Uncertainty and Probability

Jane Terry Nutter

TR 86-36

UNCERTAINTY AND PROBABILITY

Jane Terry Nutter
Department of Computer Science
Virginia Polytechnic Institute & State University
Blacksburg, Virginia 24061
(703)961-6931
nutter@vt.edu
nutter%vtcs1@bitnet-relay.arpa

ABSTRACT

Advocates of probability theory as a primary tool for reasoning in contexts of uncertainty and incomplete information have increased in number in recent years. At the same time, opponents have put forward a variety of arguments against using probabilities in this field. This paper examines the relationship between probability theory and reasoning in uncertainty, and argues that (contra opposing views) probability theory does have a place, but that its place is more restricted than many of its advocates claim. In particular, two major theses are presented and argued for. (1) Reasoning from probabilities works well in domains which permit a clear analysis in terms of events over outcome spaces and for which either relatively large bodies of evidence or long periods of "training" are available; but such domains are relatively rare, and even there, care must be taken in interpreting probability results. And (2) some generalizations with which AI applications must concern themselves are not statistical in nature, in the sense that statistical generalizations neither capture their meanings nor even preserve their truth values. For these contexts, different models will be needed.

INTRODUCTION

Probability estimates have been used to aid decision making in AI systems for over ten years now (see e.g. [Shortliffe and Buchanan 1975] and [Duda, Hart, and Nilsson 1976]), and have been under fire even longer (see e.g. [McCarthy and Hayes 1969]). Recently, the increased attention devoted to common sense reasoning and reasoning in contexts of uncertainty has fueled the debate, and clearly defined battle lines have emerged. Supporters point to a well-developed, well-understood, rigorous formalism for dealing with uncertainty (see especially [Cheeseman

1985] and [Ginsberg 1985]). Opponents continue to object, as McCarthy and Hayes did in 1969, that applying Bayesian methods to calculate probabilities requires information that is not generally available ("where do the numbers come from?"), and which, insofar as it is available, is usually tainted (for problems with human estimates of probabilities, see [Kahneman, Slovic, and Tversky 1982]). In addition, they charge that using probabilites suppresses important information (see e.g. [Cohen and Grinberg 1983] and [Sullivan and Cohen 1985]); that statistical analyses fail to distinguish uncertainty from inherent "fuzziness" (see e.g. [Zadeh 1981]); that judgements of typicality and normic generalizations underlie much reasoning from uncertainty and are not probabilistic (see e.g. [Rosch 1975], [Rosch and Mervis 1975], [Rosch, Mervis, Gray, Johnson, and Boyes-Braem 1976] on the non-probabilistic nature of typicality, [Scriven 1959] and [Scriven 1963] on nomic generalizations, and [Nutter 1982] on kinds of generalizations which AI systems must deal with); and so on. Even supporters of statistical approaches separate into those who prefer straightforward Bayesian analysis (e.g. [Cheeseman 1985]), those who prefer the Dempster-Shafer approach (e.g. [Yen 1986], Ginsberg 1984], [Ginsberg 1985], [Strat 1984], and [Yu and Stephanou 1984]; for the statistical theory, see [Dempster 1968] and [Shafer 1976]), and those who prefer some other variant of the Bayesian approach (e.g. [Snow 1986]). Others, notably Judea Pearl, are working on developing logical theories to permit logic-style inferences about probabilites, especially in the realm of reasoning about independence (see [Pearl 1986]), while continuing to argue that probability provides an epistemically adequate and effective framework for the general problem of reasoning in uncertainty (e.g. [Henrion 1986]).

Both sides present their positions forcibly and plausibly, but not always with attention either to their opponents' points or to independent substantiation. As a consequence, arguments on both sides of the fence have tended to produce more heat than light. The purpose of this paper is to attempt to lower the temperature while illuminating the terrain. In the end, this paper argues neither for nor against the use of statistical analysis in AI systems. Instead, it argues that statistical methods are applicable, but only in some cases, and that care must be taken to identify those cases correctly, to fulfill all requirements for reliable results, and to make certain that what is represented as a probability is indeed probabilistic in nature.

