think, and hence, does not address the issues on which, in our opinion, credible contemporary debate should focus. As illustrated by Shafer's hypothetical "ploxoma" example, his "challenge" is unconvincing, for he casts himself first in the role of challenger, then of umpire, and finally reverts to challenger, proclaiming himself well satisfied with the result.

Lindley oversimplifies by identifying Bayes with the use of probability. In fact, numerical probabilities which are both syntactically and semantically very close to Lindley's probabilities are essential to three alternative approaches ("classical statistics," "upper and lower probabilities," and "belief functions") which he criticizes. But surely it is first necessary to understand the various styles of reasoning with probability implied by each of these systems, before either choosing among them or judging which circumstances are appropriate for each. In our view, moreover, the belief function system is very close to Bayes, and indeed includes Bayesian models as special cases, so it is not easily rejected in favor of Bayes except by arguments whose artificiality is painfully obvious from the belief function standpoint.

We make no attempt here to defend classical statistics, upper and lower probability systems, or fuzzy logic, where the last seems fundamentally different from the others, but we can accept that each may have an appropriate place in valid and useful formal analyses. Instead, we comment briefly on the flexibility which belief function theory adds to Bayesian theory in its ability to incorporate evidence. Then we discuss at greater length the connection between belief functions and decision theory.

Lindley repeats verbatim his discussion of the Shafer (1982) "ploxoma" example, as he says, to provoke further discussion. As matters stand, Shafer has not modified his original representation, which contains three belief function components: (a) a range .05 to .15 to describe the prior probability of "virulent ploxoma," (b) a vacuous belief function to describe the inability of the "ordinary ploxoma" experiment to distinguish between x_2 and x_3, and (c) a 25% discounting applied to the experimental data about ordinary ploxoma. Evidently, these features were introduced to illustrate the flexible forms of uncertainty representation encompassed by the belief function paradigm. Lindley's response is to suggest converting Shafer's analysis to Bayesian form by altering (a) to a single prior probability .1, (b) to the indifference prior assigning .5 to each of x_2 and x_3 given that one or the other has occurred, and (c) to $E(\gamma_i \mid \beta_i) = \beta_i$ where $(\gamma_1, \gamma_2, \gamma_3)$ refers to valid chances of (x_1, x_2, x_3) given ordinary ploxoma for the new patient George, although $(\beta_1, \beta_2, \beta_3)$ refers to the questionable experimental results.

So far, neither side has explicitly addressed their differences over (a) or (b). But Lindley does now respond to the Shafer (1982) rebuttal of his altered (c), by allowing that we might have "no confidence at all" in the ordinary ploxoma study, in which case "$E(\gamma \mid \beta)$ would not depend on β." That is, 100% discounting means that the Bayesian constructs $E(\gamma_i)$ from wherever Bayesians construct such priors, thus implicitly introducing other sources of information processed together in the Bayesian's head to produce the prior. Presumably, if the Bayesian were, more realistically, to adopt less drastic discounting, the same prior about $(\gamma_1, \gamma_2, \gamma_3)$ would still be assessed and combined with information from the data via another assessed prior $f(\gamma \mid \beta)$. Thus, Bayesian analysis is not at all simple in execution, if one takes it seriously.

Belief function methodology does introduce more complexity into the class of available representations of uncertainty, although not typically into the task of assessing specific representations. Lindley criticizes the mathematical generalization as lacking "necessity" in the sense of William of Ockham. The important question is whether the added flexibility is necessary in practice to permit satisfactory representation of an analyst's state of uncertainty about the real world. We believe that it is literally impossible to answer the question outside the context of real examples based on attempts to construct formal representations of uncertainty reflecting actual uncertain knowledge of the real world. Because their example is purely hypothetical, neither Shafer nor Lindley is able to discuss in any specific way the construction of their specific models.

Important points about Lindley's "challenge" are, first, that a meaningful test must deal with real examples, and, second, that the umpire must rely on some assessment of help given to third party clients. Bayesian decision analysis has been out in the field for about 30 years, and in our (subjective) assessment has achieved only limited penetration into what might seem to be its natural markets. If this failure to penetrate were simply due to ignorance of the techniques on the part of practitioners, then Lindley's proselytizing might have a point. A more plausible explanation is that practical construction of realistic Bayesian models is typically very difficult. Because belief function analysis does not pretend to the vague and difficult goal of integrating all evidence available to the analyst, but instead attempts only to represent explicit and limited packets of independent evidence, the process of constructing belief function models is inherently simpler, and on this score belief function methods have excellent prospects for success in practice. Of course, computational difficulties are another matter, and here the tradeoffs are less clear, because

computationally neither approach has advanced beyond its infancy. All in all, there would seem to be sound practical reasons for seeking to relax Bayesian constraints to obtain more flexible, explicit, and realistic representations of uncertainty.

Since the 1950s, both Bayesian and frequentist decision theorists have agreed that Bayesian expected loss is the appropriate numerical criterion for comparing possible acts. Such a Bayesian analysis introduces a precise ordering among acts, in the sense that the set of all acts is partitioned into subsets where the analyst is indifferent within a subset but has a complete preference order between subsets. Since general belief functions replace expected loss by upper and lower expected losses, moving to a belief function framework implies a tradeoff. On one hand the analyst simplifies inputs to specific sources of evidence, while paying on the other hand by having only partial orderings among acts. The partial orderings arise because the single numerical Bayesian expectation is replaced by numerical upper and lower expectations.

Lindley evidently does not wish to allow partial ordering, but do the standard normative arguments which he presents really prohibit it? For example, the scoring rule argument posits an artificial decision problem, where the acts are possible choices of a numerical measure of uncertainty about some unknown binary state, and the losses are heuristic quality assessments of the numerical measure of uncertainty. The conclusion for this decision problem, as for decision problems generally, is that Bayesian decision rules are the only admissible decision rules. More precisely, the conclusion is that, if there is a rule that selects a single act from the available set of acts, the rule must minimize expected loss under some probability distribution. Note, however, that the condition "if there is a rule" tacitly prejudges the question at issue. For if we choose to report only a partial ordering, we are, in effect, opting to specify no rule, so the admissibility result becomes irrelevant.

