In Defense of Probability

Peter Cheeseman
SRI International
333 Ravenswood Ave., Menlo Park, California 94025

Abstract

In this paper, it is argued that probability theory, when used correctly, is sufficient for the task of reasoning under uncertainty. Since numerous authors have rejected probability as inadequate for various reasons, the bulk of the paper is aimed at refuting these claims and indicating the scources of error. In particular, the definition of probability as a measure of belief rather than a frequency ratio is advocated, since a frequency interpretation of probability drastically restricts the domain of applicability. Other sources of error include the confusion between relative and absolute probability, the distinction between probability and the uncertainty of that probability. Also, the interaction of logic and probability is discusses and it is argued that many extensions of logic, such as "default logic" are better understood in a probabilistic framework. The main claim of this paper is that the numerous schemes for representing and reasoning about uncertainty that have appeared in the AI literature are unnecessary—probability is all that is needed.

1 Introduction

A glance through any major AI publication shows that an overwhelming proportion of papers are concerned with what might be described as the logical approach to inference and knowledge representation. It now widely accepted that many knowledge representations can be mapped into (first order) predicate calculus, and the corresponding inference procedures can be reduced to a type of controlled logical deduction. However, examples of human reasoning (judgements) are full of such terms as "probably", "most", "usually" etc., showing that many patterns of human reasoning are *not* logical in form, but intrinsically probabilistic.

The claim that many patterns of human reasoning are probabilistic does not mean that the underlying "logic" of such patterns cannot be axiomatized. On the contrary, a basis for such an axiomatization is given in section 3. The claim is that when such an exercise is performed, the resulting patterns of inference are different in form from those found in analogous logical deductions. A characteristic difference is that in probabilistic inference all the relevant inference paths ("proofs") connecting the evidence to the hypothesis of interest must be examined and "combined", while in logic it is sufficient to establish a single path between the axioms and the theorem of interest. Also, the output is different, the former includes at least one numerical measure, the latter simply true or false.

Unfortunately, the logical style of reasoning is so prevalent in AI that many have attempted to force intrinsically probabilistic situations into a logical straight-jacket with predictable limited success. Two conspicuous examples of this are "Default Logic" [19] and "Non-Monotonic Logic" [15] discussed in more detail below. These methods are appropriate for dealing with some situations where limited knowledge is available. The same cannot be said for those who invent new theories for reasoning under uncertainty, such as "Certainty Factors", "Schafer/Dempster Theory", "Confirmation Theory", "Fuzzy Logic", "Endorsements" etc.

These theories will be shown below to be at best unnecessary and at worst misleading (not to mention confusing to the poor novice faced with so many possibilities). Each one is an attempt to circumvent some perceived difficulty of probability theory, but as shown below these difficulties exist only in the minds of their inventors. However, these supposed difficulties are common misconceptions of probability, generally springing from the inadequate frequency interpretation. A major aim of this paper is to put forward the older view (Bayes, Laplace etc.), that probability is a measure of belief in a proposition given particular evidence. This definition avoids the difficulties associated with the frequency definition and answers the objections of those who felt compelled to invent new theories.

An analogy can be draw between the situation in AI in the late 1970s, where Pat Hayes, in a paper entitled "In Defence of Logic" [10], found it necessary to take a broadside at the proliferation of new representation languages (with associated inference procedures) that proported to solve difficulties with the logical approach. He showed that far from being "nonlogical" it is possible to cast such languages into an equivalent logical form, and by doing so provide a clear semantics. In addition, he pointed out the obvious but unpopular fact that logic has been around for

a long time and has a considerable body of research and experience that no new theory can match. Similarly today we have a set of new theories for dealing with uncertainty, despite the fact that probability theory has been around for three centuries and, as shown below, is sufficient for the task.

Any text on probability presents a formal calculus for manipulating probabilities according to a consistant set of axioms. Many disputes concerning probability are centered on the *interpretation* of the terms in the formal system, since an interpretation (model theory) is necessary if the theory is to be applied. Others dispute that the formal axioms under any interpretation really capture their intuitive expectations for uncertain inference. This paper argues for a particular interpretation of the probability formalism and that the result is sufficient for all uncertain inference in AI. Since Bayes' theorem is integral to the use of probabilities the terms Bayesian and probabilistic are used interchangeably.

2 Some Misconceptions of Probability

This section discusses and hopefully exorcises the most common misconceptions of probability.

