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 $0 \le p(A \mid H) \le 1$ 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 $p(A_1 \mid H) = ½$ or $p(A_2 \mid A_1 \cap H) = ⅔$ or $p(A_1 \cap A_2 \mid H) = ¼$, 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 $p(D_i \mid \overline{\Delta})$, 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 $\text{Bel}(A) = a$, $\text{Bel}(\bar{A}) = b$ 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 $(1 \le i \le n)$. 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 $(1 \le i \le n$ 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

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

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.