HOW TO DO WITH PROBABILITIES WHAT PEOPLE SAY YOU CAN'T

Judea Pearl

September 1985
CSD-850031

Technical Report
CSD-850031
R-49
Revision I
September 1985

HOW TO DO WITH PROBABILITIES WHAT PEOPLE SAY YOU CAN'T*

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California
Los Angeles, CA 90024
(judea@UCLA-locus)
(judea@LOCUS.UCLA.EDU)

To be presented at the Second Conference
on Artificial Intelligence Applications
December 11-13, 1985, Miami, Florida

*This work was supported in part by the National Science Foundation, Grant #DCR 83-13875.

HOW TO DO WITH PROBABILITIES WHAT PEOPLE SAY YOU CAN'T

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California
Los Angeles, CA 90024

ABSTRACT

Ever since McCarthy and Hayes proclaimed probabilities "epistemologically inadequate" for reasoning with partial beliefs, research in this area has consisted primarily of nonnumerical approaches, attempting to enrich first-order logic with modal operators that capture the notions of default, likelihood, and knowledge. This paper addresses the problem from the opposite extreme: devising new representations to probabilistic models that emphasize the qualitative aspects of the reasoning process and minimize its sensitivity to numerical inputs. We find that, although numbers per se are bad summarizers of implicit knowledge, they can be very useful in processing that which has been explicated. Probabilistic networks of conceptually related propositions, where the numbers serve to regulate and propel the flow of information, allow reasoning about uncertainty to be as knowledge-intensive, accurate, and psychologically plausible as the level of details we care to explicate.

The paper describes a mechanism for maintaining and propagating beliefs in such networks, which facilitates concurrent, distributed, and coherent inferences, and fully conforms to the axioms of probability theory. Using this mechanism as a model of reasoning we find that many arguments against the use of probabilities are no longer valid, while others expose a core of problems that must eventually be confronted by every formalism of partial beliefs.

1. INTRODUCTION

Probability theory is shunned by most researchers in Artificial Intelligence. New calculi, claimed to better represent human reasoning under uncertainty, are being invented and reinvented at an ever-increasing rate. A major reason for the emergence of this phenomenon has been the objective of making reasoning systems *transparent*, i.e., capable of producing *psychologically meaningful* explanations for every intermediate step used in deriving the conclusion. Admittedly, traditional probability theory has erected cultural barriers against meeting this requirement. For example, scholarly textbooks on probability theory create an impression that to construct an adequate representation of probabilistic knowledge we must literally start by defining a *joint distribution function* $P(x_1, \ldots, x_n)$ on all propositions and their combinations, and that this function should serve as the sole basis for all inferred judgments. As a result, even simple tasks such as computing the impact of an evidence e on a hypothesis h via

$$P(h \mid e) = \frac{P(h, e)}{P(e)} = \frac{\sum_{x_i \neq h, e} P(x_1, \ldots, x_n)}{\sum_{x_i \neq e} P(x_1, \ldots, x_n)}$$

appear to require a horrendous number of meaningless arithmetic operations, unsupported by familiar mental processes. Another example is the striking disparity between traditional numerical definitions of independence (e.g. $P(h, e) = P(h) \cdot P(e)$) and the ease and conviction with which people identify conditional independencies, being so unwilling to provide precise numerical estimates of probabilities.

However, other representations of uncertain knowledge are available, which provide a more faithful model of human reasoning, and still comply with the basic tenets of probability theory. *Dependency-graph* representations, in which the links signify direct probabilistic dependencies among semantically-related propositions, are the most appealing candidates because they are robust to numerical imprecisions. They permit people to express essential qualitative relationships and preserve them despite sloppy assignment of numerical estimates. An integral part of dependency-graph models of reasoning is the assumption that the basic steps invoked while people query and update their knowledge correspond to local mental tracings of links in these graphs and this, in turn, determines what kind of operations people consider "psychologically meaningful". *Bayesian networks* offer an effective formalism for describing and controlling such graph operations.

Section 2 summarizes the properties of Bayesian networks and of a Belief Maintenance System (BMS) that performs inferences within such networks [Pearl, 1985a]. The impact of each new evidence is viewed as a perturbation that propagates through the network via local communication among neighboring concepts. We show that in reasonably sparse networks such autonomous propagation mechanism can support both predictive and diagnostic inferences, that it is guaranteed to converge in time proportional to the network's diameter, and that every proposition is eventually accorded a measure of belief consistent with the axioms of probability theory.