2.1 Probability is a Frequency Ratio

Rather than give an historical account of the different theories (interpretations) that have been applied to probability (e.g. [8]), the following definition is put forward as one that withstands all previous criticisms:

The (conditional) probability of a proposition given particular evidence is a real number between zero and one, that is a measure of an entity's belief in that proposition, given the evidence

Several corollaries follow directly from this definition. Firstly, there is no such thing as *the* probability of a proposition, since the probability value depends on the evidence used to derive it. This implies that if new evidence is utilized, the probability value assigned to the proposition will generally change. The only exception to this variability is when the probability is zero or one, because then there is no longer any uncertainty and further evidence makes no difference. Secondly, different observers with different evidence (information) will assign different probabilities. There is no contradiction inherent in this—the apparent contradiction comes from the idea that every proposition has a unique probability. A third consequence of the above definition is that probabilities are inherently subjective in the sense that the value depends on the believer's information, but they are objective in the sense that the same (ideal) believers should arrive at the same value given the same information.

This definition differs sharply from the still commonly held frequency definition of probability:

The probability of an event (hypothesis) is the ratio of the number of occurrences (n) in which the event is true to the total number of such occurrences (m)

This definition has some immediate problems that many other critics have noted. For a start, this definition restricts probability to domains where repeated experiments (e.g., sampling) are possible, or at least conceivable. Also, the probability of an event under this definition is undefined if there are no prior examples ($m = 0$)—thus limiting its usefulness. Even worse are cases where, for example, there has been one success ($n = 1$) and one trial ($m = 1$), giving a probability of one for the next event!—that is on the basis of a single trial the probability of the next event is known with certainty. In most circumstances this is nonsense—those who defend the frequency ratio definition escape into "the law of large numbers" which essentially says that given a large number of (repeatable) trials, the true probability lies within given error bounds with high probability. This restriction bans small sample cases from the realm of probabilistic ("frequency") analysis, but works well for the large sample case. Given the success of the frequency definition in areas where it is applicable, it is fortunate that there is a strong connection between the measure of belief definition of probability and the frequency ratio definition. It has been shown by Jaynes [12] that under certain conditions (e.g., repeatable trials) the *expectation* of the frequency ratio is necessarily equal to the probability. However the measure of belief definition applies to the small sample case as well.

Philosophers have been arguing the "correct" definition of probability for centuries, and some have defined up to five different meanings for probability [8], including: "statistical probability" (i.e., the frequency ratio definition); "probability = propensity" (i.e., probability used for prediction); "logical probability" (i.e., the degree of confirmation of a hypothesis based on logical analysis) and "subjective probability". The measure of belief definition subsumes all these supposedly different concepts. For example, the probability of set membership ("the probability of A being a B") and the probability of future events ("the probability that H will happen, given E") are not different kinds of probability but just the observer's belief in the corresponding proposition given the evidence. Similarly, it makes no difference to the belief in a proposition whether the probability is the result of logical analysis (e.g., the probability of a number being prime) or the result of empirical observations (e.g., the probability of surviving a car accident). The philosophical distinctions and alternative definitions of probability obscure rather than enlighten understanding of probability.

2.2 Bayesian Analysis Requires Vast Amounts of Data

This particular fallacy has appeared so often that its truth is rarely questioned. The reason for this fallacy follows directly from the frequency ratio definition. This says

that the probability of a proposition, such as "This patient has a particular infection given his particular set of symptoms", can be computed from the number of patients that have previously exhibited that combination of symptoms. Clearly, in practice, the set of previous patients with a particular combination of symptoms is going to be very small or zero, so by the frequency definition of probability this conditional probability cannot be computed.

In anything but the most trivial cases, the basic problem that the Bayesian (or any other) approach must deal with is that the available information is not sufficient to determine any particular conditional probability, as in the above example. That is, the probability space associated with a particular problem is usually highly under-constrained by the known probabilities, so it is impossible to calculate directly any particular conditional probability [3]. The normal way around this difficulty is to make additional assumptions that supply the missing constraints. The most common assumption is "conditional independence", as advocated in [2], [7] and [16]. The conditional independence assumption has been generalized by Lemmer and Barth [14] to include conditional independence between groups of propositions, and all these forms of independence assumptions are subsumed under the maximum entropy assumption [3], [13].