Lindley abuses the theory of belief functions by substituting belief into a scoring measure as though it were a simple probability. It may therefore be helpful to sketch what we see as the right way to think about belief functions in a decision-theoretic framework. To illustrate, consider a decision problem with decision space $D = \{d_1, d_2, d_3\}$, outcome space $W = \{w_1, w_2\}$ and loss function given in Table 1. Let $D^* = \{d(p_1, p_2, p_3) \mid p_1, p_2, p_3 \geq 0, \sum_{i=1}^3 p_i = 1\}$, where $d(p_1, p_2, p_3)$ denotes the randomized decision that selects decision d_i, $i = 1, 2, 3$, with probability p_i. Following DeGroot (1970),

$$L(w_1, d(p_1, p_2, p_3)) = 10p_2 + 20p_3 \tag{1.1}$$

and

$$L(w_2, d(p_1, p_2, p_3)) = 45p_1 + 20p_2 + 10p_3$$

where $L(w, d)$ denotes loss. From Figure 1 it is clear that a decision $d(p_1, p_2, p_3)$ is admissible in the ordinary decision theory sense if either $p_1 = 0$ or $p_3 = 0$, thus including the pure decisions d_1, d_2, and d_3. The minimax decision is easily computed to be $d(0, \frac{1}{2}, \frac{1}{2})$ where

$$L(w_1, d(0, \tfrac{1}{2}, \tfrac{1}{2})) = L(w_2, d(0, \tfrac{1}{2}, \tfrac{1}{2})) = 15.$$

Suppose our knowledge about the outcome is represented by the belief function Bel over W. Let $\{\mu\}_{\text{Bel}}$ be the collection of probability measures μ over W that satisfy

$$\text{Bel}(A) \leq \mu(A) \leq Pl(A)$$

for all $A \subset W$. For $d, d' \in D^*$, we say d is *uniformly dominated* by d' with respect to Bel if

$$E\{L(w, d) \mid \mu\} \geq E\{L(w, d') \mid \mu\}$$

for all $\mu \in \{\mu\}_{\text{Bel}}$ and there exists $\mu^* \in \{\mu\}_{\text{Bel}}$ such that

$$E\{L(w, d) \mid \mu^*\} > E\{L(w, d') \mid \mu^*\},$$

where $E\{\cdot \mid \mu\}$ denotes expectation computed based on μ. We call a decision *permissible* against Bel if it is not uniformly dominated by another decision in D^* with respect to Bel. Hence, a decision is admissible if it is permissible against the vacuous belief function.

Suppose Bel is

$$\begin{aligned} m(\{w_1\}) &= .6, \\ m(\{w_2\}) &= .2, \\ m(\{w_1, w_2\}) &= .2. \end{aligned} \tag{1.2}$$

It follows that $\{\mu\}_{\text{Bel}} = \{\mu_t \mid .6 \leq t \leq .8\}$, where μ_t denotes the probability measure that assigns probability t to w_1 and probability $1 - t$ to w_2. For the pure decisions,

$$E\{L(w, d_1) \mid \mu_t\} = 45(1 - t),$$

$$E\{L(w, d_2) \mid \mu_t\} = 10t + 20(1 - t) = 20 - 10t,$$

$$E\{L(w, d_3) \mid \mu_t\} = 20t + 10(1 - t) = 10 - 20t.$$

Figure 2 plots $E\{L(w, d_1) \mid \mu_t\}$, $E\{L(w, d_2) \mid \mu_t\}$, and $E\{L(w, d_3) \mid \mu_t\}$ for $.6 \leq t \leq .8$. It shows that d_3 is uniformly dominated by d_2 with respect to Bel. Since $E\{L(w, d(p_1, p_2, p_3)) \mid \mu\} = \sum_{i=1}^3 p_i E\{L(w, d_i) \mid \mu\}$, it is straightforward to prove that $d(p_1, p_2, p_3)$, $p_3 > 0$, is uniformly dominated by $d(p_1, p_2 + p_3, 0)$ with respect