The body of this paper, then, consists of three sections. The first describes some elementary aspects of probability theory as analyzed through the discipline of statistics, to form a basis for discussion. The second argues for using statistical analyses in certain cases, and characterizes some of the features which an AI application and its domain must have for statistical analysis to be useful. The final section describes some instances of uncertainty which, it is argued, cannot be represented as probability.

PRELIMINARY REMARKS ON PROBABILITY

What are probabilities?

For the purposes of constructing AI systems, the philosophical nature of probabilities matters less than what kinds of phenomena classical and Bayesian probability analyses model. However, since there have been vehement disputes on this issue, perhaps a brief discussion is in order. The discipline of statistics begins investigating probabilities in any particular instance by defining (at least loosely) a space of outcomes, that is, mutually exclusive observations of test results. Events are sets of outcomes from that space. When probability theorists refer to probabilities, they typically mean (in the first instance) event probabilities, that is, the likelihood that the outcome of a particular test will belong to the set which defines the event. This likelihood is traditionally defined in terms of frequency: given a "sufficiently large" number of tests, what is the proportion of outcomes in the event set to total outcomes?

This frequency view has been attacked for centuries. A recent and persuasive account of several criticisms can be found in [Cheeseman, 1985]. Probably the single strongest argument (from an AI standpoint) against the frequency view is that taking this approach, each event has exactly one correct probability. But for AI purposes, such a probability is neither attainable nor, in some cases, even interesting; rather, we are interested in the probability of an hypothesis *given the current evidence*. A further objection is that the frequency theory "restricts probability to domains where repeated experiments (e.g. sampling) are possible, or at least conceivable" [Cheeseman, 1985]. In addition, the concept of "long run frequency" has bothered people for centuries. How long? How do you know? Why should "large numbers" (how large?) have special properties?

These objections can be met without deserting a frequency-based approach. The probability of any hypothesis on the basis of the current evidence can be -- and in normal statistical practice is -- interpreted as the conditional probability of the hypothesis given the conjunction of events which that evidence reflects. In other words, while the frequency view provides a single, well-defined probability for every event over the space, it also provides a mechanism for representing precisely the relativized probabilities we are most interested in, and these are exactly the probabilities that statisticians investigate. Second, classical statistics texts contain chapters on game theory and decision theory which describe in detail techniques for estimating probabilities on the basis of very small samples (see e.g. [Freund and Walpole 1980], Chapter 9, or almost any other freshman text). So not only does classical statistics recognize that this can be done, the theory instructs the interested in how to do it; only, it also warns not to place great faith in the accuracy of such estimates.

The hardest question to meet is the philosophical question of the significance of the Law of Large Numbers: what does it mean to talk about "long run" frequencies? Classical statistics provides some tests for whether an actual sample is large enough; but that cannot answer the philosophical question. The best that can really be said here is that other approaches have their own philosophical questions which they cannot answer either.

The classical alternative to the frequency view is the subjective probabilities view, which derives from the views of the eighteenth-century English clergyman Thomas Bayes. According to this approach, probabilities measure subjective certainty levels. There are two options here which should be distinguished. The first is well-defined, and clearly subjective (in the sense in which a philosopher would use the term): the probability of an event given the current evidence is the measure of the degree to which a (particular specific "real live") individual believes that the event will occur on the basis of that evidence. The problem with this interpretation is evident: people will believe all sorts of things, and different things at different times, for different reasons or none at all. There is no reason to suppose that one person's "probability" in this sense will match another's, and no grounds outside individual psychology for a *science* of probability at all.

It is unlikely that many supporters of subjective probabilities ever actually meant that, although they often seem to say it:

> ... 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.*
>
> ([Cheeseman 1985], 1003; emphasis in original).

The second alternative, and the view that is actually held, is that probabilities measure how much *an ideal, rational subject ought* to believe that an event will occur, given the evidence. This second option relativizes probabilities (to evidence), but has no subjective element (they do not depend on who does the believing). This approach has two difficulties, both as obvious and as pressing as the problem the frequency theory has with understanding the long run. First, what makes someone an ideal rational subject? On this view, probability cannot be considered well-defined until those properties are spelled out. Second, how (other than by measured frequencies) can we establish the degree to which such a subject ought to believe that a given event will occur?

How do probabilities behave?