Section 3 shows that the current trend of abandoning probability theory as the standard formalism for managing uncertainty is grossly premature--taking graph propagation as the basis for probabilistic reasoning nullifies most objections against the use of probabilities in reasoning systems. For example, the graph representation allows us to:

1. Construct consistent probabilistic knowledge-bases without collecting "massive amounts of data".

2. Admit judgmental evidence at any level of abstraction.

3. Ensure that evidence in favor of a hypothesis not be construed as partially supporting its negation.

4. Postpone judgement.

5. Distinguish between various types of uncertainty.

6. Trace back the sources of beliefs and produce sound explanations.

7. Optimize the acquisition of data.

2. BELIEF MAINTENANCE USING PROBABILITIES

2.1 Bayesian Networks

Bayesian Networks are directed acyclic graphs in which the nodes represent propositions (or variables), the arcs signify the existence of direct causal influences between the linked propositions, and the strengths of these influences are quantified by conditional probabilities (Figure 1).

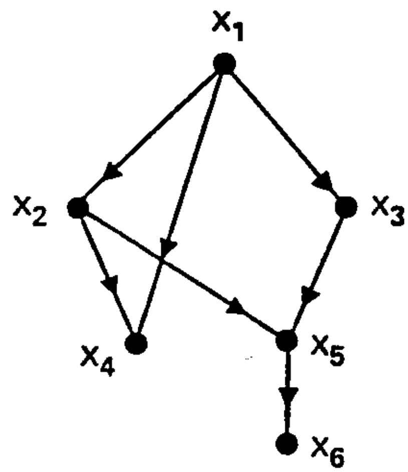

Figure 1

Thus, if the graph contains the variables $x_1, \ldots, x_n$, and S_i is the set of parents for variable x_i, then a complete and consistent quantification can be attained by specifying, for each node x_i, an assessment $P'(x_i \mid S_i)$ of $P(x_i \mid S_i)$. The product of all these assessments,

$$P(x_1, \ldots, x_n) = \prod_i P'(x_i \mid S_i) \tag{1}$$

constitutes a joint-probability model which supports the assessed quantities. That is, if we compute the conditional probabilities $P(x_i \mid S_i)$ dictated by $P(x_1, \ldots, x_n)$, the original assessments are recovered. Thus, for example, the distribution corresponding to the graph of Figure 1 can be written by inspection:

$$P(x_1, x_2, x_3, x_4, x_5, x_6) = P(x_6 \mid x_5)\, P(x_5 \mid x_2, x_3)\, P(x_4 \mid x_1, x_2)\, P(x_3 \mid x_1)\, P(x_2 \mid x_1)\, P(x_1).$$

An important feature of Bayes network is that it provides a clear graphical representation for the essential independence relationships embedded in the underlying model. The criterion for detecting these independencies is based on *graph separation*: namely, if all paths between x_i and x_j are "blocked" by a a subset S of variables, then x_i is independent of x_j given the values of the variables in S. Thus, each variable x_i is independent of both its siblings and its grandparents, given the values of the variables in its parent set S_i. A path is "blocked" if it contains an instantiated variable between two diverging or two cascaded arrows. A different criterion holds for converging arrows: the connection between two arrows converging at node x_k is normally "blocked", unless x_k or any of its descendants is instantiated. In Figure 1, for example, x_2 and x_3 are independent given $S_1 = \{x_1\}$ or $S_2 = \{x_1, x_4\}$, because the two paths between x_2 and x_3 are blocked by either one of these sets. However, x_2 and x_3 may not be independent given $S_3 = \{x_1, x_6\}$; because x_6, as a descendant of x_5, "unblocks" the head-to-head connection at x_5, thus opening a pathway between x_2 and x_3.

2.2 Belief Propagation in Bayesian Networks

Once a Bayesian network is constructed, it can be used to represent the generic causal knowledge of a given domain, and can be consulted to reason about the interpretation of specific input data. The interpretation process involves instantiating a set of variables E corresponding to the available evidence and calculating its impact on the probabilities of a set of variables H designated as hypotheses. In principle, this process can be executed by an external interpreter who may have access to all parts of the network and may use its own computational facilities to store and manipulate intermediate results. An extreme example would be to calculate $P(H \mid E)$ using the ratio definition $P(H, E)/P(E)$ (see Introduction). However, the sequence of steps followed by such an interpreter would seem foreign to human reasoning, and would not be defensible by a psychologically meaningful explanation.

