DECISION SCIENCE CONSORTIUM, INC.

ALTERNATIVE THEORIES OF INFERENCE IN EXPERT SYSTEMS FOR IMAGE ANALYSIS

Technical Report 85-1

Marvin S. Cohen, Stephen R. Watson, and Eamon Barrett

Prepared for:

U.S. Department of the Army
Engineer Topographic Laboratories
ETL-CS-I
Fort Belvoir, Virginia 22060

Contract No. DACA72-84-C-0005

January 1985

Decision Science Consortium, Inc.
7700 Leesburg Pike, Suite 421
Falls Church, Virginia 22043
(703) 790-0510

The views, opinions, and findings contained in this report are those of the author(s) and should not be construed as an official Department of Defense position, policy, or decision, unless so designated by other official documentation.

TABLE OF CONTENTS

Page

1.0 INTRODUCTION..1-1

2.0 INFERENCE METHODS FOR EXPERT SYSTEMS............................2-1

2.1 Concepts of Uncertainty....................................2-3

2.1.1 Chance vs. imprecision.................................2-3
2.1.2 Chance vs. incompleteness..............................2-3
2.1.3 Imprecision vs. incompleteness.........................2-4

2.2 A Framework for Evaluating Theories of Uncertainty........2-4

2.2.1 Why a framework?.......................................2-4
2.2.2 Components of evaluation...............................2-7
2.2.3 What is validity?......................................2-8
2.2.4 Implications for knowledge engineering.................2-9

2.3 Current Status of Methods for Handling Uncertainty........2-10

2.3.1 MYCIN..2-11
2.3.2 Other developments.....................................2-12

2.4 Bayesian Probabilities....................................2-13

2.4.1 Using probability theory for inexact reasoning.....2-13
2.4.2 Feasibility: Quantity of inputs....................2-15
2.4.3 Computational tractability..........................2-16
2.4.4 Validity: Axiomatic derivation.....................2-17
2.4.5 Plausibility of instances...........................2-18
2.4.6 Semantics: Behavioral specification................2-21
2.4.7 Naturalness of inputs..............................2-23
2.4.8 Concepts of uncertainty............................2-24
2.4.9 Summary..2-25

2.5 Belief Functions..2-26

2.5.1 Nature of the theory...............................2-26
2.5.2 Dempster's rule....................................2-28
2.5.3 Feasibility: Quantity of inputs....................2-30
2.5.4 Computational tractability..........................2-31
2.5.5 Validity: Semantics................................2-31
2.5.6 Face validity......................................2-33
2.5.7 Concept of validity................................2-33
2.5.8 Shafer's canonical example.........................2-35
2.5.9 A Bayesian foundation for belief functions?........2-40
2.5.10 Role of the assumptions in constructing
an analysis..2-45
2.5.11 Plausibility of instances: Conflict
of evidence..2-46
2.5.12 What is "conflict of evidence"?...................2-52
2.5.13 Summary...2-54

TABLE OF CONTENTS (Continued)

Page

2.6 Fuzzy Set Theory....................................2-54

2.6.1 Nature of the theory....................................2-54
2.6.2 Applications of fuzzy set theory to inference......2-55
2.6.3 Fuzzy implication....................................2-55
2.6.4 Fuzzy probabilities....................................2-57
2.6.5 Feasibility....................................2-58
2.6.6 Validity....................................2-58
2.6.7 Semantics: Where do the numbers come from?........2-58
2.6.8 Inference: What are the appropriate connectives?....................................2-60
2.6.9 Plausibility of instances....................................2-61
2.6.10 Summary....................................2-61

2.7 Non-Monotonic Reasoning....................................2-62

2.7.1 Nature of the theory....................................2-62
2.7.2 Dependency-directed backtracking....................................2-62
2.7.3 Example of informal non-monotonic reasoning........2-66
2.7.4 Application of a non-monotonic system..............2-68
2.7.5 Feasibility....................................2-71
2.7.6 Face validity....................................2-71
2.7.7 Plausibility of instances: Conflicting evidence....................................2-72
2.7.8 Summary....................................2-77