The mathematics for measuring probabilities is the same on both these competing definitions: Bayes's Theorem is a theorem of classical statistics, for example. The significant differences come in questions of when it is legitimate to apply those formulas, and what they can be taken as establishing. In this regard, it seems (at least to me) that the frequency analysis has an advantage: designers of A.I. systems ought to care less whether their systems "ought" to believe their answers than how often those answers are right. Especially for systems making decisions or consultations with practical consequences, we should be measuring and maximizing that if we measure anything. But whatever philosophical view of probabilities is embraced, the mathematics always agrees with long run frequency expectations in all situations in which we can make sense of them.

Several of the mathematical features of event probabilities and their measurement are counterintuitive enough to be worth mentioning. First, experiments structured by statisticians *always* assume more than is known. Statistical experiments define a hypothesis about the likelihood of an event, and then compare actual observations against the predictions of the hypothesis. This has implications frequently overlooked in AI debates. One relates to the controversy over the so-called assumption of maximum entropy: the policy of assuming all events independent unless a connection has been found (see [Cheeseman 1985]). Opponents claim that this involves assuming more than is known, since the events in question may be dependent; supporters respond that the assumption of maximum entropy provides "a neutral background against which any systematic (non-random) patterns can be observed.... [W]ithout this prediction, it is difficult to detect if the current information is incomplete, and thus to discover new information" ([Cheeseman 1985], 1004). Any hypothesis provides a background for detecting deviation; and no experiment can be run without some hypothesis. The real question, then, is which hypotheses yield the best results without extensive "training"; this question must be answered by experiment, not argument.

Another, and for AI more serious, implication is that the outcome of an appropriate experiment must be observable independent of the statistical prediction. This is a problem for medical expert systems, for instance. A system which is trying to solve problems at the level of "diagnose infectious disease" is often predicting outcomes which cannot be independently determined by straightforward observation (if they could, we wouldn't need the systems). This has serious consequences concerning the "trainability" of such systems; we will return to this later.

Finally, some simple properties of probabilities should be noted. For independent events -- that is, events with the property that whether an outcome belongs to one has nothing to do with whether it belongs to another -- the joint probability (probability that all events will occur) is the product of the probabilities of the events. Since all probabilities are between zero and one, it follows that the joint probability of several independent events is always smaller than the probability of any one of them, unless all but one of them have a probability of one or at least one has a probability of zero. For dependent events, the joint probability is at most the maximum of the individual event probabilities, and it is that only if the corresponding event entails all the others; this has important consequences which we will return to later. Notice that the joint probability for dependent events may be zero even though none of the individual probabilities is, and it will always be so if at least two of the events are mutually exclusive. More subtly, the joint probability of, say, six events may be zero even though no two of them are mutually exclusive, if, say, five of them together exclude the sixth. Similarly, the probability that an outcome will fall into at least one of several independent events is the sum of the probabilities of the events in question. If they are dependent, it is at least the maximum of the individual probabilities, and at most their sum. Notice that a false assumption of independence *under*estimates the probability of disjunctions *and over*estimates the probability of conjunctions. In a long chain of reasoning involving both, it is not at all clear that these offsetting errors would be easy to detect and isolate.

APPLICATIONS FOR PROBABILITIES

Where decisions or predictions must be made on the basis of partial information, and where there is enough information to tell what outcomes are most likely given what is known, probability theory can be used to make these decisions and predictions accurately and responsibly. The mechanism is available, it is well-defined and well-understood, it gives good results, and it is the only mechanism we have with those properties. These facts alone suffice to show that probability has a place in reasoning in uncertainty, and henceforth in this paper the point will be taken as established. (For those who would like to see more on this issue, [Ginsberg 1985] presents both arguments in favor and a description of a proposed system for implementing probability reasoning.) But implementing probability-based reasoning requires more than computing some mathematical formulas. This section examines some of those requirements, and resulting consequences for AI systems which implement probability-based reasoning.

Where do the numbers come from?

Any application must consider where the system gets its information. There are two possiblities: a system may proceed from known probability values and distributions, or it may begin with initial probability estimates which are recognized as possibly inaccurate and which are refined in the light of further evidence (in Bayesian terminology, these estimates are called "prior probabilities" or "priors"). The first choice provides better initial results, and is easier to implement. Unfortunately, it requires a depth of knowledge in the application domain which is almost never attainable, and so the second approach is the one most often taken.