A more transparent interpretation process results when we restrict the computation to take place at the *knowledge level itself*, not external to it. That means that the links in the network are the only pathways and activation centers that direct and propel the flow of data in the process of querying and updating beliefs. Accordingly, we imagine that each node in the network is designated a separate processor which both maintains the parameters of belief for the host variable and manages the communication links to and from the set of neighboring, logically related, variables. The communication lines are assumed to be open at all times, i.e., each processor may at any time examine the messages received from its neighbors and compare them to its own parameters. If the compared quantities satisfy some local constraints, no activity takes place. However, if any of these constraints is violated, the responsible node is activated to revise its violating parameter and transmit new messages to its neighbors. This, of course, will activate similar revisions at the neighboring nodes and will set up a multidirectional propagation process, until equilibrium is reached.

The fact that evidential reasoning involves both top-down (predictive) and bottom-up (diagnostic) inferences has caused apprehensions that, once we allow the propagation process to run its course unsupervised, pathological cases of instability, deadlock, and circular reasoning will develop [Lowrance, 1982]. Indeed, if a stronger belief in a given hypothesis means a greater expectation for the occurrence of its various manifestations and if, in turn, a greater certainty in the occurrence of these manifestations adds further credence to the hypothesis, how can one avoid infinite updating loops when the processors responsible for these propositions begin to communicate with one another asynchronously?

The key to maintaining stability in bi-directional inference systems lies in storing with each proposition an explicit record of the sources of its belief. Thus, in addition to its measure of total belief, each proposition also maintains a list of parameters, called *support list*, each representing the degree of support that the host proposition obtains from one of its neighbors.

2.3 Maintaining The Support List

The problems associated with asynchronous propagation of beliefs, have simple solutions if the network is singly connected, namely, if there is one underlying path between any pair of nodes. These include trees, where each node has a single parent [Pearl, 1982], as well as graphs with multi-parent nodes, representing events with several causal factors [Kim and Pearl, 1983]. We shall first describe the propagation scheme in singly connected networks and then show how it can be modified to handle loops.

Consider a fragment of a singly connected Bayesian network, as depicted in Figure 2. Let variable names be denoted by capital letters, e.g., A, B, X, Y, and their associated values by subscripted lower case letters, e.g., $a_1, a_2 \cdots$. With each variable A we store the following three parameter lists:

1. $P(A \mid B, C)$ -- The fixed conditional probability matrix which relates the variable A to its immediate causes (parents).

2. $\pi_A(B)$ -- The current strength of *causal* (or *prospective* support, contributed by each in-

coming link, e.g., from B to A, where:

$$\pi_A(B) = P(B = b_j \mid \text{evidence connected to } A \text{ via } B) \qquad j=1,2,... \tag{1}$$

3. $\lambda_X(A)$ -- The current strength of *diagnostic* (or *retrospective*) support contributed by each outgoing link, e.g., from A to X, where:

$$\lambda_X(A) = P(\text{evidence connected to } A \text{ via } X \mid A = a_i) \qquad i=1,2,... \tag{2}$$

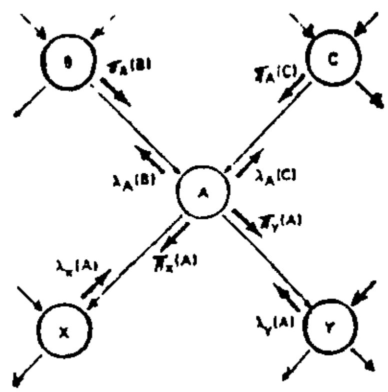

Figure 2

The absolute magnitudes of the elements in each of the λ vectors are arbitrary; only their ratios count. In the case of bivalued (propositional) variables, only a single parameter, the likelihood ratio, is needed.

The π–λ parameters are delivered to each node by the corresponding neighbors and are sufficient for calculating the current belief over the values of A:

$$\text{Bel}(A) = P(A = a_i \mid \text{all evidence})$$
$$= f_b\,[P(A \mid B, C), \lambda_X(A), \lambda_Y(A), \pi_A(B), \pi_A(C)]. \tag{3}$$

Similarly, the π–λ parameters stored at A are sufficient for calculating the appropriate messages (also π–λ parameters) that A should deliver to its corresponding neighbors. These calculations involve all the stored parameters except the one obtained from the port receiving the message.

