PROBABILISTIC INFERENCE and PROBABILISTIC REASONING*

by

Henry E. Kyburg, Jr.
University of Rochester
Rochester, New York 14627

Commonsense Reasoning **3900 words**

There are two profoundly different (though not exclusive) approaches to uncertain inference. According to one, uncertain inference leads from one distribution of (non-extreme) uncertainties among a set of propositions to another distribution of (non-extreme) uncertainties among those propositions. According to the other, uncertain inference is like deductive inference in that the conclusion is detached from the premises (the evidence) and accepted as "practically certain;" it differs in being non-monotonic: an augmentation of the premises can lead to the withdrawal of conclusions already accepted.

We show here, first, that probabilistic reasoning and probabilistic inference are distinct; second,that probabilistic inference is what both traditional inductive logic ("ampliative inference") and non-monotonic reasoning are designed to capture, third, that acceptance is legitimate and desirable, fourth, that statistical testing provides a model of probabilistic acceptance, and fifth, that a generalization of this model makes sense in AI.

*Research underlying this paper has been partially supported by the Signals Warfare Center of the U. S. Army.

1. Probabilisitic Inference and Probabilistic Reasoning

Uncertainty enters into human reasoning and inference in at least two ways. It is reasonable to suppose that there will be roles for these distinct uses of uncertainty also in automated reasoning.

One role for uncertainty concerns choices among alternative actions. For good reasons, a conception of uncertainty that did *not* satisfy the probability axioms, but that was used for computing expected utilities, would lead the agent into decisions under which the agent would come out on the short end in every eventuality. (The Dutch Book Theorem, Ransey, 1950.) Uncertainty, construed in this way, is what we need for computing the expectations that are fed into decision rules, the most common and persuasive of which is the rule to maximize expected utility. (This rule may not always be applicable.)

Uncertainty need not be represented by a single classical probability function; if uncertainties are represented in some more general fashion, what the Dutch Book argument shows is that there should exist some classical probability function that is compatible with that more general representation. It has been argued (Levi, 1980; Kyburg, 1987) that the most general form of representation for these uncertainties is that of a set of classical probability functions, having a convex hull, defined over an algebra of propositions. Such a representation includes as special cases belief functions (Shafer, 1976) and most interval representations of uncertainty. Less general forms of convexity, for example, in which a set of parameterized probability functions is taken to be convex over values of the parameter, are also of interest. (A well-known example concerns "exchangeable" random quantities, which may generally be regarded as mixtures of independent random quantities. The analysis of these distributions provides the ilntuitive justification for the use of subjective probabilities.) Manipulating probability and utility representations, seeing which ones are consistent with which, or which imply which, constitutes one form of probabilistic reasoning. It is this kind of reasoning that Nilsson (1986) considers to be the appropriate subject matter of probabilistic logic.

In addition to merely representing uncertainty and employing it in decision theory, we are concerned with how uncertainties are modified or updated in response to evidence. The classical way of doing this, for classical probabilities, is by means of Bayes' theorem: if E becomes known, is accepted as evidence, then the new or updated probability P' of any statement H in our algebra becomes the old likelihood of E multiplied by the ratio of the old probability of H to the old probability of E:

$$P'(H) = P(E|H) * (P(H)/P(E))$$

This is called confirmational conditionalization. A more general procedure is Jeffrey Conditionalization (Jeffrey, 1965), which applies when we undergo some experience or make some observation whose import is exactly to lead us to shift some probability from $P(E)$ to $P'(E)$. Then the new probability of every other statement becomes:

$$P'(H) = P(H|E)*P'(E) + P(H|\sim E)(1 - P'(E))$$

Confirmational conditionalization can be extended to the more general approach that represents uncertainty by convex sets of classical probabilities: it can be shown that if each classical probability function in a convex set of probability functions is updated by conditionalizing on the evidence E, the result will be a new convex set of classical probability functions, provided E does not have zero probability on all the original probability functions (Kyburg, 1987).

There are other ways in which one might want to update probabilities than by conditionalization -- certain forms of direct inference, in which probabilities are derived from knowledge of statistics or chances, have been shown to conflict with conditionalization, for example (Levi, 1980). But while any of these procedures have a perfect right to be called 'probabilistic reasoning,' they are not what I mean by probabilistic inference.

In inference, in general, one begins with certain statements or propositions (representations of states of affairs), premises, and goes through a process that leads to another statement, the conclusion. From "Tosses of this coin are

independent and heads occur half the time," we infer, not probabilistically, but deductively, that if we know that heads has occurred on the first toss of a pair, then the probability that heads also occurred on the second toss of the pair is a half. We infer, not probabilistically, but deductively, that triples of tosses consisting of three heads occur an eighth of the time. We infer, not probabilistically, but deductively, that if our sample of n from P is random with respect to R, then the probability is at least 0.91 that the proportion of P that are R lies within $3/(4n)^{1/2}$ of the observed sample population.