"Prior probabilities" are not probabilities: they are guesses. The only interpretation of probabilities under which priors qualify is the extreme subjective view discussed above, on which anyone's assessed level of commitment is *a fortiori* a probability, but not an interesting one, since on this view, a science of probabilities is probably impossible and in any case would have nothing to do with what is or is not likely to happen.) If the priors are bad guesses, then results based on them will be bad results, even if all other assumptions hold. There are two ways to mitigate this problem: base priors on data and experiments so as to start with good ones, or validate or train the system to improve bad ones.

Basing priors on data and experiments is straightforward and the most reliable course, when it can be done. Where data from reasonably representative samples already exist, those data can be used to establish priors. Where such data do not already exist, classical experiments based on the outcome distribution and the predictions of a testable hypothesis are designed. For some A.I. system domains, this procedure is feasible. The domain for PROSPECTOR, for instance, is small and reasonably well understood. We have fairly good information on the occurrence rates of different minerals, and it is easy to imagine, at least, what it would be like to have reliable estimates of joint and conditional probabilities over many of the relevant events. Since knowledge of conditional probabilities entails knowledge of which events are independent of each other, this knowledge also obviates to a great extent the need for assumptions like maximum entropy.

Unfortunately, domains in which this kind of information is available are rare. By contrast, if we consider a medical domain, the number of possible outcomes is huge, their distributions are less well known, their interactions are frequently unknown, and reliable data are notoriously hard to come by. The number and scope of experiments necessary to establish such data are overwhelming. In cases such as this, some other course must be taken. The usual technique at present is to rely on opinions from experts, formally or informally elicited, from individual experts alone or panels in consultation, by any of a variety of strategies.

There are many reasons for doubting the accuracy of such estimates. First, people in general and trained scientists in particular are lousy at estimating probabilities. The classic studies establishing this ([Tversky and Kahneman 1971] and [Tversky and Kahneman 1974]) were published over a decade ago, and many more have followed, confirming and extending the results (see e.g. the articles in [Kahneman, Slovik, and Tversky 1982]). Second, even assuming that the experts on the panel are sophisticated enough to avoid the most common kinds of error, they lack the information they would need to make accurate estimates. (The problem isn't that the experts refuse to give us this information; they don't have it either.) The question is what to do about it.

Option One: Do nothing. It is a theorem that repeated application of Bayesian analysis to a given event yields a sequence of priors which converges on the frequency probability, however abysmal the original prior. So if we start with whatever estimates we have available and let the system improve them as it goes, we will arrive at good results in the course of nature.

There are three problems with this. (a) The system can only improve its estimates if it has independent information on whether the outcome matches its predictions. For a medical system, this independent information is often unavailable. (That a patient got well after treatment does not as a rule confirm the diagnosis, for instance, since many treatments help with a broad range of problems, and anyhow, most patients get well anyway.) In the absence of independent information on the outcome, the system's estimates will not improve. Worse yet, they may *seem* to be confirmed because disconfirming instances, although present, go undetected. (b) Even if the system improves over time, it starts off doing badly. If we care what answers we get, we may not want to tolerate this. (It is good if a program learns the right treatment to recommend; it is less good if it learns by recommending wrong treatments which kill nine or ten people....) (c) While the mathematics guarantees that estimates will converge, they may do it slowly. If the initial priors are bad enough, it may be a very long time before the system's predictions get much better.

Option One A: Do nothing, but report ignorance. Many researchers propose the use of ranges as opposed to point probabilities to reflect "second order" uncertainty: wide ranges reflect very uncertain estimates, while narrow ranges reflect more certain ones. Dempster-Shafer theory is then used to calculate probabilities on the basis of these ranges. This has the advantage that a user who receives an answer with an attached probability of, say, [0.41, 0.99] has been warned that the system really doesn't know, whereas a user who receives the same answer with an attached probability of 0.7 has not. But there is no more a reliable guide for setting the ranges than there is for setting the priors, so that they too may prove misleading; and a bad answer with a warning is still a bad answer. This approach can be combined with the measures below for improving the quality of the priors; whether and how much improvement will result remains to be seen.

Option Two: Validate the system's predictions. While it may not be possible to check an expert system's judgements against actual outcomes, it is possible to check them against the judgements of actual experts. Comparison against human performance may not give results as good as "true" priors would, but it meets any standard that could reasonably be expected. However, in this as in any other context, care must be taken to ensure that validation is not tainted.