For example:

$$\lambda_A(B) = f_\lambda[P(A \mid B, C), \lambda_X(A), \lambda_Y(A), \pi_A(C)], \tag{4}$$

$$\pi_X(A) = f_\pi[P(A \mid B, C), \lambda_Y(A), \pi_A(B), \pi_A(C)]. \tag{5}$$

The combining functions f_b, f_λ, and f_π involve only inner products and component-by-component products [Kim, 1983].

The impact of new evidence propagates through the network by uniform local computations which may be concurrent, asynchronous, or activated by some goal-oriented strategy. Upon receiving an activation signal, each processor examines the π–λ parameters stored, then recomputes and transmits the π–λ messages for its neighbors. Eqs. (4) and (5) demonstrate that a perturbation of the causal parameter, π, will not affect the diagnostic parameter, λ, on the same link and vice versa. The two are orthogonal to each other since they depend on two disjoint sets of data. Therefore, any perturbation of beliefs due to new evidence propagates through the network and is absorbed at the boundary without reflection. A new equilibrium state is reached after a finite number of updates which, in the worst case, is equal to the diameter of the network.

This architecture lends itself naturally to hardware implementation capable of real-time interpretation of rapidly changing data. It also provides a reasonable model of neural nets involved in cognitive tasks such as visual recognition, reading comprehension, and associative retrieval, where unsupervised parallelism is an uncontested mechanism.

Special provisions are necessary to support propagation in networks containing loops (like the one in Figure 1), where parents of common children also possess common ancestors. If we ignore the existence of loops and permit the nodes to continue communicating with each other as if the network was singly-connected, it will set up messages circulating indefinitely around the loops and the process will not converge to a coherent equilibrium.

The method that we found most promising, called *conditioning* [Pearl, 1985b], is based on the ability to change the connectivity of a network and render it singly connected by instantiating a selected group of variables. In Figure 1, instantiating x_1 to some value would block the pathway x_2, x_1, x_3 and would render the rest of the network singly connected, where the propagation techniques of the preceding paragraphs are applicable. Thus, if we wish to propagate the impact of an observed data, say at x_6, to the entire network, we first assume $x_1 = 0$, propagate the impact of x_6 to the variables $x_2, \ldots, x_5$, repeat the propagation under the assumption $x_1 = 1$ and, finally, linearly combine the two results weighed by the posterior probability $P(x_1 | x_6)$. It can also be executed in parallel by letting each node receive, compute, and transmit several sets of parameters, one for each value of the conditioning variable. This mode of propagation is not foreign to human reasoning. The terms "hypothetical" or "assumption-based" reasoning, "reasoning by cases," and "envisioning" all refer to the same basic mechanism of selecting a key variable, binding it to some of its values, deriving the consequences of each binding separately, and integrating these consequences together.

3. PROBABILITY IN SELF DEFENSE

3.1 Constructing Consistent Probabilistic Knowledge-Bases Without Collecting "Massive Amounts of Data"

The stigma of "requiring massive amounts of data" has remained with probabilistic models since the dark ages of statistical tyranny, when probability was perceived to be primarily a measure of relative frequency, supported solely by statistical tests. With the advent of the subjective views of probability and the increased awareness in the informational value of expert opinion, the charge has been somewhat reduced. It was replaced by "requiring an overwhelmingly large number of expert judgments" which, practically speaking, may be as unattainable. Indeed, if one takes the view that every variable in the Bayes network may depend on all the other variables, a complete graph ensues, and the number of parameters required for quantifying a graph with n binary variables is $O(2^n)$.

Fortunately, people do not perceive the world to be that cruel and unmanageable but, rather, adapt internal representations which constitute workable approximations in certain domains. No expertise could otherwise be developed. Thus, the question of whether probability theory is appropriate for capturing human expertise depends not on whether the world actually complies with the approximations made by the probabilistic model builder, but rather on whether the language of probability permits the model builder to express his/her approximations in a natural and consistent way.