The use of the maximum entropy assumption (or its specializations) raises the question of its validity. For maximum entropy, it has been shown that the probability generated is the one which has the maximum number of possible worlds consistent with the known information, and in this sense is the "best" value [12]. In some circumstances, such as occur in statistical mechanics, the probability of the system being in a state with entropy significantly less than the maximum is vanishingly small. Maximum entropy implies that if a non maximum value is chosen, then more information is being assumed than was available—i.e., the maximum entropy gives the "least commitment" value or the one that distributes the uncertainty as evenly as possible over the set of possibilities. Conditional independence (the most common form of maximum entropy) is not just another assumption, as implied in [18], it is the only consistant assumption that can be made in the absence of any information about possible interactions. However, these desirable properties do *not* mean that information is being generated out of nothing.

What maximum entropy is doing is providing a neutral background against which any systematic (non-random) patterns can be observed. That is, if the current (probabilistic) information is incomplete, the predictions using this information and maximum entropy will differ significantly from future observations. When such differences are detected the response should not be to throw out maximum entropy (as many authors advocate), but to utilize this additional information. Maximum entropy *is* making stronger predictions than the current information warrants because it is assuming the current information is complete. However, without this prediction it is difficult to detect if the current information is incomplete, and thus difficult to discover new information. Also many decision making situations require probability values, so that some additional principle, such as maximum entropy, is necessary in these circumstances to select point values even when the value is poorly known. This justification for the maximum entropy assumption is really the old problem associated with the use of prior probabilities in the Bayesian approach as discussed in the next subsection.

2.3 Prior Probabilities Assume more Information than Given

This statement appears in numerous AI publications, especially those expounding the Schafer-Dempster approach to uncertainty. For example:

"Bayesians might attempt to represent ignorance by a function assigning 0.25 to each of the four possibilities, assuming no prior information. ...such a function would imply more information given by the evidence than is truly the case."—[1].

"A Likelihood represented by a point probability is usually an over statement of what is actually known, distorting the available precision".—[9]

Yet those that make these claims fail to show a single unfortunate consequence that follows from this supposed assumed information. To illustrate the situation, consider the following examples. In the first example, you are told there is a normal dice and asked what probability you would assign to the next throw yielding a "6". The Maximum Entropy answer is $\frac{1}{6}$, since this distributes the uncertainty as evenly as possible over the set of possibilities. In the next example you are told there is a *loaded* dice (but not which numeral is favoured) and are asked what is the probability of a "6". Again the answer representing your state of knowledge is to assign $\frac{1}{6}$. The difference between these two situations is that in the first example your knowledge of dice mechanics and symmetry implies that after having seen the outcome of many throws you do not expect to change your state of knowledge (i.e., the probability assignment). However, for the loaded dice, you do expect the probabilities assigned to the different faces to change as a result of further trials.

Those who reject the Maximum Entropy approach argue that in the second example, the initial assignment of $\frac{1}{6}$ was assuming more than you know because after many trials you ended up with a different assignment (i.e., the initial assignment was incorrect). This objection arises from the mistaken idea that there is such a thing a *the* probability of a proposition instead of the idea that probability represents a state of knowledge. Of course the probability assignment to a proposition will change as more information is gained without inconsistency with previous assignments. The idea that there is a unique probability associated with a particular proposition comes from situations where all observers have the same information (e.g., physics), and so they all have the same measure of belief (assuming ideal observers).

However, not just any prior probabilities will do. If non-

equal priors are chosen, this implies that you have information about the different possibilities. Put another way, the equal prior assignment gives a neutral background against which deviations from your state of maximum uncertainty can be detected. It is because it is *not* assuming more information than given that the maximum entropy assignment is used. Looking at the example in reverse, if someone assigns equal probabilities to a set of possible outcomes, they are telling you they are completely ignorant about the next outcome (apart from how many possibilities there are). Note that in these two examples, our knowledge about our knowledge of the probabilities (i.e., the probability distribution of the probability) is the main difference.

A more subtle criticism of the use of the principle of indifference that has historically plagued probability theory is illustrated by the following example. Assume there are five "concepts" (a, b, c, d, e), then the principle of indifference will assign prior probability $\frac{1}{5}$ to each. If you are now told that concept a is actually f or g, then you should reassign probabilities of $\frac{1}{6}$ to each of (f, g, b, c, d, e). This apparent arbitrariness of the prior probabilities through regrouping and relabelling is put forward as a reason for rejecting use of priors at all. The arbitrariness of the probability assignment only arises in this example because the "concepts" are meaningless, so any grouping is just as meaningless as any other. If the problem is undefined, probability theory (or any other theory) cannot say anything useful. However, as soon as the concepts are identified with possibilities in the real world, the arbitrariness disappears. When each possibility in a problem corresponds to a physically realizable possibility, we no longer have the freedom to count arbitrary groupings of such outcomes as if they are a separate outcome—i.e., we can no longer arbitrarily redefine the problem [11].