In ordinary deductive logic, the process of inference is such as to preserve truth: if the premises are true, so is the conclusion. Note that the probabilistic reasoning mentioned above fits into this deductive pattern.

In most applications of deductive inference, we do not know that the premises are true: they may be warranted, firmly believed, accepted as practically certain. Sometimes the premises are merely accepted hypothetically, or for the purposes of argument.

Often valid deductive inference provides warrant for accepting its conclusion. But not always: confronted with the validity of the inference: "All swans are white; Sam is a black Australian swan; Therefore Sam is white," we opt for the rejection of the universal premise rather than for the (inconsistent) inclusion of the consequence.

Spelling out the conditions under which valid deductive inference provides warrant for its conclusions is not trivial. It is not trivial because it depends on spelling out the justification for the premises of a deductive argument. This much is quite clear.

2. Ampliative Inference.

What is controversial is whether or not there is any form of inference other than deductive inference. Is there any way of arguing from premises to conclusion that is not (necessarily) truth preserving, and if there is, why would one want to do it anyway? Of course there is a tradition in philosophy that considers "inductive inference," "ampliative inference," and the like (Kneale and Kneale, 1962).

David Israel (1986) suggests that there is no other logic,

but that real inference proceeds in non-deductive ways. This is just what the philosophical tradition of inductive or ampliative inference, or the more recent philosophical concern with scientific inference has been concerned with. Whether or not there is a "logic" of such inference is indeed controversial. In either event, it is something worth looking for; if it doesn't exist, the search for it will nevertheless prove enlightening.

In artificial intelligence, this search is to be found in the search for representations of non-monotonic reasoning (such as circumscription), non-monotonic logics, default logics, and the like. We want to be able to infer that Tweety can fly. Since the kinds of inference under investigation do not preserve truth, we have to be able to back up: if we enlarge the premiss set, we may have to shrink the conclusion set. Non-monotonic inference is not generally taken to be probabilistic, but work on non-monotonic logic suggests that there is interest in inference rules -- that is, rules that lead from premises to the acceptance of a conclusion -- that need not be truth preserving. Many people want to be able to detach conclusions from their premises. Not all approaches to non-monotonic logic allow full detachment; de Kleer's ATMS (de Kleer, 1986), for example, requires that tags reflecting the assumptions used in carrying out an inference be carried along with the conclusions.

3. Why Accept?

Despite the fact that some people are interested in non-deductive inference, we may still sensibly ask why they should be: Why should we accept any statements that are not (say) mathematical or logical truths? It might be thought that we couldn't use conditionalization for updating without acceptance: after all, when we update on evidence E, we take the probability of E to be one. And once a statement has a probability of 1 (or of 0) that probability can never be changed by conditionalization. But there are other ways to handle updating: Jeffrey's rule, for example, or various net-propagation procedures, such as Pearl's (Pearl, 1986).

In principle, there is no reason that human or machine knowledge in a certain domain should not be represented by a complete algebra of statements and a classical probability distribution (or a set of classical probability distributions) over

them, in which no empirical statement ever receives a probability of 0 or 1. Such a system would have no need for a probabilistic rule of inference.

While such a system would be conceptually simple, it would not reflect the way in which people function epistemologically. I find myself willing to assert categorically a large number of propositions that could conceivably be false. Perhaps more to the point, The progress of science depends on the categorical acceptance of such statements as: measurement *m* is not in error by more than so and so much. The efforts of engineers in designing a tool or a product depend on their ability to take as <u>given</u> such possibly erroneous statements as those concerning strength of materials, conductivity, ... Such statements we take as <u>premises</u> in deductive arguments, whose conclusions we therefore also accept. I am willing to take such statements as <u>evidence</u> -- e.g., I take as evidence the statement that about 50% of tosses result in heads; it is relative to this evidence that I assign a probability of heads on the next toss of about .5.

Our empirical scientific knowledge is expressed, not in probabilities (for the most part) but in categorical statements. There is a sense in which we may want to say that our science is uncertain; but there is no obvious probability we associate with the principle that the vector sum of the forces acting on a static body must be zero. We do not take measurement to result in statements such as "with probability 0.9, the reading 4.30 was obtained," (for we regard that as a matter of record, and therefore of probabilitiy 1) nor do we report the result of the measurement as an unbounded normal probability distribution: the mean of the normal distribution of observations of the quantity measured is (estimated to be) 4.30. We report (with confidence .99) that the value is 4.30 ± .02. We use this interval for the next step in our reasoning or design or decision process.