The incremental process by which Bayesian networks are constructed is ideal for this modelling task. It permits people to express explicitly those qualitative relationships perceived to be essential, and helps preserve these qualities *despite* the sloppy assignments of numerical estimates. The addition of any new node x_i to the network only requires that the expert identify a set S_i of variables, which "directly bear" on x_i, quantify the strength of this local relation, and make no commitment regarding the effect of x_i on other variables, outside S_i. Even though each judgment is preformed locally, their sum total is guaranteed to be consistent.

In many domains the resulting graphs are sparse, (i.e., each variable having a small number of parents), because people conceptualize causal relationships by forming hierarchies of small clusters of causal factors. But even in cases where the number of parents k is sizable, estimating $P(x_i \mid S_i)$ usually requires less than the theoretical number of 2^k parameters. This is because the interactions among the factors in each cluster are normally perceived to fall into one of a few prestored, prototypical structures, each requiring about k parameters. Common examples of such prototypical structures are: noisy OR gates (i.e., any one of the factors is likely to trigger the effect), noisy AND gates, and various enabling mechanisms (i.e., factors identified as having no influence of their own except enabling other influences to become effective).

Admittedly, when an interaction is so complex that it cannot be approximated by any of the restored prototypes, then 2^k parameters are required for specifying $P(x_i \mid S_i)$. In such cases, the same number of parameters would also be needed in any uncertainty formalism, no matter how clever (unless, of course, it serves to hide the existence of other approximations).

3.2 Admitting Judgmental Evidence at All Levels of Abstraction

Imagine having constructed a complex Bayes network representing all factors and beliefs in an intricate murder trial, then a pathological report R arrives stating that "there is probably an 80% chance that the victim was indeed murdered". How can this judgmental statement be incorporated into the existing knowledge base, assuming that the network already contains the proposition h = "the victim was murdered", but does not contain any of the pathological findings upon which the report was based. The difficulty in admitting R stems from the need to reconcile two apparently conflicting statements: the evidence obtained prior to R may have imparted on h a belief $P_1(h)$ while the report R states $P_2(h) = .80$. What should the updated value of $P(h)$ be, and how should it affect the rest of the network?

This difficulty was one of the motivations behind the development of the Certainty Factors (CF) formalism [Shortliffe and Buchanan 1975] and made the designer of PROSPECTOR [Duda *et al.*, 1976] resort to various interpolation techniques. The Bayesian solution to this problem is rather simple: the impact of new evidence e on a proposition h inside the network should not be expressed in terms of absolute probability but, rather, as a likelihood ratio.

$$\lambda = \frac{P(e \mid h)}{P(e \mid \neg h)}$$

The reason is that in order to properly assimilate an absolute-probability report we must know exactly how much of the knowledge contained in the network (e.g., crime rate information, witness testimonies, etc.) was also consulted in the preparation of that report. The likelihood ratio λ, on the other hand, being a truly local relation between e and h, is independent of any such consideration.

If the report is phrased as a likelihood ratio statement, we simply incorporate it by adding a new link pointing from h to a "virtual" variable e, quantify the link by λ, and propagate its impact as if e was a confirmed propositional statement. If the report is phrased in terms of absolute probabilities we must extract from it the value of λ by either asking the reporter to reveal his starting probability $P(h)$ prior to observing e, or by assuming that the reporter starts the observations at some standard prior, say $P(h) = 1/2$. In our case, this assumption leads to

$$\lambda = \frac{.8}{.2} = 4.$$

It is interesting to note that an identical assumption has tacitly been incorporated into the calculus of certainty factors if one interprets CF to stand for $(\lambda-1)/(\lambda+1)$ [Heckerman, 1985; Grosof, 1985].

3.3 Ensuring That Evidence in Favor of a Hypothesis Would Not Be Construed as Partially Supporting Its Negation

Such apprehensions were part of the desiderata in the development of the CF model. Shortliffe & Buchanan [1976] observed that an expert who provides the rule $R: e \Longrightarrow h(0.7)$ "may well agree that $P(h \mid e) = 0.7$, but he becomes uneasy when he attempts to follow the logical conclusion that therefore $P(\text{not } h \mid e) = 0.3$". The expert claimed that the observations were evidence (to degree 0.7) in favor of h and should not be construed as evidence (to degree 0.3) against h.