TABLE 1
Loss function

	d_1	d_2	d_3
w_1	0	10	20
w_2	45	20	10

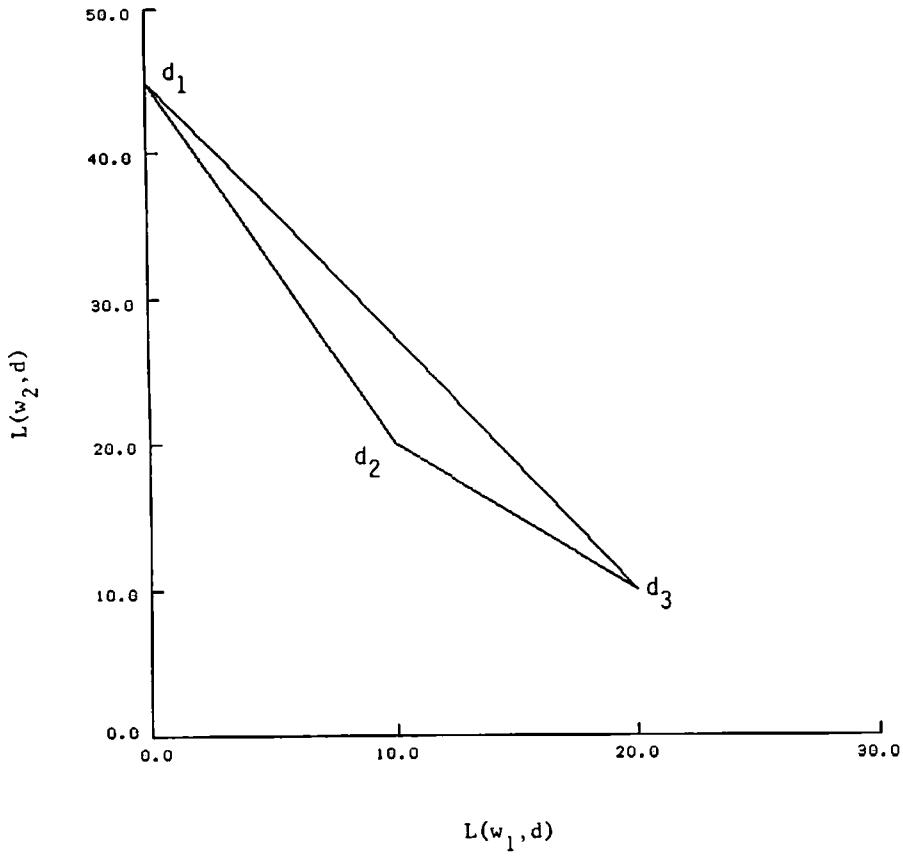

FIG. 1. $L(w_1, d)$ *versus* $L(w_2, d)$.

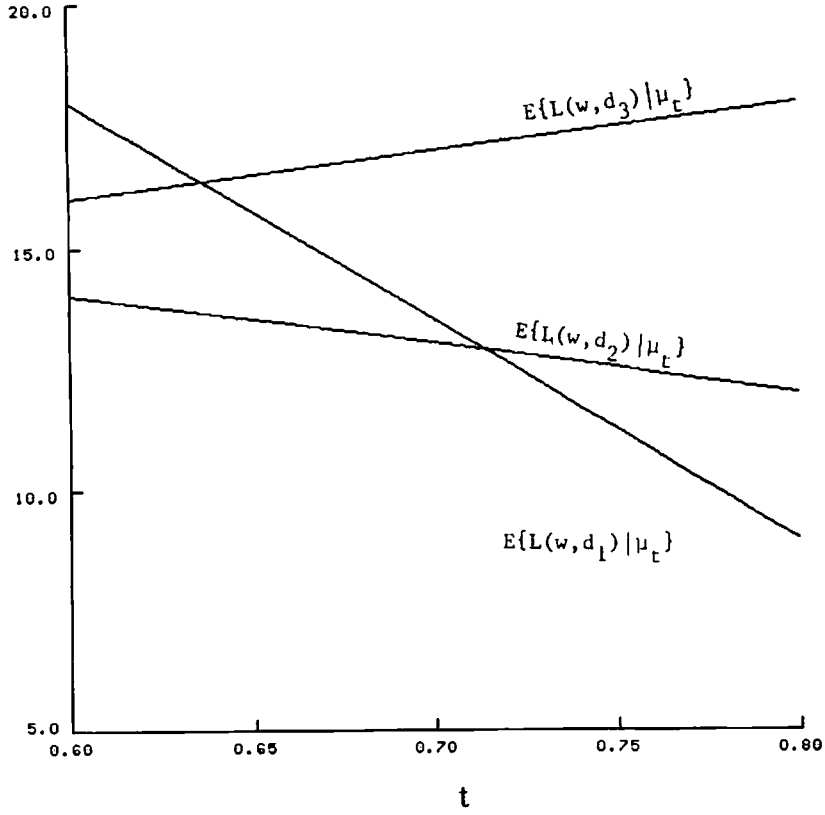

FIG. 2. *Expected loss.*

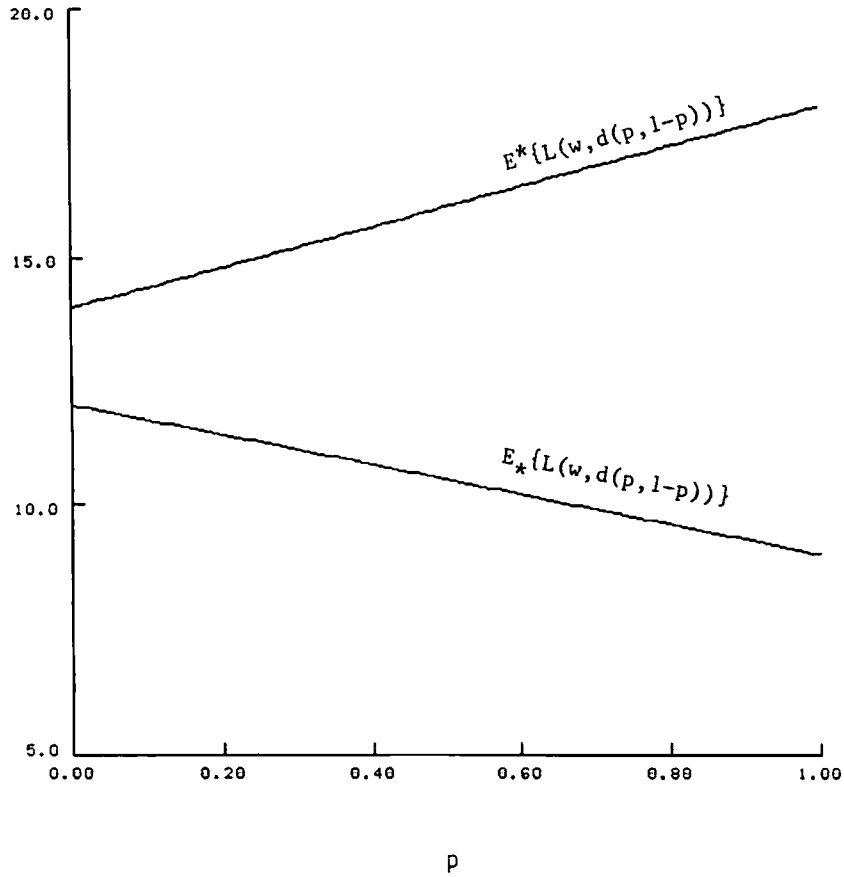

FIG. 3. *Upper and lower expected loss.*

to Bel. The permissible decisions against Bel are $d(p, 1-p, 0), 0 \le p \le 1$.