In particular, the system's performance should be compared with the performance of several experts (the more the better) on the same cases (not just on cases which are judged similar). In addition, the experts in question should not know how the system is constructed, what its basic assumptions (including its priors) are, what questions it asked, what conclusions it reached, or how it reached them. In practice, this means that validation must take place in an environment other than the one in which such a system would be used: if the considered opinions of several experts were routinely available, there would be no need for the expert system. Finally, since the point of validation is to tune the system's priors, this phase must be pursued with the attitude that in case of disagreement, and in the absence of overwhelming evidence to the contrary, the experts are right and the system is wrong.

Option Three: Train the system. If outcomes are independently observable, the fact of convergence can be used to put a system through an initial training phase. This amounts to the "do nothing" approach, but pursued "off-line" and with careful supervision until there are signs of convergence. Given observable outcomes, this lets the mathematics work for the designers, while permitting intervention if radically unstable results show that a particular estimate was so bad that it will take a long time to converge. It should be noted, however, that if many factors are involved, or if original estimates are very bad, this course may require very large amounts of data before responses become reliable enough for the system to "go on-line" responsibly.

What can the numbers tell you?

The most likely hypothesis is usually not the best explanation. The reason for this has to do with the intuitively paradoxical fact that a hypothesis which covers half the information and ignores (is indifferent to) the rest is more probable than one which gives exactly the same explanation of the first half of the data and then goes on to explain the rest. Recall that the probability of a conjunction is at most equal to the maximum of the probabilities of its conjuncts, and only reaches that limit when the conjunct with the highest probability entails all the others. It follows directly from this that more specific hypotheses are mathematically guaranteed to have lower probabilities than any less specific hypotheses which they entail: adding information to a hypothesis reduces its probability. Always. Ten out of ten.

So, consider the following example, once again in the medical domain. Suppose (our data) that a patient has a fever, a sore throat, white spots on the tonsils, nausea, diarrhoea, and vomiting. Now consider two possible hypotheses: (*A*) the patient has strep throat; (*B*) the patient has strep throat and a gastro-intestinal virus. For the reasons outlined above, *A* is (necessarily) the more probable hypothesis; but *B* is intuitively the better explanation.

In many cases, what we really want from the system is the best explanation. This means something like, the hypothesis which best covers the facts while maintaining a reasonable probability. If the "answer space" has more than one level of granularity, this is different from the most probable hypothesis, because more specific explanations are preferred to more general ones, so long as they do not become "intolerably unlikely". This is not to say that a system which is looking for the best explanation cannot use probability-based reasoning to advantage. But such a system requires a more sophisticated answer selection mechanism than "most probable hypothesis": whatever else goes on, at the end of the calculations, some reasoning takes place -- to select the best explanation as opposed to most likely hypothesis -- which is not probabilistic in nature. Hence, *pace* Cheeseman, even in situations which admit of a clear statistical analysis, it takes more than just probabilities to reason in uncertainty.

What do the numbers cost?

Probabilities have been proposed as a limiting threshhold to keep down the cost of inference (see e.g. [Ginsberg 1985]). It is not clear, however, that they can fulfill this function in the way that their supporters suggest. For instance, Ginsberg [1985] suggests that inferences can be limited by setting a threshhold with the property that once a conclusion's probability reaches that limit, it may be taken as established without further investigation. For the sake of argument, say that the limit in question is 0.98. Suppose we know the following: anything which is A has a probability of 0.99 of also being B; x is A; anything which is C has a probability of 0 of also being B; x is C. Now we ask whether x is B. The answer, of course, is no. But if the system first finds the rule about As and the fact that x is A, giving a probability of 0.99 for x being B, it will cut off inference there. So the argument that using probabilities allows early termination needs to be taken with a grain of salt. The "best" (highest hit rate) hypothesis for *any* rare event is *always* the hypothesis that it never happens. That isn't useful if we are trying to predict, detect, and reason about rare events. Non-numerical inference mechanisms can terminate inference early, for instance on the basis of resource limitation (see e.g. [Donlon 1982]); only, the system may miss an answer it would otherwise get. Likewise, a system which threshholds on probabilities can terminate early; but it may get the wrong answer. For more on threshholding problems, especially with regard to the Dempster-Shafer approach, see [Dubois and Prade 1985].