This paradox, like the problem of the preceding section, stems from an attempt to give the rule $R: e \Longrightarrow h$ (0.7) an absolute probability interpretation instead of extracting from it likelihood-ratio information. Probability theory dictates that if the rule R is to be treated as a stable, modular relationship between e and h, invariant to other information in the system, then it may only convey likelihood-ratio information. The posterior probability $P(h \mid e)$, by contrast, is also sensitive to the prior probability $P(h)$ just before observing e. Accordingly, the Bayesian practitioner will attempt to explicate the meaning of R in terms of λ, hoping to uncover the invariant relations intended by the expert.

The reason that the example seems paradoxical is that by the phrase: "evidence in favor of a hypothesis", we expect to see an *increase* in the probability of the hypothesis from $P(h)$ to $P(h \mid e)$ (with $P(h \mid e) > P(h)$). On the other hand, viewing the absolute probability $P(\text{not } h \mid e) = 0.3$ as a fixed property of the rule R somehow conveys the false picture that $P(\text{not } h)$ should increase by some positive factor no matter what its initial value was. No expert would construe a dramatic drop of $P(\text{not } h)$ from, say, 0.99 to 0.3 as *supporting* the negation of h [Cheeseman,1985]; however, an increase from $P(\text{not } h) = 0.01$ to 0.3 would rightly be perceived as contrary to the spirit of the rule R. The likelihood ratio formulation has a built-in protection against such confusion because it conveys only *change* information; evidence in favor of h (i.e., $\lambda > 1$) will always produce $P(h \mid e) > P(h)$ while evidence opposing h will be characterized by $\lambda < 1$ and result in $P(h \mid e) < P(h)$.

3.4 Postponing Judgment

People often perceive of a piece of evidence as supporting a set of hypotheses S without providing any information concerning the relative likelihood of the individual hypotheses in the set. The need to express an increase of belief in S, while postponing judgment regarding S's constituents, a major force in the development of the theory of Belief Functions (BF) [Shafer, 1976], has been cited repeatedly as a unique feature of the BF approach [Barnett, 1981] [Gordon and Shortliffe, 1985] and, unfortunately, has led people to ignore the fact that probability theory proper is equipped with identical, if not superior, capabilities.

In Sections 3.1 and 3.2, we emphasized the fact that in the Bayes-network formulation the statement "evidence e bears directly on h" simply means that e is connected to the network via a single link, between h and e. Thus to specify the effect of e on the entire knowledge base, the expert need only quantify the relation between e and h (using the likelihood ratio) but, otherwise, make no judgment whatsoever regarding other propositions in the system. It is hard to imagine a more extreme form of noncommitment.

Critics have sometimes claimed that probabilistic formalisms are incapable of representing the notions of "total ignorance" or "non-commitment" because an even distribution of probabilities among elements implies an uneven distribution among sets of these elements, and this, presumably, reflects more knowledge than one is willing to admit. The likelihood-ratio formulation, though, escapes this criticism and captures many of the properties we normally attribute to the notion of "no commitment". For example, if an evidence e imparts a likelihood-ratio λ to a set S of mutually exclusive hypotheses, then the probability of every proper subset S' of S should be modified by the same multiplicative factor. Therefore, if we take the view that the conversational utterance "committing weight" refers to deciding the factors by which the probabilities are to be modified, a different picture emerges. Unlike absolute probabilities which are additively conserved, the modifying factors (or weights) are not conserved; we may have a situation where each singleton hypothesis in S is accorded the weight w and, simultaneously, every set of hypotheses (in S) also draws the weight w, perfectly reflecting the notion of neutrality. Thus, equal distribution of likelihood-ratio weights amounts to giving complete freedom to other evidence (as well as to the evidence summarized by prior probabilities) to shape the final beliefs accorded the subsets of S.

3.5 Distinguishing Various Types of Uncertainty

Consider the following three statements [Raiffa, 1968]:

S_1 - I do not know anything about Baseball.

S_2 - Both teams are strong, and I know Baseball.

S_3 - Either team has equal chance (50%) of winning tomorrow's game.

S_1 and S_2 reflect totally different types of knowledge, yet when it comes to summarizing that knowledge in probabilistic terms they seem to be encoded by the same lifeless quantifier: 50%, as in S_3. Such realizations have caused many researchers to seek richer representations of uncertainty outside the framework of probability theory.