For a decision $d \in D^*$ define lower and upper expected loss with respect to Bel to be (cf. Dempster, 1967)

$$E_*\{L(w, d)\} = \inf_{\mu \in |\mu|_{\text{Bel}}} [E\{L(w, d) \mid \mu\}]$$

and

$$E^*\{L(w, d)\} = \sup_{\mu \in |\mu|_{\text{Bel}}} [E\{L(w, d) \mid \mu\}].$$

Figure 3 plots the upper and lower expected loss of the permissible decisions $d(p, 1-p, 0)$ as functions of p. Since $d(0, 1, 0) = d_2$ has the minimum upper expected loss, we call d_2 the miniupper decision against Bel.

Notice that the miniupper decision against a vacuous belief function is the minimax decision and the miniupper decision against a Bayesian belief function is the corresponding Bayes decision. Hence, the miniupper method is a generalization of minimax and Bayes. Under more general settings where there can be more than two outcomes, it can be shown that the task of finding the miniupper decision can be reformulated as a linear programming problem. Details will be given in a coming technical report.

We are not necessarily endorsing the miniupper decision here. Indeed, in the above example, we have no reason to fault someone who chooses d_1 over d_2. The point is that some guidance toward rational decisions can be made even if uncertainty is represented by belief functions instead of distribution functions.

ACKNOWLEDGMENTS

This work was supported in part by Office of Naval Research Contract 00014-85-K-0496 and Army Research Office Grant DAAL03-86-K-0042.

ADDITIONAL REFERENCES

DEMPSTER, A. P. (1967). Upper and lower probabilities induced by a multivariate mapping. *Ann. Math. Statist.* **38** 325–339.

DEMPSTER, A. P. and KONG, A. (1986). Uncertain evidence and artificial analysis. Research Report S-108, Dept. Statistics, Harvard Univ.

KONG, A. (1986). Multivariate belief functions and graphical models. Ph.D. dissertation, Dept. Statistics, Harvard Univ.

Comment

Glenn Shafer

I found it a pleasure to read the articles by Dennis V. Lindley and David Spiegelhalter. They present an elegant case for the use of Bayesian (i.e., conditional probability) methods in expert systems. Lindley provides a concise summary of arguments he and others have developed over the last several decades in support of the claim that rationality demands the use of conditional probability. Spiegelhalter supplements this with an account of what is actually being accomplished using conditional probability in diagnostic systems, and he also contributes some interesting new arguments.

In response, let me first express my admiration for the practical work Spiegelhalter reports on. The GLADYS system is especially attractive, because it brings close to reality the kind of probability calculation philosophers have always considered ideal—the calculation of conditional probabilities on the basis of massive and directly relevant frequency data. I share Spiegelhalter's excitement with the prospect that widespread use of microcomputers will enable us to attain this ideal much more often than we have in the past.

WHY GENERALIZE PROBABILITY?

Spiegelhalter emphasizes capabilities of the Bayesian language that are often overlooked. Weights of conflict can be used to monitor Bayesian analyses, and weights of evidence can be used to explain the results. There are Bayesian definitions of imprecision and ignorance. We do not need to generalize from Bayes to belief functions in order to formalize these concepts.

The desire to generalize Bayes does not spring, however, from dissatisfaction with the ideal of conditional probability. It springs from the realization that this ideal is sometimes unattainable. Directly relevant frequencies are often unattainable. Sometimes we can make decent conditional probability arguments even without such frequencies, but sometimes we cannot. Sometimes we simply lack evidence for some of the probability judgments that a given conditional probability analysis requires.

The only satisfactory description of uncertainty, Lindley tells us, is probability. He is no less correct than the man who believes that the only satisfactory household is one with a dozen servants. It's wonderful if you can afford it.

STANDARDS OF RATIONALITY

What should we say about the claim that rationality demands we make Bayesian analyses regardless of the availability of the ingredients? For my own part, I find that every argument for this claim boils down to another appreciation of the beauty of the Bayesian ideal.

Lindley believes that Savage's axioms are so self-evident that their violation would look ridiculous. But in fact these axioms derive their appeal from the Bayesian ideal rather than vice versa. If we did not have the picture of conditional probability and expected utility in mind, we would not even be able to understand most of Savage's axioms (Shafer, 1986b).

The idea of a scoring rule also derives from the Bayesian ideal rather than vice versa. It has relatively little force in abstraction from that ideal. If we intend to assign a number to each of two complementary events and accept a penalty for each event based on the number's distance from one if the event happens and its distance from zero if it fails, then we should make the two numbers add to one. But how would we explain this game to a naive listener? We would say that the numbers are supposed to be like probabilities—close to one for events that are expected to happen and close to zero for events that are expected to fail. The game fits the picture of additive, or frequency-like, probability, and it is incomprehensible outside that picture. It does not fit the theory of belief functions, where a degree of belief close to zero indicates inadequate evidence for the event, not assurance that the event will fail.

Another argument for Bayes is based on the relatively sharp preferences given by expected utility calculations. We can calculate upper and lower expectations from belief functions, but these will not give a definite preference between two alternatives as often as the Bayesian calculation will. But would we expect such sharp preferences were it not for our fascination with the Bayesian ideal? Would we really expect an analysis of our evidence and pre-existent preferences to tell us always exactly what to do, leaving no occasion for caprice? In fact, human beings, unlike Buridan's ass, are capable of choosing without sufficient reason, and they often use that capability. Building a similar capability into a computer is one of the easier tasks of artificial intelligence.

CONSTRUCTIVE PROBABILITY

In my contribution to this symposium, I say that Bayesian analyses use games of chance as canonical examples to which to compare actual evidence. Lindley says such games provide a standard by which to measure belief. There are commonalities here, but

also important differences. It is difficult to use the verb "measure" without pretending that there is a well-defined property to be measured. Talk about canonical examples encourages a more constructive attitude.