For example, consider the famous problem known as Bertrand's Paradox. In this "paradox" we are required to draw lines "at random" that intersect a circle, and wish to know the probability that the length of a chord of such a line is longer than a side of an inscribed equilateral triangle. There appears to be different answers depending how "equally possible" situations are defined. Three possibilities are to assign uniform probability density to: (a) the distance between the centers of the chord and circle, (b) the angle the chord makes with the center, and (c) the center of the chord within the circle; each possibility giving a different answer. Jaynes [11] has shown that only (a) is consistent with the requirement that the answer be invariant under infinitesimal translations and rotations—an obvious requirement coming from our understanding of the physical set-up.

Another example of the invariance argument leading to a definite prior probability assignment is to consider the probability of finding a ship within a particular square mile somewhere in the Atlantic. If this is the only information available, then an invariance argument requires assigning equal probability to equal areas, in agreement with intuition. Since the Atlantic is roughly diamond shaped, this means that the probability of finding the ship at an equatorial latitude is higher than at a polar latitude. If the ship is instructed to move to a particular latitude, but interference completely scrambles our reception of *which* latitude, then after the ship has had time to move, our knowledge is represented by assigning uniform probability to each latitude. This new assignment, based on the new "information" leads to a new probability distribution in which the probability of finding the ship near the equator is now less than near the poles. This example shows that in real problems we cannot arbitrarily assign equal prior probabilities to any dimension or combination of possibilities because to do so implies unequal assignments on other dimensions. In practice, our rich domian background knowledge usually leads to non-uniform priors, even though we may be uncertain of their values. In complex cases, there is no substitute for a careful analysis of each problem to find what the appropriate priors for that problem are.

2.4 Numbers are not Necessary

An unfortunate tendency in AI is to rediscover the wheel but call it something else so it then becomes a "new" paradigm. An example is found in Cohen and Grinberg [5], who shows, convincingly, that in many situations it is necessary to keep track of the evidence that was used to arrive at a particular (conditional probability) judgment, so that the judgment can be revised if new evidence requires it. Their work calls attention to the fact that a computed probability number is just a *summarization* of all the evidence that was used to derive it (for convenience in decision making), and so does not contain information about its origin. However, it still a *conditional* probability and the conditions of its derivation can also be important. This utilization of probabilistic dependencies is unfortunately given the new name "endorsements", and from its success in explaining observed judgements under uncertainty, the conclusion is reached that numbers are not necessary for such judgements at all!

This conclusion has validity in restricted circumstances—in particular, it is possible to construct a theory of *relative* probabilities (e.g., [8]) that only uses information of the form P_1 is-more-probable-than P_2. Deductions in such a theory do not use numbers and can keep track of their dependencies in a style similar to "endorsements". However, the best that such a theory can say is that "this proposition is the most probable given the evidence"—it cannot indicate any absolute strength in its conclusion. It often happens that the most probable alternative is itself highly unlikely, but non-numeric approaches are unable to express such a result. The bottom line is that judgment under uncertainty *can* be done without using numbers if the user is in a decision making situation where he has only to choose among a set of alternatives. If he has the option of not selecting at all (e.g., because the most likely alternative is still too improbable), then non-numeric approaches are not sufficient.

2.5 More than one Number is Needed to Represent Uncertainty

Many of the alternative theories of uncertainty start with the observation that a single number (a probability value) does not represent all the uncertainty about a proposition—in particular, it does not indicate the accuracy with which the probability value itself is known (i.e., the probability of the probability). Similarly, Schafer [20] distinguishes between uncertainty (roughly a probability) and ignorance (no knowledge of the probability). However, even though one can make these distinctions, basic questions about their utility remain. Ultimately, the utility of any theory of uncertainty comes from the coupling it provides between evidence (information) and decision making (or prediction). A theory of uncertainty is useless without a model theory that indicates how to map evidence into an uncertainty measure and how to use this uncertainty measure to make predictions (or decisions). To decide whether particular distinctions of types of uncertainty are useful or not, we must examine whether they make any difference to the theory's decision making behavior.

The theory of optimal decision making using point probability and utility values is well known. This would seem to imply that a point probability is sufficient to represent uncertainty. However, this theory makes the assumption that the probabilities used in the analysis are known to sufficient accuracy. Probability theory can be extended so that a probability density function is assigned to a sentence instead of a point value, or higher order moments of the density function can be given. However, a result of decision analysis is that *exactly the same decision is reached whether a point value or a density function is used.* This situation is similar to that in mechanics, where a complex body can be replaced by a point mass at the center of gravity to give the same results. However, knowledge of the probability density function is important for sensitivity analysis as in the following example.