As a matter of practicality, no one, I suspect, has ever tried to represent a significant piece of knowledge or expertise in the form of a distribution ofer a complete field of sentences. It would be perverse. The universe of <u>all</u> possibilities is just too large to handle this way. We must cut it down to size by ignoring some possibilities, or we will use all our resources doing

probability cojputations rather than solving our real problem. When we measure a rod by a method M whose distribution of error is normal with a mean of zero and a standard deviation of.01, we don't worry about the finite probability that the reading is off by more than .05. As for the distribution of error itself, we don't even keep the data: the hypothesis was confirmed well enough. Maybe the mean is really 10^{-6} rather than 0. Maybe the standard deviation isn't exactly .01. But the probability of a significant difference is too small to bother about. This is probabilistic inference in action.

4. Models of Probabilistic Inference.

In testing a statistical hypothesis, the standard goal is to devise a rule that will erroneously reject that hypothesis no more than α of the time. Such a test will lead you to a false rejection no more frequently than α (Lehman, 1959). Of course α is a free parameter; but we choose α to be small enough that the possibility of making this sort of error in a given context does not worry us. The size α we choose reflects how seriously we take the mistake in question. If it is very serious, we want to be very sure (but we can't ask for a guarantee) that it won't happen.

It is very bad form to say of a hypothesis that has been rejected at the level that the probability is at most α that it was falsely rejected. But as Birnbaum has pointed out (Birnbaum, 1969), while we can learn not to say such things, it is hard to know what else to think.

Consider the simplest and most elegant of all forms of statistical inference: you have a normally distributed quantity X, but you don't know the parameters of its distribution. Nevertheless, since you iknow that it is normally distributed, you know the distribution of the quantity $t = N^{1/2}(x - \mu)(s^{-1})$, where x and s are the sample mean and standard deviation, and μ is the unknown population mean. Knowing the distribution of t, you can therefore compute the probability, for example, that

$$x - t_{\phi} N^{1/2} < \mu < x + t_{\phi} N^{1/2}.$$

If you pick some probability level that makes you feel comfortable

under the circumstances, and you are indifferent between over and underestimating μ, then you will have an exact interval estimate of the unknown mean μ, indexed by f_p: a level of fiducial probability or practical certainty.

Note that this inference is non-monotonic: on observing a further sample from the population it may well be that some different interval for μ will be acceptable at the index f_p.

Or consider the most common form of confidence interval inference: you have a binomial population with an unknown parameter r; you draw a sample from the population, and observe a relative frequency f; you construct a class of intervals (p_l, p_u) such that whatever the true value of r may be, the probability is at least p that the sample frequency will fall in the corresponding interval. But it will have done this *if and only if* r lies between a certain maximum and a certain minimum value. These values determine what is called a confidence interval, and in particular, a $100p$% confidence interval, since its limits require the specification of an acceptable p.

Outside of statistics, consider Levi (1967). Levi is concerned with the circumstances under which one ought to add a hypothesis to one's corpus of knowledge. The famous *Rule A* for doing so involves, in addition to the probability of the hypothesis, a measure of the epistemic content of the hypothesis, and a further parameter q, which varies from 0 to 1 and functions as an index of caution.

In artificial intelligence Matthew Ginsberg (1985) applies a technique much like that of binomial confidence interval inference (the main difference being that he uses a rougher approximation) to the problem of inferring an interval characterizing the reliability of a default rule in non-monotonic logic. In order to do this, he finds it necessary to introduce a parameter g, analogous to the fiducial parameter f_p, which he calls "gullibility."

Finally, in my own work (1961, 1974) I have adopted a "purely probabilistic" rule of acceptance. That is, a body of knowledge is indexed by a "level of acceptance"; statements whose probabilities (relative to a body of knowledge of even higher rank) are greater than this level of acceptance may be accepted.

5. Probabilistic Acceptance.

The simplest and most natural idea for acceptance in AI is just to accept those statements whose probability exceeds a certain critical number. This number may have to be changed to reflect different circumstances -- it will be context dependent -- but so, we may suppose, are α, g, q, p, and f_p context dependent.

In what way is acceptance level context dependent? One natural answer is that acceptance level depends on what is, or might be expected to be, at stake. If the range of stakes we are contemplating is limited -- for example, it can't be more than 10 to 1 -- then probabilities greater than .9 are behaviorally indistinguishable from probabilities of 1, and probabilities of less than .1 are indistinguishable from probabilities of 0.