One aspect of the constructive nature of Bayesian probability judgment, emphasized by Shafer and Tversky (1985), is the fact that we must construct our starting point. We must construct a probability distribution before we can condition it or multiply it by likelihoods. Bayesian theorists often assert categorically that every new experience must be treated in terms of its likelihood. Lindley, for example, declares that "an AI system faced with uncertainty about A_2 and experiencing A_1 has to update its uncertainty by considering how probable what it has experienced is, both on the supposition that A_2 is true, and that A_2 is false." But since a person may get around to constructing "initial" probabilities only after experiencing A_1, he or she has the option of treating A_1 as part of the evidence for those initial probabilities. Consider Lindley's investigator, who has discovered evidence that a criminal is left-handed. Instead of treating this evidence in terms of its likelihood, the investigator uses it directly in constructing a probability distribution.

There are problems, of course, where the construction can all be done in advance and then applied to many cases. GLADYS deals with this kind of problem; the same framework is applied to one patient after another. If I understand Spiegelhalter correctly, he believes that the bounded nature of expert systems means that this is the only kind of problem with which they can deal.

A finite system that permits construction can, however, deal with an unbounded range of situations. This is one of the fundamental points of the generative theory of grammar. The constructive nature of human reasoning makes us capable of exploring ever new realms of experience, and the ambition of AI is to duplicate this capability. Rule-based expert systems are one attempt to do so. These systems do not handle probabilistic reasoning very well, and many AI theorists would conclude from this that probabilistic reasoning has little role in genuine intelligence. In order to prove them wrong, we must do more than retreat to bounded systems like GLADYS. We must take the problem of automating construction seriously.

ADDITIONAL REFERENCE

SHAFER, G. (1986b). Savage revisited (with discussion). *Statist. Sci.* 1 463–501.

Comment: A Tale of Two Wells

Dennis V. Lindley

The main issue is whether uncertainty should be described by probability, belief functions, or fuzzy logic; not just in artificial intelligence and expert systems, but generally. Are we to be probabilists, believers, or fuzzifiers? Or do we need some mixture of all three disciplines? To me, the important distinction between the methods rests on the rules of combination of uncertainty statements. Do we operate with the calculus of probability, the rules of belief functions, or with those of fuzzy logic? In my paper the challenge was made "that anything that can be done by these methods (belief functions and fuzzy logic) can better be done with probability." This reply will address one such challenge and I hope to show that Dempster's rule for belief functions does not behave as well as Bayes rule. My discussion is therefore chiefly addressed to Shafer and Zadeh. The omission of any discussion of Spiegelhalter's contribution arises because I agree substantially with it, and highly regard it. I wish that his program for dyspepsia had been more Bayesian and that he had recognized that uncertainty about a probability is usually a reference to the desirability of obtaining more data, so that his conflict ratio should really reflect this. To return to the challenge.

In 1685 the then Bishop of Bath and Wells wrote a paper in which the following problem was discussed. Two witnesses separately report that an event is true. Both are known to be unreliable to the extent that they only tell the truth with probabilities p_1 and p_2 respectively. What reliability can we then place, in the light of the witnesses' testimonies on the truth of the event? The Bishop's answer was $1 - (1 - p_1)(1 - p_2)$. The following is a precis of his argument. If the event is false, both witnesses must have lied, an event of probability $(1 - p_1)(1 - p_2)$. Consequently one minus this is the required reliability.

The result retains its interest today because the

Bishop's rule of combination of the two pieces of evidence is the same as Dempster's rule used in belief-function calculus. So here we have a challenge: I maintain probability can do better. (Readers will notice that the example is similar to that of Slippery Fred, used by Shafer in his paper, but is somewhat simpler. It was introduced by Shafer in the oral discussion of the original papers.)

Almost 80 years later, in 1763, the rector of Tunbridge Wells, Thomas Bayes, introduced his rule, presumably being unaware of the Bishop's proposal. This is now known as Bayes' rule (of probability), which we now apply to the Bishop's problem.

Let A denote the event whose truth is in question, and write a_1 and a_2 for the statements by the two witnesses that A is true. Since A is uncertain and a_1, a_2 are known assertions, we have to calculate $p(A \mid a_1, a_2)$, the probability that A is true, given both a_1 and a_2, using the rules of the probability calculus. This probability, and all those subsequently calculated, are judgments by some person. When, later, the Bishop's values, p_1 and p_2, are used, it will be supposed that, suitably interpreted, they are accepted as his by this person.

It is easier to work with the odds rather than the probability. These satisfy Bayes' rule

$$\frac{p(A \mid a_1, a_2)}{p(\bar{A} \mid a_1, a_2)} = \frac{p(a_1, a_2 \mid A)p(A)}{p(a_1, a_2 \mid \bar{A})p(\bar{A})}. \tag{1}$$

On the far right we have the original odds on A before the witnesses gave their evidence. Write $p(A) = \pi$, so that the odds are $\pi/(1 - \pi)$. Also on the righthand side, in the numerator, occurs the probability $p(a_1, a_2 \mid A)$. This is the probability, were A true, that *both* witnesses would report it so; that is, tell the truth. The problem as formulated tells us nothing about this but there is a strong hint of independence in the original presentation—notice the multiplication $(1 - p_1)(1 - p_2)$—so if it is presumed here we might write $p(a_1, a_2 \mid A) = p(a_1 \mid A)p(a_2 \mid A)$, and similarly in the denominator, $p(a_1, a_2 \mid \bar{A}) = p(a_1 \mid \bar{A})p(a_2 \mid \bar{A})$.

Next consider one term in the numerator, $p(a_1 \mid A)$. This is the probability that the first witness will say the event is true when indeed it is true: in other words, tell the truth. But this is not the only way he could tell the truth: he could announce A was false when indeed it was false. This is $p(\bar{a}_1 \mid \bar{A}) = 1 - p(a_1 \mid \bar{A})$, which occurs in the denominator. The Bishop's argument used p_1, the probability of telling the truth, and to apply the rector's approach it is necessary to relate p_1 to $p(a_1 \mid A)$ and $p(\bar{a}_1 \mid \bar{A})$. If t_1 is the event that the first witness tells the truth, then

$$p(t_1) = p(t_1 \mid A)p(A) + p(t_1 \mid \bar{A})p(\bar{A}).$$

But t_1 when A is true (false) is $a_1(\bar{a}_1)$, so

$$p_1 = p(a_1 \mid A)\pi + p(\bar{a}_1 \mid \bar{A})(1 - \pi)$$

on inserting the Bishop's value p_1 for $p(t_1)$. The simplest assumption is that $p(a_1 \mid A) = p(\bar{a}_1 \mid \bar{A})$; that is, truth is just as likely when A is true as when it is false. It then easily follows that the common value is p_1.