If you are given a black box and told that it will put out a string of decimal digits and are asked what is the probability that the first digit will be say 7, the standard principle of indifference answer is (.1). If, later, after seeing 10,000 digits of which 1000 were 0, 1000 were 1, etc., in no noticeable order, you are again asked to give the probability that the next digit will be 7, you will still answer (.1). This last answer, by standard information theory, implies that all the evidence gave no information whatever—you are still as uncertain about the probability of the next event as you were before seeing the "evidence". However, something has clearly changed between these two cases—it is the expectation that further evidence will significantly change our probability assignment (i.e., will provide real information). This changed expectation can be captured as a standard deviation about the probability value which is very large initially and becomes quite small (about .003) after seeing the 10,000 trials.

This example implies that if you are in a decision making (or prediction) situation *and obtaining more evidence is not an option* then a single number (the probability) is a sufficient representation of your uncertainty. However, if obtaining more information is a possible option, then a measure of how informative this information is likely to be (e.g., the standard deviation) is required. Thus, how many numbers are needed to represent uncertainty depends on the questions you are trying to answer with the uncertainty representation. To always calculate two numbers, as done in the Schafer-Dempster approach, is often overkill, and in some cases, under-kill.

2.6 The Bayesian Approach Doesn't Work—So Here is a New Scheme!

As described above, various authors have found fault with Bayesian probability, and their response has been to invent new representations and inference procedures that purport to remove particular difficulties. However, these *ad hoc* theories do not have a well established model theory to show how to go from real data to the internal uncertainty representation and then to map the final uncertainty representation into a well defined decision theory. Because of this missing interpretive framework, and because of their rejection of prior probabilities, they have produced all sorts of misleading conclusions. The following examples are illustrative:

Example 1

"Translated to the notation of conditional probability, this rule $(s_1, s_2, s_3 => h1)$ *seems to say* $P(h_1|s_1, s_2, s_3) = 0.7$ *where* h_1 *is the hypothesis that the organism is* Streptococcus, s_1 *is the observation that the organism is gram-positive,* s_2 *that it is a coccus, and* s_3 *that it grows in chains. Questioning of the expert gradually reveals, however, that despite the apparent similarity to a statement regarding a conditional probability, the number 0.7 differs significantly from a probability. The expert may well agree that* $P(h_1|s_1, s_2, s_3) = 0.7$, *but he becomes uneasy when he attempts to follow the logical conclusion that therefore* $P(not\ h_1|s_1, s_2, s_3) = 0.3$. *He claims that the three observations are evidence (to degree 0.7)* in favor *of the conclusion that the organism is a* Streptococcus *and should not be construed as evidence (to degree 0.3)* against Streptococcus. *We shall refer to this problem as Paradox 1 ..."*—[1]

The authors then conclude, on the basis of this "paradox", that one should gather and evaluate separately the evidence for an hypothesis and the evidence against it. This spurious argument only arises by ignoring prior probabilities and the consequent misrepresentation of the situation to the expert. The prior probability of an infection being caused by a particular bacterium is low, for the sake of argument we will assume it to be .01. After seeing the evidence (s_1, s_2, s_3) the expert is willing to update his probability (i.e., his belief) to 0.7. Another way of saying the same thing is that the probability (belief) in the negation of the hypothesis (that the organism is not *Streptococcus*) drops from a prior of .99 to .3. Thus, either way, the evi-

dence is being used to strongly support the hypothesis, and not (as claimed above) being construed as evidence against the hypothesis. Given the misrepresentation of the situation, it is not surprising that the expert felt uneasy with the way his evidence was being used.

This example shows the danger of ignoring prior probabilities when dealing with uncertainty, and also shows its considerable advantages when used properly. As a basic principle of inference one should use whatever information is available, and this includes prior probabilities. Perhaps the main sources of opposition to the use of prior probabilities is that they are subjective estimates of the expert, and it has been shown (e.g., [21]) that people are not very good at estimating probabilities. However, the expert does not necessarily have to supply the priors—once the hypothesis space is defined, the equiprobable assignment (i.e., the principle of indifference) or relevant data can be used instead. If the expert has prior information (e.g., some infections have higher prior probabilities than others) then he should give this information to the system (in the form of non-uniform priors), because to not do so is to ignore useful information. The fact that these subjective estimates will be poorly known is no excuse for not using them. Fortunately, the final probability values calculated on the basis of extensive new information are not very sensitive to the exact value of the priors.