3.0 THE NON-MONOTONIC PROBABILIST: AN APPLICATION OF BELIEF FUNCTIONS, FUZZY LOGIC, AND NON-MONOTONIC REASONING............3-1

3.1 Contrast Between Probabilistic and Qualitative Approaches to Conflict Resolution........................3-1

3.2 Functional Outline of a Proposed System: The Non-Monotonic Probabilist....................................3-2

3.2.1 Rule-based belief function module..................3-3
3.2.2 Non-monotonic reasoning as an embedding context....3-5
3.2.3 Belief functions as a controlling context for non-monotonic reasoning........................3-7

3.3 The Non-Monotonic Probabilist as an Interactive System....................................3-9

3.4 Fuzzy Measures....................................3-11

3.4.1 Conflict....................................3-12
3.4.2 Support lists....................................3-13
3.4.3 Assumptions....................................3-15
3.4.4 Foundations....................................3-15

TABLE OF CONTENTS (Continued)

Page

3.4.5 Suppositions....................................3-16
3.4.6 Dependency-directed backtracking...................3-16
3.4.7 Conflict as the control over revision..............3-17

3.5 Conclusion...3-17

4.0 SUMMARY AND PROSPECTS...................................4-1

4.1 The Requirement for a Non-Monotonic Probabilist...........4-1
4.2 Main Results..4-2
4.3 Next Steps..4-4

APPENDIX A: APPLICATION OF ALTERNATIVE INFERENCE THEORIES TO PROBLEMS OF IMAGE UNDERSTANDING

References

1.0 INTRODUCTION

In recent years expert systems have been designed to replicate human reasoning in an increasing sphere of inference and decision-making tasks (Hayes-Roth et al., 1983; Buchanan and Duda, 1982). Expert systems have now been developed for medical diagnosis and treatment (e.g., Shortliffe, 1976), geological exploration (e.g., Duda et al., 1979), chemical analysis (Lindsay et al., 1980), military planning (Engelman et al., 1979), and other areas of specialized human skill.

In other areas, however, such as image analysis, the infiltration of expert system techniques has been relatively slow. One reason, at least, is that predominantly mathematical or statistical methods appear to be appropriate for such tasks as filtering or pattern matching against pixel data. The result has been a failure thus far to integrate satisfactorily such "bottom up" methods with requirements that promise to be more adequately met by expert system technology: e.g., the incorporation of intelligence information or explicit general knowledge in the process of image analysis and image understanding, and the resolution of conflicts between alternative sources of evidence or analysis (cf., Rosenfeld, 1984).

The objective of our research has been to address this problem on both a theoretical and a practical plane. Our theoretical goals were:

- to explore the feasibility of developing improved mechanisms for expert system inference, and
- to provide a better general understanding of inference mechanisms for expert system applications.

In our subsequent effort, we have (a) developed a heuristic framework for the evaluation, selection, and/or design of inference methods in expert systems; (b) critically scrutinized, within that framework, a variety of alternative schemes for handling uncertainty--those associated with Bayes, Shafer, Zadeh, and nonmonotonic logic; and (c) identified shortcomings and recommended modifications or extensions of those technologies. A major thrust of this part of our work is that requirements exist within expert system technology itself which will (or should) drive it toward a closer accommodation with mathematical or statistical methods; and, conversely, that the intelligent and flexible automation of probablistic methods will require techniques of qualitative reasoning traditionally associated with artificial intelligence. This work is reported in Section 2.0 below.

On the practical side, we have developed the high-level conceptual design of a new inference mechanism, incorporating and extending many of the findings of our theoretical work. This system, the Non-Monotonic Probabilist (NMP), utilizes Shaferian belief functions, fuzzy measures, and non-monotonic reasoning--where different concepts of uncertainty call for them. Probabilistic inference is embedded within a framework of qualitative reasoning which is in turn controlled by measures of the credibility of inferential argument. "Fuzzifying" these measures, in turn, ensures a simple but graded process of high-level control. Our work on this system has established the feasibility of a flexible and "intelligent" deployment of probabilistic methods in image understanding. This work is reported in Section 3.0 below.