In addition, the training process may prove more expensive than it first appears. Ginsberg [1985] is one of the few presentations of a system which goes into detail on the process of getting tests to improve the quality of probability judgements, so the following remarks will be made in the context of his system's behavior. But these costs result from necessary steps if the system trains "on-line": any system which counts on this must either sacrifice accuracy or pay these prices. Whenever an inference establishing a probability for a proposition is made, Ginsberg's system keeps a record of all evidence which was used. Then every time new evidence affects the probability of a proposition, every inference in which that proposition was used is retraced to update the conclusion's probability. Unfortunately, until the number of tests gets large, none of the probabilities in the system are reliable. So for most of its early life, the system must recompute the probability for virtually all its inferences every time it sees anything. Using cut-offs makes this worse, by the way, since in that case the system really cannot just retrace the proofs that went through; it should also recheck those that were terminated early, of which presumably no record exists. In effect, this means that rather than retracing a known path, it must perform the entire inference again from scratch.

Also, A.I. systems rarely perform controlled experiments, testing selected observable variables against the predictions of some hypothesis. Instead, they get information and reason forward from it. This means that to get the full benefit of training "tests", systems which train must analyze every new piece of information to determine which of the events they know about this datum may fit. Consider such a system, when it first meets Fred the Flamingo. Not only is it meeting a bird that flies -- it is meeting something pink that flies; something over three feet tall that is pink; a female named Fred (as it happens); and so on, and so on, and so on.... Combining this need to extract events from new data with the need to retrace and correct probabilities based on previous inferences, the computational cost should be clear.

WHERE PROBABILITIES DON'T BELONG

Pace Cheeseman (and many, many others), not all that is not universal is probabilistic. For instance, if, as Cheeseman claims, the by now tormented example "Birds fly" really means "Most birds fly", then birds don't fly in the spring. In the nesting season, baby birds outnumber adults. Baby birds don't fly (and as we all know, neither do some of the grown ups). Hence in the nesting season, "Most birds fly" is false, while "Most birds cannot fly" is true. By the way, we can do even better with "Birds lay eggs," which is out-and-out false (year round) of more than half of the population (none of the males do, for starters). So if Cheeseman is right, anyone who says in the spring that birds fly, or at any time that birds lay eggs, is mistaken. This is nonsense.

"Birds fly" must be decoded with respect to typicallity. If typicallity can be modeled by any statistical notion, it is category cue validity, not probability. (For a development of this theory, see [Rosch 1975], [Rosch and Mervis 1975], and [Rosch, Mervis, Gray, Johnson, and Boyes-Braem 1976].) "Birds lay eggs," on the other hand, is not statistical at all. It is shorthand for a genuine, accept-no-substitutes universal -- but not for "For all x, if x is a bird, then x lays eggs". Instead, it is in a class with the non-universal generalizations "Mammals bear young alive" (duck-billed platypi are egg-laying mammals) and "Reptiles and fish lay eggs" (garter snakes and sharks bear live young). By the way, these generalizations cannot be translated straightforwardly into probability claims counting over species instead of individuals: *no* species *either* bears live young *or* lays eggs; only (female) individuals belonging to species do.

The typicality-based uncertainty involved in generalizations like "Birds fly" centers on whether an individual has a property typical of things of its kind. There is another kind of uncertainty, also related to typicality, but centering instead on the extent to which a given property applies to an individual. This is the issue of vagueness. The kind of inferences justified on the basis of degree-of-applicability are different from the kinds based on either typicality or probability. The difference between measure-of-membership and typicality is subtle, but real. Typical birds fly. But how typical a bird Tweety is does not measure how well Tweety flies or how even how likely Tweety is to fly (hummingbirds are atypical in many ways, but spectacularly good fliers).