Applying the same argument to the second witness, we easily have from (1) that

$$\frac{p(A \mid a_1, a_2)}{p(\bar{A} \mid a_1, a_2)} = \frac{p_1 p_2 \pi}{(1 - p_1)(1 - p_2)(1 - \pi)},$$

whence,

$$p(A \mid a_1, a_2) = \frac{p_1 p_2 \pi}{p_1 p_2 \pi + (1 - p_1)(1 - p_2)(1 - \pi)}. \tag{2}$$

It is this result that can be compared to the Bishop's $1 - (1 - p_1)(1 - p_2)$.

To reach the Bayesian result (2) some assumptions have been made. We list these and comment upon them.

I: $p(A) = \pi$ is known, and relevant to the answer.

Its relevance seems indisputable. Even the testimony of very reliable witnesses (p_1 and p_2 near 1) would leave some doubt in a person's mind about A if initially he thought it most improbable (small π). Conversely, unreliable witnesses would still leave him having appreciable probability for A if initially π was near 1. Since it is relevant, its value must be included in the calculations. This is perhaps the Bishop's main mistake: to fail to appreciate the importance of π.

II: a_1 and a_2 are independent, both given A, and given $\bar{A}$.

Notice that the independence assumption is quite subtle. It demands independence both when the event is true and when it is false—but not unconditionally. It is easy to imagine circumstances where one independence holds but not the other. Suppose A is the event that a defendant in a court of law truly committed the crime with which he has been charged. If A is true, two witnesses may collude in providing him with an alibi; if A is false, no such collusion is needed. So a_1 and a_2 may be independent given $\bar{A}$, but not given A.

The Bishop almost certainly was tacitly assuming independence in 1685. It is also supposed in the modern belief function treatment, and Dempster's rule only realistically applies when it obtains. The

Bayesian approach works without independence: it has only been assumed here for simplicity and comparison with beliefs. What the Bayesian view does is to force one to consider the subtle nature of the dependence between the witnesses.

III: $p(a_i \mid A) = p(\bar{a}_i \mid \bar{A}), \quad (i = 1, 2).$

This asserts that the witnesses are equally reliable whether A is true or false. Again it is easy to imagine circumstances where this is not true. In some cultures there is a tendency for witnesses to say what they think will please the listener. So if A is the event "the airport is near," veracity is more likely when A is true than when it is false. Consequently one cannot be sure that $p(a_i \mid A)$ and $p(\bar{a}_i \mid \bar{A})$ are both p_i.

The Bishop certainly did not recognize the distinction, as have many writers after him. The Bayesian approach does not demand the equality: it merely forces one to recognize that two types of veracity are possible.

Applied to the Bishop's problem, the rector's approach forces one to consider one's initial belief in the event, the nature of the dependence between the witnesses, and the two forms of reliability that arise. We suggest that, on reflection, it will be admitted that all three features are relevant to the final answer. Even if the independencies and the equalities of the reliabilities are admitted, as the Bishop and the modern equivalent tacitly do, the result is still different from the Bishop's. It is of interest to enquire when they are equal. Equating (2) and $1 - (1 - p_1)(1 - p_2)$ easily gives after a little algebra the condition that

$$(1 - \pi) = p_1 p_2 \pi + (1 - p_1)(1 - p_2)(1 - \pi).$$

The righthand side is $p(a_1, a_2)$, the unconditional probability that both witnesses assert A is true, so that the Bishop and rector only agree (under assumptions II and III) if

$$p(\bar{A}) = p(a_1, a_2).$$

In words, the probability that the event is false has to be equal to the probability that both witnesses assert its truth. This is surely unreasonable.

I put it to the readership: my challenge has survived, probability does do better. Let us support the rector of Tunbridge Wells and not the Bishop of Bath and Wells: let us favor truth and not the establishment. (Bayes was a minister in the unestablished church.)

ACKNOWLEDGMENTS

I am grateful to Richard E. Barlow for useful comments on a first version of this tale, to Sir Richard Eggleston for illuminating discussions on the legal problems with two witnesses, and to Glenn Shafer for drawing my attention to the Bishop's paper.

Comment

David J. Spiegelhalter

It is fairly predictable that I should agree wholeheartedly with Professor Lindley's lucid justification of probability as the correct paradigm for handling uncertainty in expert systems (but how strange it is to see him cast in the role of defender of orthodoxy!). In particular, his emphasis on remembering the background evidence H is crucial to avoid any conception that there is a single "true" probability of an event, and the frequent references to the operational meaning of probability gives a practical as well as a theoretical justification. However, playing the devil's advocate, I see two main reasons why the artificial intelligence community may not be convinced by the argument.

Firstly, he turns all statements expressing uncertainty into expressions of probability concerning (at least theoretically) verifiable events, whereas many constructors of expert systems would prefer to keep their propositions deliberately imprecisely defined in order to look more like human reasoning, and do not provide an operational means of verification. Secondly, even if verifiable events *are* being considered, the scoring rule argument presumes a certain type of evaluation procedure which many might claim was rarely appropriate, since the criteria for the "success" of an expert system may only require a very coarse handling of uncertainty.

Nevertheless, the theoretical arguments concerning optimality and coherence are only one weapon in the armoury. Pearl (1986b), in a recent strong advocacy of probability, uses no normative criteria but concentrates on the power of the theory in adequately modeling complex evidential reasoning, and I feel, in the end, it will be the intuitive appeal and flexibility of probabilistic reasoning that will change the current climate.