To bridge the gap between theory and practice, we have developed and compared specific applications of Bayesian, Shaferian, and fuzzy methods to three representative problems in the field of image analysis: the incorporation of general knowledge or intelligence information, filtering and template matching, and "probabilistic relaxation." A description of this work is contained in Appendix A.

Finally, Section 4.0 summarizes the main line of argument leading to the development of NMP and describes the prospective application of a system like NMP.

2.0. INFERENCE METHODS FOR EXPERT SYSTEMS

In typical expert systems applications, the highest available standard of reasoning in the relevant area of knowledge is expert practice itself, rather than a formal theory, algorithm, or search technique. As a result, much of the effort in expert systems development consists in the extraction of relevant knowledge from human experts for translation into machine-usable form. A second consequence, whose importance is only now being fully understood, is the need to represent uncertainty, to implement processes of inexact reasoning, and to incorporate some form of "metaknowledge": i.e., knowledge about the strengths and weaknesses of the system's own knowledge base.

A variety of alternative frameworks now exist for representing and reasoning about uncertainty. Among the most prominent are Bayesian probability theory, belief functions (Shafer, 1976), and fuzzy set or possibility theory (Zadeh, 1965, 1972). There is also considerable interest in non-numerical methods of inexact reasoning, such as non-monotonic logic (Doyle, 1979). Uncertainty calculi of these types can contribute to a variety of expert system functions; for example: (1) to combine different items of evidence or lines or reasoning in drawing a conclusion; (2) to control the allocation of computational resources among different lines of reasoning or knowledge resources; (3) to generate requests for additional data or judgments from users; (4) to halt computations when acceptable results are obtained; and (5) to explain to users how a conclusion was arrived at and what its credibility is.

The selection of a framework for accomplishing these functions will also have an impact on knowledge acquisition. The choice of such a framework will help structure the dialogue between knowledge engineer and domain expert, determining what questions are asked and how they are answered (cf., Shafer and Tversky, 1983). This process is seldom (if ever) the literal "transfer" of information, or rules, from expert to system. Much of the relevant knowledge is (as yet) unverbalized and only implicit in expert action and intuition. The value of frameworks for representing uncertainty must be assessed in part, therefore, by the way they influence the quality and quantity of the information an expert provides (Cohen, Mavor, and Kidd, 1984).

Unfortunately, there has as yet been little systematic research on the impact of alternative inference frameworks either on knowledge acquisition or on expert system functioning. In part, this can be attributed to the pragmatic urgency of getting systems up and running. In part, it may be due to a bias against numerical methods in the artificial intelligence tradition (as noted by Shafer, 1984a). Finally, however, it may be due to a set of real methodological obstacles. For example:

(1) Alternative frameworks for uncertainty differ in the degree to which appropriate normative justifications have been achieved; they differ also in the demands they impose on the expert for assessments, in the computational burden they impose on the system, and in the ease with which they represent distinctions and yield conclusions which are natural to a particular expert or user. Evaluation, in short, must be multidimensional. But it is by no means clear how tradeoffs among these competing considerations should be resolved.

(2) The theories themselves are in a process of evolution. To some extent, the success of an application depends on the ingenuity of the developer as much as on the intrinsic worth or potential of the theory.

(3) Alternative frameworks often appear to differ in the concept, or kind, of uncertainty which they attempt to capture (e.g., chance, imprecision, or completeness of evidence). On the other hand, defenders of each theory tend to regard the other theories, in some instances, as special cases of their own, and in other instances as invalid. Thus, it is seldom clear whether these theories are best regarded as competitors or as alternative tools with different, but complementary functions.

These three methodological challenges will be a recurring focus of Section 2.0. In Section 2.1 we amplify the notion that different concepts of uncertainty may be involved in expert reasoning, and in Section 2.2 we lay out a provisional multidimensional framework for evaluating alternative theories of inference and pinpointing areas in need of improvement. All this is by way of prelude to an examination of alternative systems of uncertainty in Sections 2.3 through 2.7.