More importantly, because more often confused, degree-of-applicability does not work like probability, although they tend both to be measured on a zero-to-one scale and reported alike. For instance, consider the following two claims about Oscar the Ostrich:

(i) Oscar is a (typical) bird at 0.6
(ii) Oscar is male at 0.5

Claim (i) says that Oscar is not very birdlike, presumably on the grounds that ostriches aren't. It is the sort of claim that Zadeh's fuzzy set theory was originally developed to handle; it attempts to measure the extent to which an individual falls within the bounds established by a fuzzy concept. Claim (ii) is a probability claim, presumably reflecting that the system doesn't know whether Oscar is male but does know that Oscar is a bird, and that half of all birds are male, making the chances that Oscar is male 50-50. But that is not to say that Oscar is half male: it is consistent for the system to hold (ii) and also to hold that any given bird is either completely male or not at all. The two claims look superficially alike, but they cannot reasonably be taken the same way: (i) says that Oscar is not a very typical bird; (ii) does not say that Oscar is not very male. In case (i), the result may not reflect incomplete information at all. It reflects a fundamental fact about how Oscar relates to a vague concept. In case (ii), the information is incomplete and can be completed by a single experiment (look at Oscar and see). If no difference in representation reflects this basic difference in content, the system will reason incorrectly a good part of the time.

Translations of common generalizations into probabilities do not preserve truth values, and translations of degree-of-applicability claims do not preserve inferences. So neither preserves meaning. Hence wherever else they can be used, probabilities cannot be used to understand all generalizations and expressions of uncertainty in natural language understanding, or in any system which gets its data in natural language. Natural language understanding requires inference in contexts of uncertainty all over the place, including inference from (previously processed or understood) non-statistical generalizations. Hence there are instances of inference in contexts of uncertainty which are not amenable to analysis as probabilities.

CONCLUSIONS

Probability theory is an important tool for many AI applications which involve uncertainty. In cases where outcome likelihood is at stake, and where the necessary data are available, it is the best known mechanism. But it is also *hard*. It requires a detailed analysis and understanding of the domain, and either a great deal of data (for deriving priors or training) or an extensive validation procedure, if the answers obtained are to be reliable. There are no short cuts.

Also, statistics cannot provide a panacea for all problems of uncertainty, generality, vagueness, and ignorance. No matter how rigorous a mathematical model, it can only be expected to give reasonable answers if the situations to which it applies actually conform to its underlying assumptions. The existence -- and prevalence -- of non-statistical generalizations guarantees that other mechanisms for dealing with uncertainty must also be investigated. It is true that the existence of a well-defined, long-studied, thoroughly articulated theory is strong argument for its use -- but not for its extension, willy-nilly, into other fields. Sometimes, one theory is developed before another because the first makes sense and the second doesn't; other times, as for instance with physics and biology, we simply make progress on the easier problem first.

Ultimately, A.I. systems which reason in uncertainty need ways to combine all these modes of uncertainty. In particular, we need to distinguish representations of probabilities, fuzzy membership, and generalizations based on typicality. One such approach would represent the first two as functions which take properties and yield either numbers or second order relations, and represent the third using some form of default reasoning (I have argued elsewhere for a simple monotonic extension to first order logic; see [Nutter 1983]). Axioms and rules can then make use of information when and as it is available, without misrepresenting that information and so making wrong inferences from it. Obviously this scheme is utopian; so what can we do meanwhile? Some application domains are particularly amenable to one of these forms of inference, and can do without the others; in these cases, we make choices, hopefully understanding the limitations and trade-offs. On the science front, we can develop as many models as we can, with as close attention to the phenomena to be modeled as possible. And meanwhile, we can remember that given the diversity of kinds of reasoning involved, almost anyone who claims to have the one and only key to reasoning in uncertainty is almost certainly wrong.

ACKNOWLEDGEMENTS

My thanks to Dr. Janet C. Rice of the Department of Biostatistics and Epidemiology, School of Tropical Medicine and Public Health, Tulane University, for her patience in looking over early versions of this paper to check for gross statistical bloomers. As always, credit goes to her only for accuracy; any errors lie solely at the author's door.

REFERENCES

Cheeseman, P., "In defense of probability", *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (IJCAI-85), 1985, 1002-1009.

Cohen, P.R. and Grinberg, M.R., "A framework for heuristic reasoning about uncertainty", *Proceedings of the Eighth International Joint Conference on Artificial Intelligence* (IJCAI-83), 1983, 355-357.

Dempster, A.P., "A generalization of Bayesian inference", *J. Royal Statistical Society* **30** (Series B), 1968, 205-247.