Professor Shafer's historical perspective puts the current discussion in an appropriate context, and emphasizes that many of the issues raised in expert system research are by no means novel. The interest in belief function methodology is understandable, as it appears to provide a means of avoiding full subjective assessment of a joint probability distribution, and—by formulating "uncertainty" in terms of reliability of evidence—it seems to attach uncertainty directly to the *rule* rather than the consequences of the rule. All this is very attractive, but users of the methodology also have to take on board a rule of combination that can lead to somewhat unintuitive results (Zadeh, 1986), problems in providing an operational interpretation of the numerical inputs and outputs, and a considerable computational burden.

Shafer does show how computationally efficient schemes are available on simple trees, but this is an extremely restrictive class of model, excluding both multiple causes of the same event, and an element being a member of two taxonomic hierarchies (for example, "gallstones" may also be part of a "dyspepsia" taxonomy). In contrast, efficient probabilistic schemes are now being devised for general graphical structures.

This still leaves the ability of belief functions to deal with "unknown" or "unknowable" probabilities. From a historical point of view, it would be easy to slip into the "likelihood versus Bayesian" debate at this point. But I believe the objective of constructing expert systems enables us to avoid such arguments. In such technological applications, there is real understanding of the problem to be exploited, and from a purely pragmatic point of view, unknown probabilities just do not occur—an assessment can always be obtained by careful questioning. Of course, the subject may not feel too confident in his assessment, and will not be able to list a set of independent sources of evidence for his opinion. But the opinion is there and can be used, although, as Professor Lindley emphasizes, in certain circumstances the imprecision may be relevant. As Professor Shafer points out; explanation of a system's conclusions may be provided at many levels, and probability judgments that have not been "constructed" on specified evidence can, if necessary, be identified. Provided a system's predictive performance is being monitored by scoring rules, it seems quite reasonable in a medical area to exploit "informed guesses" rather than rely on a legalistic paradigm that models unreliable "witnesses."

ADDITIONAL REFERENCE

ZADEH, L. (1986). A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combinations. *Artificial Intelligence Mag.* 85–90.

Rejoinder

Glenn Shafer

Watson and Dempster and Kong underline the point that belief functions are a form of probability. I can only say that I agree wholeheartedly.

I still have some bones to pick, on the other hand, with Spiegelhalter and Lindley.

Spiegelhalter's comments on the computational situation are misleading. He suggests that computationally efficient schemes for belief functions are available only for a very restrictive class of models, whereas efficient Bayesian schemes "are now being devised" for very general models. In fact, most Bayesian computational schemes have belief-function generalizations. It is true that the Bayesian special cases usually require less computation: Bayesian models require more complicated inputs than belief-function models, and there is less need for computation when you begin with more information. But the trade-off between complexity of input and complexity of computation differs from case to case, and belief-function computations are manageable in a greater variety of situations than Spiegelhalter suggests.

In my article, I discussed Judea Pearl's work on propagating Bayesian belief functions in trees, and I noted that Pearl's Bayesian scheme is a special case of a general scheme for propagating belief functions in trees. This general scheme has now been described in some detail by Shafer, Shenoy, and Mellouli (1986). In recent unpublished work, Pearl and Spiegelhalter have made progress in dealing with Bayesian networks that are not trees. Similar work is also underway for belief functions, with the most important contribution so far being Augustine Kong's dissertation at Harvard (Kong, 1986). In the last chapter of this dissertation, Kong shows how the belief-function scheme of Shafer and Logan (1985) can be adapted to handle multiple diseases with no additional computational cost.

Spiegelhalter cites Zadeh in support of the view that Dempster's rule of combination can lead to unintuitive results. For a reply to Zadeh's arguments, see Shafer (1986a).

The Bishop of Bath and Wells whose work on probability Lindley discusses was named George Hooper. Hooper actually became a bishop only in 1703, long after his work on probability was published. Details about Hooper's life and work are given by Grier (1981). Hooper gave two rules for combining testimony, a rule for concurrent testimony and a rule for successive testimony. I have discussed these rules and their Bayesian counterparts elsewhere (Shafer, 1978, 1986c).

Hooper's rules were widely admired in the 18th century; they appear, for example, in Diderot's *Encyclopedie*. The Bayesian analysis that Lindley reviews, together with a corresponding analysis for the case of successive testimony, displaced Hooper's rules in the early 19th century (see Shafer, 1978). But this Bayesian account of "the probability of testimony" quickly became a laughingstock. It was roundly and justly denounced both by logicians critical of probability, such as John Stuart Mill, and by probabilists who preferred a frequentist interpretation, such as Antoine-Augustin Cournot.

The theory of belief functions does not require us to go back to Hooper's rules. Instead it provides a framework that includes both Hooper's analyses and the Bayesian analyses as special cases, along with many intermediate possibilities. The virtue of this flexibility is that we can tailor our analysis to our actual evidence. If we have significant prior evidence, we can use it. If we have evidence for causal dependence between the witnesses, we can use it. If we have instead evidence for dependence in our uncertainties about the witnesses, we can use it. By relating the numbers we offer to actual evidence in this way, we can hope to escape the ridicule that so wounded subjective probability in the 19th century.

ADDITIONAL REFERENCES

GRIER, B. (1981). George Hooper and the early theory of testimony. Dept. Psychology, Northern Illinois Univ.

KONG, A. (1986). Multivariate belief functions and graphical models. Ph.D. dissertation, Dept. Statistics, Harvard Univ.

SHAFER, G. (1978). Nonadditive probabilities in the work of Bernoulli and Lambert. *Arch. Hist. Exact Sci.* **19** 309–370.

SHAFER, G. (1986c). The combination of evidence. *Internat. J. Intelligent Systems* **1** 155–179.

SHAFER, G., SHENOY, P. and MELLOULI, K. (1986). Propagating belief functions in qualitative Markov trees. Working paper no. 186, School of Business, Univ. Kansas.