Example 2 (Fuzzy Sets, Fuzzy and Possibilistic Logic)

"., it is a standard practice to rely almost entirely on the techniques provided by probability theory and statistics, especially in applications relating to parameter estimation, hypothesis testing and system identification. It can be argued, however, as we do in the present paper, that such techniques cannot cope effectively with those problems in which the softness of data is nonstatistical in nature–in the sense that it relates, in the main, to the presence of fuzzy sets rather than random measurement errors or data variability."—Zadeh, [23]

This quote captures some of the motivation that underlies fuzzy sets (and their further development—fuzzy and possibility logic)—namely, the fallacy that probabilities are necessarily frequencies. The concept of vague set boundaries has no obvious frequency interpretation, so Zadeh invented fuzzy sets to capture this vagueness idea. Actually, there *is* a probabilistic (degree of belief) model for vague sets that also supplies a computable quantitative measure for the "best" (most informative) vague classification. Normally, a set is defined by a criterion that distinguishes members from non-members without allowing for partial membership. This concept of sets has been widely critisized by philosophers (e.g., Wittgenstein) largely because sets in common use do not have sharp boundaries. The alternative probabilistic model is to define a set by a "prototype" and expectations of divergence from the prototypical features shown by members of the set. That is each object has a numeric "degree of membership" given by how likely it is that the observed features would have occurred given that it is a member of that set. The best classification of the object is that which maximizes the probabilistic "similarity" measure, and it is quite possible for an object to be so dissimilar from any prototype that it forms a new set. Also, an object can be simultaneously probabilistically similar to more than one set. The underlying theory of probabilistic set membership is given in [22].

Other errors found in the AI literature include the notion that the final conditional probability value of a proposition depends on the order in which the evidence is introduced [20]; that hypotheses, such as the possible diseases a patient might have, are mutually exclusive [2]; that a piece of evidence whose conditional probability differs considerably from that of other evidence should be rejected [17] (instead of rejecting the corresponding hypothesis); etc.

2.7 Summary of Conceptual Confusions

The authors that reject probabilities as a formalism for dealing with uncertainty in AI are usually a victim of one or more of the following confusions.

- Relative versus Absolute Probabilities—To decide the most probable of a set of hypotheses is only a relative evaluation sufficient for some tasks, but decision analysis requires (absolute) conditional probability values.
- Separation of Probability and Utility—The importance (utility) of an hypothesis is often confused with its probability, since both are required for decision making.
- Probabilities are a Measure of Belief in a Proposition—This definition does not require a frequency interpretation, but applies to any well defined situation and summarizes all the evidence for that proposition.
- Probability versus Uncertainty about the Probability—The (conditional) probability P of a proposition is the user's measure of belief in that proposition, but information about the accuracy of P is fully expressed by a probability density function over P.
- Probability is not a special case of Logic—Probabilistic reasoning is often cast incorrectly in a logical form, as discusses in Section 3.
- Prior Probabilities should be used—Failure to use prior probabilities can lead to erroneous conclusions, especially when there is a large number of possibilities.
- Ambiguous Probabilities—If they occur, it is a sign that the problem is not fully defined, not that probability theory is inadequate.

3 Logic and Probability

Formally, probability can be regarded as a generalization of predicate calculus, where instead of the truth value of a

formula given the evidence (context) having only the values 0 (false) or 1 (true), it is generalized to a real number between 0 and 1. This generalization can be achieved by creating new propositions of the form "The probability of F is X", where F is an arbitary well formed formula in predicate calculus. Once it has been accepted that:

- The generalized truth value (degree of plausibility) of a formula can be represented by a real number.
- The extremes of this scale must be compatible with logic.
- An infinitesimal increase in the plausibility of A given new evidence implies an infinitesimal decrease in the plausibility of $\neg A$.
- The plausibility should not be dependent on the order of evaluation.
- Where possible, all the available evidence should be used in evaluating the plausibility.
- Equivalent problems should have the same plausibility.

then it has been shown by [6] that all the degrees of freedom have been used up. That is, all the standard Kolmogorov "axioms" of probability (Addition, Multiplication, Baye's etc.) follow as logic consequences. This implies that fuzzy set theory (which rejects the additivity axiom) is necessarily violating one or more of the above requirements. Any formalism for representing "plausibility" by a real number is either equivalent to probability theory (but perhaps differing in interpretation) or not satisfying the above basic criterion. Even formalisms that do not use a single real number (e.g., [20]) can be captured by higher order probability theory (i.e., probabilities of probabilities etc.). Probability theory provides the basic procedure for computing uncertainties in real situations, but it is often not obvious how to apply it in a particular situation—in particular, the assignment of prior probabilities has historically been the main sources of difficulty.