Donlon, G., "Using resource limited inference in SNePS", SNeRG Technical Note No. 10, Department of Computer Science, SUNY at Buffalo, 1982.

Dubois, D. and Prade, H., "Combination and propagation of uncertainty with belief functions -- a reexamination", *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (IJCAI-85), 1985, 111-113,

Duda, R.O., Hart, P.E., and Nilsson, N.J., "Subjective Bayesian methods for rule-based inference systems," *Proceedings of the 1976 National Computer Conference*, AFIPS Conference Proceedings Vol 45, New York, 1976, 1075-1082.

Freund, J.E. and Walpole, R.E., *Mathematical Statistics*, third edition, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1980.

Ginsberg, M.L., "Non-monotonic reasoning using Dempster's rule", *Proceedings of the National Conference on Artificial Intelligence* (AAAI-84), 1984, 126-129.

Ginsberg, M.L, "Does probability have a place in non-montonic reasoning?", *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (IJCAI-85), 1985,107-110.

Henrion, M, "Uncertainty in AI: is probability epistemologically and heuristically adequate?', seminar presented at M.I.T., Boston, MA, 10 November 1986.

Kahneman, D., Slovic, P., and Tversky, A., eds, *Judgement under Uncertainty: Heuristics and Biases*, Cambridge University Press, New York, 1982.

McCarthy, J. and Hayes, P., "Some philosophical problems from the standpoint of artificial intelligence", in *Machine Intelligence 4*, B. Meltzer and D. Michie, eds, Edinburgh University Press (Edinburgh) 1969, 463-502.

Nutter, J.T., "Defaults revisited, or 'Tell me if you're guessing'", *Proceedings of the Fourth Annual Conference of the Cognitive Science Society*, 1982, 67-69.

Pearl, J., "On the logic of probabilistic dependencies," *Proceedings of the National Conference on Artificial Intelligence* (AAAI-86), 1986, 339-343.

Rosch, E., "Cognitive representations of semantic categories", *J. Exp. Psych.: General* **104**, 1975, 192-233.

Rosch, E. and Mervis, C.B., "Family resemblances: studies in the internal structure of categories", *Cognitive Psychology* 7, 1975, 573-605.

Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.B., and Boyes-Braem, P., "Basic objects in natural categories", *Cognitive Psychology* 8, 1976, 382-439.

Scriven, M., "Truisms as the grounds for historical explanations", in P. Gardiner, ed, *Theories of History*, Free Press, New York, 1959.

Scriven, M., "New issues in the logic of explanation", in S. Hook, ed, *Philosophy and History,* New York University Press, New York, 1963.

Shafer, G., *A Mathematical Theory of Evidence,* Princeton University Press, Princeton, New Jersey, 1976.

Shortliffe, E.H. and Buchanan, B.G., "A model of inexact reasoning in medicine," *Mathematical Biosciences* 23, 1975, 351-379.

Snow, P., "Bayesian inference without point estimates," *Proceedings of the National Conference on Artificial Intelligence* (AAAI-86), 1986, 233-237.

Strat, T.M., "Continuous belief functions for evidential reasoning", *Proceedings of the National Conference on Artificial Intelligence* (AAAI-84), 1984, 308-313.

Sullivan, M. and Cohen, P.R., "An endorsement-based plan recognition program", *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (IJCAI-85), 1985, 475-479.

Tversky, A. and Kahneman, D., "Belief in the law of small numbers", *Psychological Bulletin* 2, 1971, 105-110.

Tversky, A. and Kahneman, D., "Judgement under uncertainty: heuristics and biases", *Science* **185**, 1974, 1124-1131.

Yen, J., "A reasoning model based on an extended Dempster-Shafer theory", *Proceedings of the National Conference on Artificial Intelligence* (AAAI-86), 1986, 125-131.

Yu, S.Y. and Stephanou, H.E., "A set theoretic framework for the processing of undertain knowledge", *Proceedings of the National Conference on Artificial Intelligence* (AAAI-84), 1984, 216-221.

Zadeh, L.A., "Possibility theory and soft data analysis", in *Mathematical Frontiers of the Social and Policy Sciences*, L.M. Cobb and R.M. Thrall, eds, AAAS Selected Symposium **54**, Westview Press (Boulder, Colorado) 1981, pp. 69-129.