2.1 Concepts of Uncertainty

How many different "kinds" of uncertainty or inexactness are there? The answer will depend on what theory (or theories) of uncertainty we ultimately choose to accept. Such a theory might derive a variety of apparently distinct notions from a single underlying principle. Nonetheless, on a more superficial plane, humans do seem to possess separate bodies of intuition, and abilities to make relatively independent judgments, concerning different sorts of uncertainty. These appear, moreover, to have different implications and roles in expert system design. Briefly delineating them will clarify what it is a theory of uncertainty could or should explain. We will distinguish among three notions:

- chance or uncertainty about the facts
- imcompleteness or quality of evidence
- imprecision or vagueness

2.1.1 Chance vs. imprecision.

The imprecision with which facts are specified is not the same as uncertainty about what the facts are. For example, the data provided by a digitized aerial photograph, consisting of a set of numbers representing gray levels at each pixel, are a precise set of data, but noise in the imaging process may make us uncertain what the "true" levels ought to be. Data such as "there is a long straight feature in the upper left of the photo" are imprecise, but entail no uncertainty. Similarly, an inference rule such as "if there is a rectangular object, then it is either a building or a field" is both an imprecise and an uncertain rule.

2.1.2 Chance vs. incompleteness.

Uncertainty about the facts is not the same as incompleteness of evidence. Consider the rule:

R1. If x is rectangular, it is a building with probability .9 or a field with probability .1.

This statement produces a high degree of certainty that x is a building, but it represents only a small portion of the obtainable evidence (viz., shape) which might bear on that question. Consider, on the other hand, the following rule:

> R2. If x is rectangular and far from a road, it is a building with probability .5 or a field with probability .5.

This statement covers more of the available evidence (i.e., shape and distance from a road), but yields a lower degree of certainty about the facts at issue.

2.1.3 Imprecision vs. incompleteness. Finally, imprecision and incompleteness of evidence are distinct. In the example above, R1 was imprecise, since x could be rectangular (and also perhaps a field or a building) to varying degrees. What if we obtain all possible data relevant to classifying x as a rectangle (i.e., a new set of very exact measurements of x's angles and sides)? Will we finally know for sure that x is or is not a rectangle? No (unless x turns out to be a perfect rectangle), since the imprecision in this example was the result of our ability to stretch the use of the term "rectangle", i.e., our willingness to tolerate a degree of deviation from perfection, not our lack of knowledge. Judgments of imprecision, in this sense, are more akin to judgments of similarity (e.g., of x to the "typical" rectangular object) than to judgments of the quality of evidence.

We conclude that there is at least a plausible case for distinguishing three notions of uncertainty. The remaining questions (to which we turn in later sections) are: (1) To what extent and in what way are each of these notions relevant to expert system design? (2) Can any of these concepts be successfully or naturally reduced to any of the others? (3) How successfully is each notion captured by current theories of uncertainty?

2.2 A Framework for Evaluating Theories of Uncertainty

2.2.1 Why a framework? Our discussion of strengths and weaknesses of alternative theories will largely be structured within the framework shown in Figures 2-1 and 2-2. The purposes of the framework are heuristic:

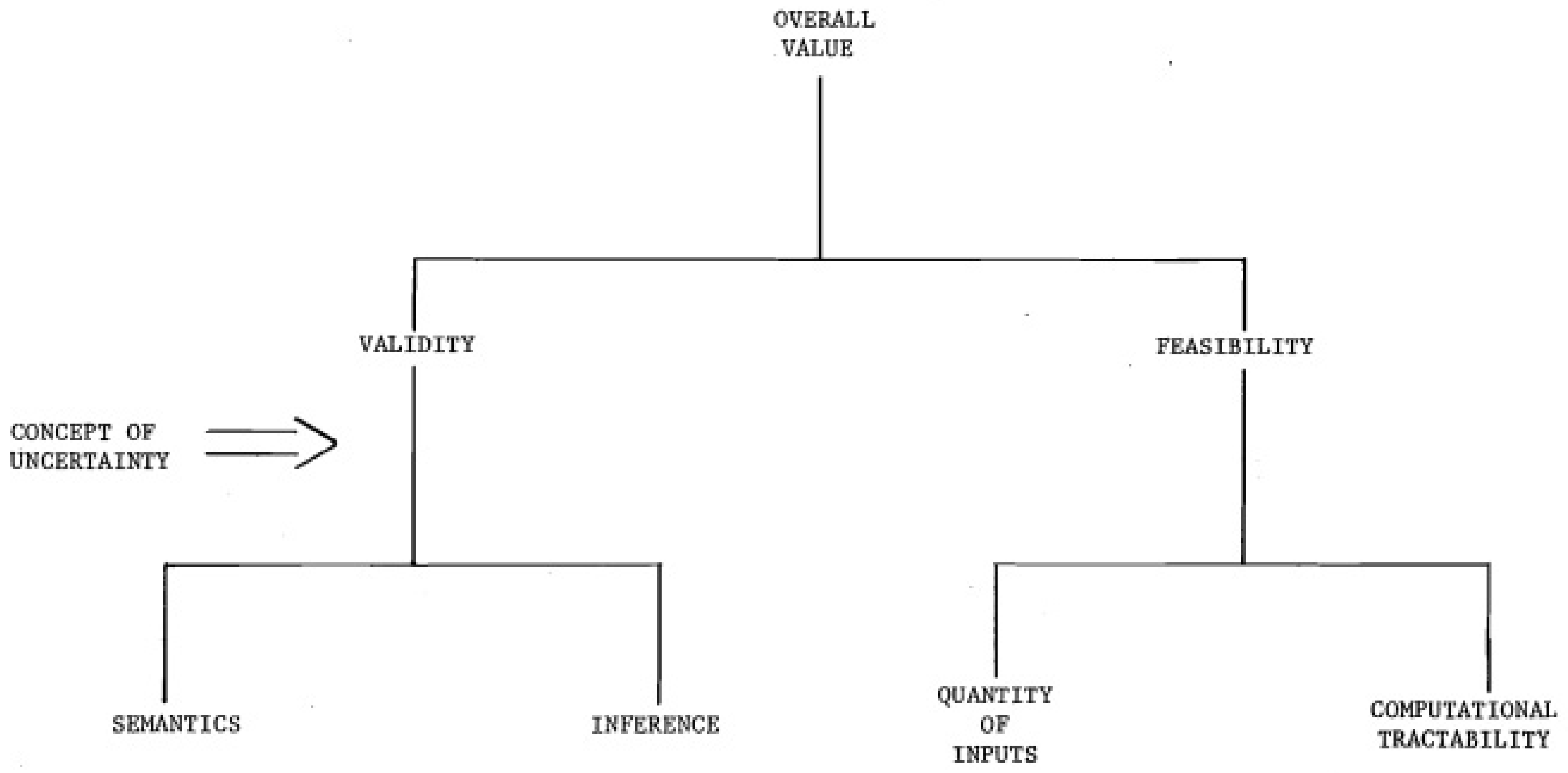

OVERALL VALUE
VALIDITY
FEASIBILITY
CONCEPT OF UNCERTAINTY
SEMANTICS
INFERENCE
QUANTITY OF INPUTS
COMPUTATIONAL TRACTABILITY