Misapplications of probability do not usually arise from dispute or uncertainty about the basic axioms but from the way they are interpreted. A purist would insist that the only propositions that can be known with certainty are tautologies (e.g., 7 is a prime number)—any empirical (contingent) proposition can only be known probabilistically, since it is based on induction. However, this insistence forbids the application of logically reasoning to anything about the real world! A reasonable compromise is to treat propositions whose probability is close to 0 or 1 as if they are known with certainty—i.e., thresholding probability values if they are "beyond reasonable doubt". The result of this approximation is to allow logical reasoning instead of probability, because it is usually easier to use. Many of the difficulties experienced by logicians in applying logic to the real world come from a failure to recognize that logic is only an approximation of probability. In particular, "Default Logic" and "Non-Monotonic Logic" are mainly concerned with belief revision when new (logically contradictory) evidence is found. While these logics are suitable for such things as theory completion (when one wishes to avoid, say, having to state all negative facts), they often attempt to force into a logical mold a type of reasoning that is *not* logical in nature. One standard example of default reasoning "All birds fly unless proved otherwise" should be "Most birds fly", which can be used as a piece of evidence in evaluating the probability of the proposition "this bird flies", along with any other relevant evidence.

In probabilistic reasoning, different pieces of evidence are combined together to change the reasoner's measure of belief in a particular proposition—a single line of reasoning, such as a logical proof, is not sufficient. In many cases, there is one piece of evidence (or line of reasoning) that dominates the final result, which is usually given as the "reason" for the result ("if there is smoke, there is fire"). Such reasoning resembles logical reasoning and is often mistaken for it, but its non-logical nature becomes clear when "contradictory" evidence is found. In probability, contradictions do not occur—all the evidence is combined to get a final probability value, so there is no need to reject evidence (although evidence can be used to reject hypotheses). Practical reasoning is usually a complex combination of logical reasoning (discovering consequences, finding the possibilities) and probabilistic reasoning (evaluating the evidence, weighting the possibilities). Likewise, AI should be using both methods where appropriate.

4 Subjective Probabilities

An important topic on the border line between AI (especially expert systems), cognitive science, psychology and philosophy is that of subjective probabilities. Given the above emphasis on probability being a measure of belief, it will come as no surprise that this paper advocates that subjective probabilities should be treated the same as any other probability (such as that from a measurement). However, there are a number of caveats that should be observed, particularly the observation [21] that people are poor estimators of probability—largely because they are victims of many of the misconceptions noted above. Rather than just accepting this situation, as the expert system community seem to, and try to work around it by better interviewing techniques and the like, the view advocated here is that we should aim for *artificial* intelligence. In particular, we should infer expert systems directly from data (as in [4]), rather than filter the same information (badly) through an "expert" and accept whatever numbers he provides. Anyone who has observed an expert giving probability estimates and then discovered he will later provide a completely different estimate, must begin to wonder about the quality of the results of such an expert system.

An artificial intelligence system that reasons under uncertainty will probably use many of the mental techniques that people use. One such technique is random sampling

in the set of possible worlds (i.e., the set of worlds that is consistent with current knowledge) to find the proportion of those worlds in which the predicate of interest is true (i.e., estimate its probability). For example, if a robot is trying to estimate the probability that a person will enter the work area during a particular operation, it should use its current world knowledge to construct (randomly) scenarios in which the event happens and others in which it does not, then using the probability of these different scenarios, to form an estimate of the events' probability. In doing this construction, logic is used extensively. For example, if is unlikely that any person could reach the work area in the time available, then the event is unlikely. When people perform similar hypothetical reasoning, they are often biased by such things as the most recent relevant events—an artificial intelligent system should be designed to avoid such biases and estimate the required probability to the accuracy desired.

An artificially intelligent system for reasoning under uncertainty should be possible based only on the basic "laws" of probability—Baye's theorem, additivity rule, multiplication rule etc., and additional principles, such as "if there is no known causal connection between two events, then assume they are independent (causal closure)" etc. In underconstrained situations, the principle of indifference (or maximum entropy) should be used to obtain the most unbiased value given the available information. No other representation or calculus is necessary for reasoning under uncertainty. This includes the problem of combining evidence from different sources (use Bayes' theorem). Note that use of Bayes' theorem requires that the system keep track of the information that was used in computing conditional probabilities for belief maintenance, in a manner very similar to truth maintenance in logic.