These attempts are equivalent to someone rejecting logic because the proposition "A is true" does not, in itself, tell you how the truth of A was established. Logic, however, does not prevent us from attaching to each proposition in the system a proof which substantiates its truth value, or, at least, pointers to the key steps in the proof. Doing so is a matter of implementation, but would not amount to rejecting logic as a guardian of consistency.

Similar considerations apply to the representation of uncertainty. A detailed description of the nature and origin of the uncertainty in a proposition Q can, but need not be attached directly to Q. A portion of the description may be stored with Q, another portion may reside in the neighboring propositions and yet another portion may be traced back to remote sections of the network. Once we have an effective mechanism of passing information through the network, it can be readily fetched from various locations and used to reconstruct the description whenever the need arises, and at whatever level of detail. The important thing is that all messages retain their probabilistic meaning, so that they can be manipulated in accordance with the consistency-guarding rules of probability calculus.

Let us be more specific about these possibilities. In our Baseball example, the network underlying S_1 would be totally different from that supporting S_2. Even though the total belief calculated for the proposition S_3 attains the value of 0.5 in both cases. The $\pi-\lambda$ parameters entering the calculations would be totally different, and these are stored explicitly, to attest the nature of the difference. In the case of S_1, Bel (S_3) would most likely be composed of $\pi = .5$ and $\lambda = 1$, indicating no evidential support. In the case of S_2, on the other hand, the proposition S_3 would most likely be imbedded in a rich network of arguments and indicators which balance each other out, namely, many of S_3's successors will contribute λ's greater than unity (favoring S_3) and many λ's smaller than unity (opposing S_3).

In light of the capability of Bayes networks to maintain and update an explicit list of source-identifying parameters (π and λ) it is hard to understand statements such as [Quinlan 1982]: "It is only in systems using a two-valued approach where the values are probabilities that there is a firm basis for detecting general inconsistency", [presumably because] "the single value combines the evidence for and against A without indicating how much there is of each". True, "the probability of A in the light of E might have the same value when no evidence in E is relevant to A as when E contains strong but counterbalancing arguments for A and against A". However, given that a Bayes network also maintains the $\pi-\lambda$ support list, it becomes extremely easy to tell these two situations apart.

3.6 Tracing Back the Origin of Belief to Produce Sound Explanations

The message-passing operations of BMS have a clear intuitive appeal and, therefore, can be directly translated into meaningful linguistic descriptions. For example, if proposition X (see Figure 2) delivers a message $\lambda_X'(A)$ to A that is significantly larger than the previous message $\lambda_X(A)$, it is an easy matter to generate the statement: "the belief in A has increased due to new evidence in favor of X". If the user insists on further identifying the evidence responsible for the change in Bel (A), it is possible to trace back the support parameters down to the bare data.

When a recommendation is finally issued by the system, it can be justified by a similar process, tracing the skeleton of an explanation subtree only along the parameters indicating substantial amounts of accumulated support. In general, such an explanation subtree may have disconnected components (i.e., a forest) indicating the existence of conflicting evidence which cancel each other's impact at higher levels of the graph. In such a case, it would be easy to coin the sentence: "Even though the occurrence of E_1 constitutes a strong indication for A, having observed E_2 sheds serious doubt on its significance in determining B".

The profile of the support list surrounding a given proposition can also be used to make a proper distinction between the terms: "*probable*", "*plausible*, and "*possible*". Event A is probable if Bel (A) is sufficiently high. "A is plausible" usually means that there are strong arguments in favor of A but not a strong evidence supporting it. Since the π parameters indicate the degree of causal support, and since causal dependence is normally associated with predictive arguments based on "first-principles", a plausible event will be characterized by high π's and low λ's. A "possible" event is one that has the potential of being confirmed by some conceivable but yet unobserved evidence. Such event can be readily identified either by simulating the effect of future observations, or by explicitly maintaining with each proposition a parameter indicating its potential for increased (or decreased) belief due to pending future evidence.

3.7 Optimizing the Acquisition of Data and Recommending Actions with Meaningful Guarantees

Techniques for accomplishing these two objectives are thoroughly discussed in the literature on Pattern-Recognition (under the category "feature-selection") and Decision Analysis [Howard, 1984], and will not be repeated here. It should be noted, though, that any formalism which supports these tasks must also permit inferences to flow both ways: from hypotheses to evidence as well as from evidence to hypothesis. In MYCIN, for example, where only the latter is activated, the control strategy cannot distinguish between a premise likely to be confirmed and a premise likely to be refuted; both are pursued with equal vigor. On the other hand, systems like PROSPECTOR [Duda *et al.*, 1976] and MEDAS [Ben-Bassat, 1980], both based on probability calculus, were successful in employing economical data-acquisition strategies.