Figure 2-1

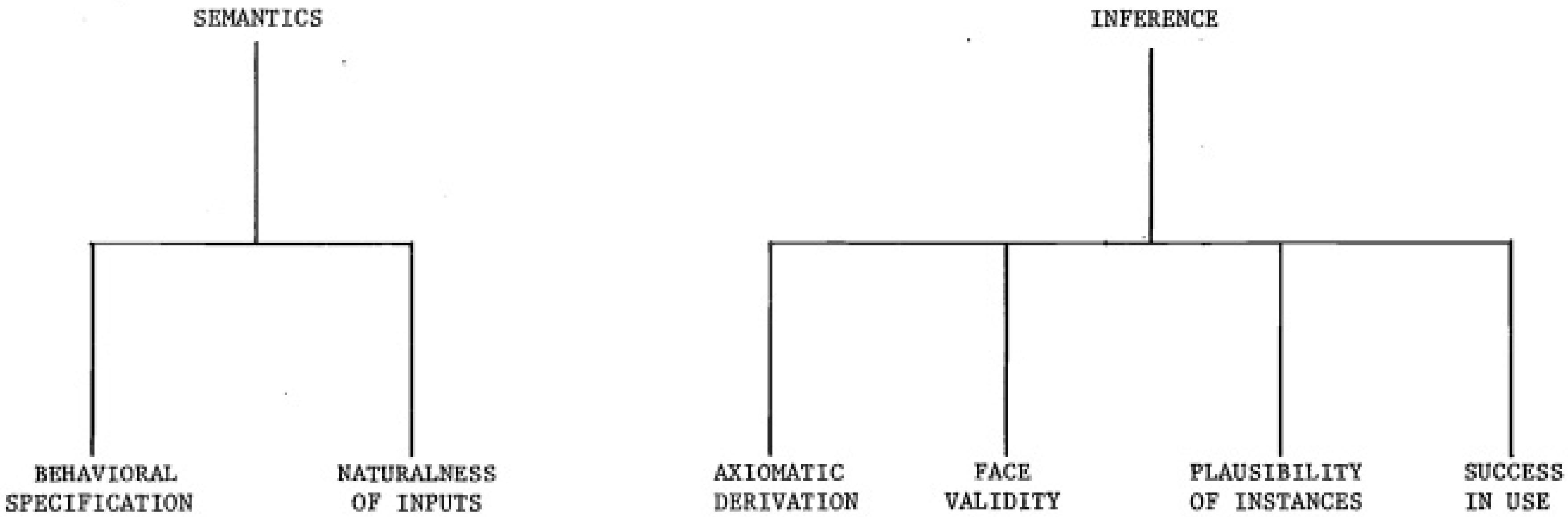

SEMANTICS
BEHAVIORAL SPECIFICATION
NATURALNESS OF INPUTS
INFERENCE
AXIOMATIC DERIVATION
FACE VALIDITY
PLAUSIBILITY OF INSTANCES
SUCCESS IN USE

Figure 2-2

- to clarify our understanding of the features involved in such an evaluation, their relationships, and the tradeoffs that must be resolved in the actual design of a system;

- to suggest directions for the modification of current methods, the development of new methods, or the synthesis of current methods, that remedy specific shortcomings while retaining existing advantages; and

- to serve (perhaps) as the eventual basis of a knowledge engineering tool for the design of inference methods in specific applications.

2.2.2 Components of evaluation. As shown in Figure 2-1, evaluative criteria fall under two main headings: validity and feasibility (corresponding roughly to benefits and costs). Under each of these are two subcategories which include factors relating to representation and reasoning, respectively. Thus, feasibility breaks down into the quantity of inputs required by the representation of uncertainty and the computational tractability of the reasoning process. Validity breaks down into the validity of the semantic representation and the validity of the process of inference or reasoning. "Concept of uncertainty" is an important conditioning parameter; i.e., the performance of a given theory of uncertainty on the various criteria included under validity will depend on the type of uncertainty which is appropriate to the application at hand.

Under validity, inference and semantics are further broken down into sets of more specific criteria, as shown in Figure 2-2. Each of these sets is a mix of formal and informal factors, i.e., criteria which seem purely mathematical or behavioral, on the one hand, and those which have a more cognitive or pragmatic aspect, on the other. Thus, under semantics, we indicate the desirability of an explicit behavioral specification for the required inputs. For example, if I assign a probability of .9 that x is a building, then according to Bayesian theory, I would be indifferent between a bet whose outcome depended on x's being a building and a bet on drawing a red ball from an urn containing 90 red and 10 black balls. As we shall see later in this section, alternative views of uncertainty have not had as much success in providing behavioral specifications for their inputs as has Bayesian probability theory. On the other hand, we also indicate under semantics the desirability that inputs take a form that is, in some sense, natural for the expert to provide. The unnaturalness of Bayesian inputs for many applications has been a strong selling point for theories attempting to supplant Bayesian probability theory.

Similarly, under inference, we include not only the existence of an axiomatic derivation, but also the face validity of the theory's basic postulates or rules, the plausibility of conclusions drawn by use of the theory in specific applications, and the successful achievement of goals by persons or systems which use the theory.