It also follows from these considerations that probabilities larger than the level of acceptance, or smaller than 1 - the level of acceptance, are just not significant as probabilities. That is, it makes no sense to bet at odds of 1000:1 on a statement that gets its probability from a statistical statement whose acceptance level is only .99. If you're only 99% sure that the coin lands heads between .48 and .52 of the time, you should not be willing to bet at odds of a thousand to one than in 12 tosses you won't get 12 heads. The constraint cuts both ways.

Most of the acceptance rules mentioned above run afoul of the lottery paradox (Kyburg, 1961). That is, each of a set of statements S_i (e.g., "ticket i will not win the lottery") may be probable enough to be accepted, and at the same time may jointly contradict other accepted statements (e.g., "there will be a winner"). The only exception is the acceptance principle advocated by Levi, which links acceptance to expected epistemic utility; only statements demonstrably consistent with what you have already accepted are candidates for future acceptance.

How serious the lottery paradox is depends on what other machinery you have. It is not deadly if you limit yourself to a probabilistic rule of acceptance. It will follow that any logical consequence of a single statement in your corpus of knowledge should also be in it; but it will not follow that every consequence of the set of sentences in your corpus will also be in it. The

latter would indeed lead to a hopeless sort of inconsistency; the former should not. If the size of the lottery is adjusted to my level of acceptance, I will answer your question about whether ticket i will win with a categorical "no." But I will answer your question about whether it is true that neither ticket i nor ticket $i + j$ will win by saying, "I don't know; but the probability is thus and so.

This seems not unreasonable. Look at the matter another way: given a (deductive) argument from premises P_1, ..., P_2 to a conclusion C, consider whether the argument obligates you to accept C. It seems natural to say that more is required than merely that <u>each</u> of the premises be accepted; I must also be willing to accept the conjunction of the premises.

Even this feature might be advantageous in A I. There is surely an epistemic difference between a conclusion reached in one step from a single premise, and a conclusion that requires a number of steps and premises. This difference disappears if the acceptability of the single premise of the first agument is no greater than that of the <u>conjunction</u> of all the premises in the second argument. A purely probabilistic rule of acceptance automatically reflects this fact.

5. <u>Conclusion.</u>

It is important to distinguish probabilistic reasoning from probabilistic inference. Probabilistic reasoning may concern the manipulation of probabilities in the context of decision theory, or it may involve the updating of probabilities in the light of new evidence via Bayes' theorem or some other procedure. Both of these operations are essentially <u>deductive</u> in character.

Contrasted with these procedures of manipulating or computing with probabilities, is the use of probabilistic rules of inference: rules that lead from one sentence (or a set of sentences) to another sentence, but do so in a way that need not be truth preserving. One could attempt to get along without probabilistic inference, but it would be very difficult.

Instances of such rules are those represented by circumscription, non-monotonic logic, default rules, etc. as well as several classes of inference rules associated with statistics, and some rules discussed by philosophers.

The simplest probabilistic rule of inference -- a high probability rule -- has some curious consequences, but it does not seem that these consequences need interfere with the useful application of the rule.

Bibliography

Birnbaum, Alan (1969) "Concepts of Statistical Evidence," in Morgenbesser et al (eds) *Philosophy Science and Method*, St. Martin's Press, New York.

de Kleer, J. (1986) An Assumption Based ATMS," *A I Journal* **28**, 1986, 127-162.

Ginsberg, Matthew (1985), "Does Probability Have a Place in Non-monotonic Reasoning?", *IJCAI 1985*, Morgan Kaufmann, Los Altos, 1985, 107-110.

Israel, David; "What's Wrong with Non-Monotonic Logic?" *Proceedings of IJCAI*, Stanford, 1980, 99-101.

Kneale, W., and Kneale, M., (1962) *The Development of Logic*, Oxford University Press.

Kyburg, Henry (1961) *Probability and the Logic of Rational Belief*, Wesleyan University Presss.

------, (1974) *The Logical Foundations of Statistical Inference*, Reidel.

------, (1987) "Bayesian and Non-Bayesian Evidential Updating;" *A I Journal* **31**, pp 271-294.

Lehman, E. H. (1959); *Testing Statistical Hypotheses*, John Wiley and Sons, New York.

Levi, Isaac (1967); *Gambling with Truth*, Knopf, New York.

---, (1980); *The Enterprise of Knowledge*, MIT Press, Cambridge.

Nilsson, N.J.; "Probabilistic Logic," *Artificial Intelligence* 28, 1986, 71-88.

Pearl, Judea (1986); "Fusion, Propagation, and Structuring in Belief Networks," *A I Journal* **29**, 1986, 241-288.

Ramsey, F. P. (1950) *The Foundations of Mathematics and Other Essays*, Humanities Press, New York.

Schafer, Glenn, *A Mathematical Theory of Evidence*, Princeton, 1976.