CONCLUSIONS

Probability theory is known to have the following unique features: 1) The basic assumptions and proposed approximations are psychologically meaningful. 2) It supports bi-directional inferences. 3) It employs parameters that, at least in principle, can be submitted to empirical tests. 4) It provides prescriptive guidance for decisions, with meaningful guarantees. In this paper we examined whether probability calculus can also meet the computational objectives of intelligent reasoning systems:

1. Knowledge-based modularity

2. Intuitive transparency of the elementary inference steps

3. Flexibility of control

The Belief Maintenance System described is shown to meet these objectives. We suggest that the next uncertainty calculus the reader is tempted to adopt or to invent be subjected to similar examinations.

REFERENCES

[1] J.A. Barnett, "Computational methods for a mathematical theory of evidence", *Proceedings of the Seventh International Joint Conference on Artificial Intelligence* (Vancouver, B.C., Canada), pp. 868-875, 1981.

[2] M. Ben-Bassat, R.W. Carlson, V.K. Puri, E. Lipnick, L.D. Portigal, M.H. Weil, "Pattern-based interactive diagnosis of multiple disorders: The MEDAS system", *IEEE Trans. on Pattern Analysis & Machine Intelligence*, PAMI-2:148-160, 1980.

[3] P. Cheeseman, "In defense of probability", *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (UCLA, 18-23 August 1985), pp. 1002-1009, 1985.

[4] R.O. Duda, P.E. Hart, and N.J. Nilsson, "Subjective Bayesian methods for rule-based inference systems", *Proceedings of the 1976 National Computer Conference (AFIPS Conference Proceedings)*, 45, pp. 1075-1082, 1976.

[5] J. Gordon & E.H. Shortliffe, "A method of managing evidential reasoning in a hierarchical hypothesis space", *Artificial Intelligence*, 26 (3), pp. 323-358, July 1985.

[6] B.N. Grosof, "Evidential confirmation as transformed probability", *Proceedings of the Workshop on Uncertainty and Probability in Artificial Intelligence* (UCLA, 14-16 August 1985), pp. 185-192, 1985.

[7] D. Heckerman, "Probabilistic interpretations for MYCIN's certainty factors", *Proceedings of the Workshop on Uncertainty and Probability in Artificial Intelligence* (UCLA, 14-16 August 1985), pp. 9-20, 1985.

[8] R.A. Howard & J.E. Matheson, *Reading on the Principles and Applications of Decision Analysis*. Strategic Decisions Group. Menlo Park, CA. 1984.

[9] J. Kim, *CONVINCE: A CONVersational INference Consolidation Engine*. Ph.D. dissertation, University of California, Los Angeles, 1983.

[10] J. Kim and J. Pearl, "A computational model for combined causal and diagnostic reasoning in inference systems", *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pp. 190-193, 1983.

[11] J.D. Lowrance, "Dependency-graph models of evidential support", COINS Technical Report 82-26, University of Massachusetts at Amherst, 1982.

[12] J. Pearl, "Reverend Bayes on inference engines: A distributed hierarchical approach", *Proceedings of AAAI Conference on Artificial Intelligence* (Pittsburgh, PA), 1982.

[13] J. Pearl, "Fusion, propagation, and structuring in Bayesian networks", Report #CSD-850022, Cognitive Systems Laboratory, Computer Science Department, University of California, 1985a.

[14] J. Pearl, "A constraint-propagation approach to probabilistic reasoning", *Proceedings of the Workshop on Uncertainty and Probability in Artificial Intelligence* (UCLA, 14-16 August 1985), pp. 185-192, 1985b.

[15] J.R. Quinlan, "INFERNO: A cautious approach to uncertain inference", Rand Note N-1898-RC, September 1982.

[16] H. Raiffa, *Decision Analysis: Introductory Lectures on Choices Under Uncertainty.* Addison-Wesley. Reading, MA, 1968.

[17] G. Shafer, *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, NJ, 1976.

[18] E.H. Shortliffe and B.G. Buchanan, "A model of inexact reasoning in medicine", *Math. Biosc.* 23: 351-379, 1975.