References

[1]Buchanan, B. G., and E. H. Shortliffe, "Rule-Based Expert Systems", Addison-Wesley, p239, 1984.

[2]Charniak, E., "The Bayesian Basis of Common Sense Medical Diagnosis", Proc. National Conf. Artificial Intelligence, Washington, pp 70-73, Aug., 1983.

[3]Cheeseman, P. C., "A Method of Computing Generalized Bayesian Probability Values for Expert Systems", Proc. Eight International Conference on Artificial Intelligence, Karlruhe, Aug. 1983, pp 198-202.

[4]Cheeseman, P. C., "Learning Expert Systems from Data", Proc. Workshop on Principles of Knowledge-Based Systems, Denver, pp 115-122, Dec. 1984.

[5]Cohen, P. R. and Grinberg, M. R., "A Theory of Heuristic Reasoning About Uncertainty", *AI Magazine* Vol. 4 No. 2, Summer 1983, pp 17-24.

[6]Cox, R. T., "Of Inference and Inquiry—An Essay in Inductive Logic", In The Maximum Entropy Formalism, Ed. Levine and Tribus, M.I.T. Press, 1979.

[7]Duda, R. O., P. E. Hart, and Nils Nilsson, "Subjective Bayesian Methods for Rule-Based Inference Systems", AFIPS Conf. Proc., National Computer Conf., Vol 45, New York, pp 1075-1082, 1976.

[8]Fine, T. L., "Theories of Probability", Academic Press Inc., 1973.

[9]Garvey, T. D., J. D. Lowrance, and M. A. Fischler, "An Inference Technique for Integrating Knowledge form Disparate Sources", Proc. 7th. International Joint Conf. Artificial Intelligence, Vancouver, pp 319-325, Aug. 1981.

[10]Hayes, P., "In Defence of Logic", Proc. 5th. International Joint Conf. Artificial Inteligence, M.I.T., pp 559-565, Aug. 1977.

[11]Jaynes, E. T., "The Well-Posed Problem", Foundations of Physics, 3, pp 477-493, 1973.

[12]Jaynes, E.T., "Where do we stand on Maximum Entropy", in "The Maximum Entropy Formalism", Levine and Tribus Eds. M.I.T Press 1979.

[13]Konolige, K., "Bayesian Methods for Updating Probabilities", Appendix D in "A Computer Based Consultant for Mineral Exploration", SRI report, Sept. 1979.

[14]Lemmer, J. F., and S. W. Barth, "Efficient Minimum Information Updating for Bayesian Inferencing in Expert Systems", Proc. National Conf. Artificial Intelligence, Pittsburgh, pp 424-427, Aug., 1982.

[15]McDermott, D. and J. Doyle, "Non-Monotonic Logic I," Artificial Intelligence, Vol. 13, Nos. 1,2, pp 41-72, April 1980.

[16]Pearl, J., and Kim, J. H., "A Computational Model for Causal and Diagnostic Reasoning in Inference Systems", Proc. 8th. International Conf. Artificial Intelligence, Karlsruhe, pp 190-193, Aug., 1983.

[17]Quinlan, J. R., "Consistency and Plausible Reasoning", Proc. International Joint Conference on Artificial Intelligence, Karlsruhe, pp 137-144, August 1983.

[18]Rauch, H. E., "Probability Concepts for an Expert System used for Data Fusion", AI Magazine, Vol. 5, No. 3, pp 55-60, Fall 1984.

[19]Reiter, R., and G. Criscuolo, "On Interacting Defaults", Proc. 7th. International Conf. Artificial Intelligence, Vancouver, pp 270-276, Aug. 1981.

[20]Shafer, G. "A Mathematical Theory of Evidence", Princeton University Press, Princeton, N.J., 1976.

[21]Tversky, A., and Kahneman, D., "Judgement under Uncertainty: Heuristics and Biases", *Science*, 185, pp 1124-31, Sept. 1974.

[22]Wallace, C.S. and Boulton, D.M. "An Information Measure for Classification", Computer Journal, 11, 2, pp 185-194, 1968.

[23]Zadeh, L. A., "Possibility Theory and Soft Data Analysis", In Mathematical Frontiers of the Social and Policy Sciences, Ed. L. Cobb and R. M. Thrall, pp 69-129.