2.2.3 What is validity? The evaluation of inference frameworks in terms of "validity" has an inevitable air of circularity, since defenders of various alternative theories typically regard different sets of criteria as relevant. Thus, we had better comment on the concept of validity which is reflected in our choice of criteria. For example, Bayesians write as though only behavioral specification and axiomatic derivation mattered (e.g. Lindley, 1982), while defenders of alternative views tend to focus exclusively on the more cognitive or pragmatic criteria (e.g. Shafer, 1981). At the other extreme from the Bayesians, L. J. Cohen (1981) argues that only the conformity of a theory with actual instances of unaided human reasoning counts toward its validity (see commentary by M. S. Cohen, 1981). Thus, the range of criteria under validity can be regarded as defining a "political" spectrum from conservative to reform. (The non-Bayesians may regard themselves as the reformers since they oppose the "prevailing" Bayesian position on pragmatic grounds, but in a more meaningful sense the Bayesians are the reformers, since they advocate that many habitual ways of thinking be rejected as cognitive illusions.)

Our own position is that all the criteria are important. Our argument is simply that no deep or principled distinction can be made among them. An axiomatic derivation lends credibility to a theory to the degree that the axioms themselves, and the assumptions in the derivation, are found to be plausible, desirable, or applicable (cf., Shimony, 1970). This is only a difference in degree from the case where a theory lacks such a derivation, but where its basic postulates themselves have face validity or plausibility. Similarly, since accepting a theory entails acceptance of inferential conclusions drawn with its aid, there is no reason why the intrinsic plausibility of those conclusions, in specific instances, should not count for or against the plausibility of the theory. Finally, since we do not regard our intuitions regarding plausibility as infallible, we must allow actual success in using a framework to achieve our goals as an additional, though highly imperfect, indication of the overall plausibility of that framework. (Intuitions of plausibility in general may be the product of an evolutionary past comprising along series of actual successes and failures.) In sum, we regard all the criteria listed under validity as tools for enhancing the overall plausibility of our system of beliefs and, ultimately, our success in action. No one of them has a privileged status, and no one can be wholly ignored for other than arbitrary or ad hoc reasons.

2.2.4 Implications for knowledge engineering. There are two important corollaries of this view for the process of knowledge engineering. First, the customary distinction between replicating expert knowledge and devising an analytic, prescriptive, or statistical model cannot be regarded as a sharp one. Adoption of a particular inference framework is a process of "bootstrapping": prior intuitions and judgments (at the level of axioms, postulates, and/or specific inferences) determine the initial design of an inference mechanism; the output of that mechanism then may lead to the reconsideration and revision of previous intuitions and judgments with which it does not agree, or to redesign of the mechanism. Builders of expert systems have tended to put more weight on "capturing" an expert's pre-existing intuitions about specific instances than on the selection of inference schemes with globally plausible properties (i.e., axioms or postulates) which might lead to some revision in those intuitions. Note, however, that in other contexts, knowledge engineers do not hesitate to impose constraints on the format in which experts are asked to report their knowledge (cf., rule-based elicitation methods, such as EMYCIN; also the description of Nii's methods in Feigenbaum and McCorduck, 1983; Buchanan et al., 1983). By formulating his knowledge within these constraints, the expert himself may achieve new insights. We would argue that constraints imposed by theories of inference should be regarded in a similar light. (Cohen, Mavor, and Kidd, 1983, contains further discussion of this point.)

Some guidance, however, can be provided to the knowledge engineer in his initial selection of an inference framework. The discussion in Section 2.1 suggested that intuitions about uncertainty fall into three relatively separable sets, corresponding to different concepts of uncertainty. Thus, a proposed theory of uncertainty cannot be evaluated in the abstract; we must consider its plausibility with respect to the appropriate set of intuitions. This suggests the following approach to a methodology of knowledge engineering:

- prior determination (through use of an evaluation framework such as the one described above) of inference mechanisms which are well-suited for specific concepts of uncertainty,
- determination on the spot, for various components in a specific application, of the concept or concepts of uncertainty that are relevant.