Rejoinder

Dennis V. Lindley

I find myself in general agreement with the contributions of Watson and Spiegelhalter. Watson is right when he says we do not have to accept Savage's axioms. But it is desirable to have an axiom system to support one's calculations and the lack of them must count against the alternatives to probability. Spiegelhalter is right when he says that ultimately it's the appeal of probability that matters: people will see that it makes good sense. Just as with Euclidean geometry, it is the operational aspect that counts, rather than Euclid. Watson queries the existence of the Great Scorer. I do not think it matters because one would wish to behave in such a way that one could not be exposed by his or her arrival. I would regard it as a serious proposal to pay meteorologists, or even medical doctors, according to their scores.

Whilst I find myself in dispute with Shafer, his arguments command respect and are not easily refuted. He contends that the axioms depend on conditional probability and expected utility, rather than that these depend on the axioms. While it is true that historically the concepts pre-date any axiom system, Savage introduced the axioms in order to justify a system, classical statistics, that denies conditional probability (of a hypothesis) and does not admit expected utility (with an expectation over unknowns); and he was much surprised when the axioms destroyed that system.

The scoring-rule argument works for almost every rule and does not depend on 0 or 1 as Shafer suggests. The preferences in Bayesian decision analysis are not necessarily sharp. If d_1 has expected utility 10.927 and d_2 10.926, then d_1 is preferred only slightly to d_2. The analysis is designed to select *an* act because only one act is typically possible.

Shafer also raises the issue of constructive probability. It is difficult, having experienced A_1, to think of probabilities for A_1 if only because probability describes uncertainty and A_1 is no longer uncertain. My response is that we should try to develop methods that

would help people to do this. If these all fail, then it will be necessary to think afresh. But forensic scientists, finding it necessary to think about probabilities for clothing stains (for example) have been able to assess them.

There remains the contribution by Dempster and Kong. They really throw the book at me and I am at a loss how to react. Certainly no response within the limits that the editor is likely to impose on me could be adequate. It is therefore perhaps best to remain silent except for one remark that touches on a point raised by others. One reason that I reject belief functions is that, at every stage, they are more complicated than probability—and that is hard enough, as Watson points out in connection with Schum's work. They involve more assessments and harder calculations. Furthermore, in my experience it is never necessary to extend the probabilistic argument in the way the theory of belief functions suggest. For example, if imprecision about a probability is relevant, then probability theory will require its assessment within its own calculus. Dempster and Kong reinforce this point when they take several paragraphs to solve the simple decision problem in their Table 1.

In conclusion may I thank those responsible for arranging the conference that led to these papers, and the editors for encouraging them to appear. I hope that readers will feel that the issues we address are important, both in theory and practice. If any readers feel they can meet the challenge it would be interesting to hear from them.

Rejoinder

David J. Spiegelhalter

By concentrating on applications, I appear to have escaped lightly in the discussion. Dr. Watson pointed out the multitude of criteria that could be used for evaluation of aids to clinical decision-making. Some order can be introduced by classifying all criteria according to whether they concern the system as *decision-maker* or as *aid*, and whether they are measures of *process* or *outcome*. Thus "internal coherence" is a process measure of the system as decision-maker, "comparison with experts" is an outcome measure as a decision-maker, "user satisfaction" is a process measure as an aid, and "effect on patients' health" is an outcome measure as an aid.

Professor Lindley was concerned about my interpretation of "uncertainty about a probability." Perhaps this phrase should not be used, since it does not differentiate between doubt in one's current beliefs due to *imprecision* in the probability assessments on which that belief is based, and sensitivity in that belief due to *ignorance* of potential future evidence. As evidence accumulates, the imprecision will generally increase as one gets into an increasingly narrow area of experience, but ignorance will be reduced. One's "point" current belief can therefore be thought of as the mean of two second-order distributions, representing what that belief might be now, and what it may become in the future.

Professor Shafer offers a vision of creative systems that can generate arguments in novel situations. He is correct that I, and my clinical colleagues, view expert systems in a much more limited sense, often having very little to do with the tenets of artificial intelligence, although exploiting their programming environments. I remain confident that probability is the appropriate tool in this area, and recent developments in strict probabilistic reasoning using local computations in general causal networks (Lauritzen and Spiegelhalter, 1987) overcome many technical problems. The parallels raised by Professor Shafer between probability/belief-function and expert-system/artificial intelligence contrasts are intriguing.

Both Professors Shafer and Dempster mention upper and lower expected losses from belief functions, which I find rather confusing. Are belief intervals to be interpreted as upper and lower probabilities or not? Suppose we adopt Dempster's decision theoretic structure after hearing "Slippery Fred's" evidence. Then $\{\mu\}_{Bel}$ obey $.8 \leq P(\text{slippery}) \leq 1.0$, which—from Shafer's original equation (3)—can easily be shown to impose the constraint $q \geq \max\{0, 4(1-2p)/(4-3p)\}$. If $p \geq \frac{1}{2}$, then $\{\mu\}_{Bel}$ is equivalent to $0 \leq q \leq 1$, which does not appear too unreasonable. However, the implicit constraints become much stronger after a crank of the rule-of-combination having seen the thermometer. Let us denote by r the probability the thermometer is right even if it is not working properly. To obtain coherently $\{\mu\}_{Bel} = .04 \leq P(\text{slippery}) \leq .05$, we require for, say $p = \frac{1}{2}$, that $(3 + 97r)/(123 - 23r) \leq$

$q \leq (23 + 77r)/(118 - 18r)$, which is a fairly narrow band around $q = r$. Thus, far from having no basis for specifying p, q and r, very strong constraints appear to have to be made in order for the decision-theoretic scheme to be coherent, were the probabilities available.

Shafer-Dempster belief intervals are widely interpreted as upper and lower probabilities in the expert-system world, but I had always thought this was an error. Now I admit to being confused.

ADDITIONAL REFERENCE

LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1987). Fast manipulations of probabilities with local representations—with applications to expert systems. Technical Report 87-7, Aalborg University Centre.