Judgments relating components of a specific expert system application to different concepts of uncertainty would thus serve as a mediating link between that application and the initial selection or design of an inference mechanism. Note that determination of the relevant concept of uncertainty in a specific application may, in part at least, be a function of explicitly identifiable features of the application: for example, the generic problem type (e.g., diagnosis, estimation, classification, monitoring, or choice of actions) and generic interactive functions (e.g., interpretations of user queries and data inputs, display of conclusions and explanations to users, alerting with regard to real time events, requests for user judgments or data, and incorporation of user overrides or revisions into the knowledge base). Thus, general guidelines linking problem types and interactive functions to concepts of uncertainty might eventually be devised.

2.3 Current Status of Methods for Handling Uncertainty

If expert systems are to replicate the performance of experts in cognitive tasks, in almost all cases some method must be found that matches the human ability to carry out inexact reasoning. In the remainder of Section 2.0, we examine a variety of calculi to that end. We will focus far less on the details of the theories than (a) on their strengths and weaknesses in the various categories outlined in Section 2.2, and (b) on potential modifications, amplifications or syntheses to redress weaknesses. After briefly discussing MYCIN, we shall move on to Bayesian probabilities (Section 2.4), belief functions (Section 2.5), fuzzy sets (Section 2.6), and non-monotonic logic (Section 2.7). The major positive contribution of this review is that numerical calculi will not adequately capture the human ability to intelligently and flexibly manipulate uncertainties unless they are embedded in a higher-order system of qualitative reasoning. This thesis provides an essential basis for the new system of reasoning to be proposed in Section 3.0. A less technical description of the various theories themselves may be found in Cohen et al., 1984.

2.3.1. MYCIN. The developers of MYCIN, by far the most familiar and influential expert system, recognized the need for an uncertainty calculus and proceeded to invent their own (Shortliffe, 1976, Chap. 4). Based on Shortliffe's calculus of certainty factors, MYCIN has had a certain degree of pragmatic success. Unfortunately, its developers as well as others have recognized an increasing number of difficulties, especially in the area of validity (Buchanan and Shortliffe, 1984).

Feasibility: Shortliffe's calculus has been demonstrably successful in this area. The required number of inputs is kept to a minimum, since complex judgments of evidential interdependencies and prior probabilities are not elicited. Inference rules are computationally consistent with a highly modular, rule-based, backwards chaining architecture.

Validity: Semantics: An original goal of MYCIN was to provide a format for expert inputs with a natural interpretation, as the degree to which a bit of evidence "confirms" a conclusion. However, no behavioral specification for certainty factors has been offered. Moreover, even on an informal level, it is unclear whether experts can have a sufficient grasp of the meaning of the numbers they are asked to assess. For example, certainty factors confound different senses of uncertainty, as well as confounding uncertainty and the importance of the hypothesis under consideration.

Axiomatic derivation: MYCIN lacks any deep normative justification. Adams (1976) has shown, moreover, that MYCIN cannot be plausibly regarded as an approximation to Bayesian methods, as Shortliffe had originally supposed.

Face validity: Numerous postulates or procedures in certainty factor theory appear ad hoc, implausible, or inconsistent. These include its disregard for interdependencies, its disregard for prior probabilities, the arbitrary cutoff on the certainty of the antecedent required to trigger a rule, and the inconsistent simultaneous use of the MIN operator and multiple rules to capture a disjunction of evidence.

Plausibility of instances: MYCIN has had some success in empirical tests which compared its performance, in prescribing therapy, with that of experts (Lu et al., 1979). In some cases, however, MYCIN's conclusions do not match intuitions. According to Buchanan and Shortliffe, with concurring evidence, results converge too rapidly on certainty even when the evidence is very weak. In an earlier version of the calculus, a very small amount of conflicting evidence could overwhelm a large amount of concurring evidence.

What concepts of uncertainty does MYCIN address? It makes no provision for impreciseness of user inputs; for example, there is no measure of the degree to which the user's description of the data matches the antecedent of a rule. As for the chance of a hypothesis being true and the quality of evidence supporting the estimate of that chance, MYCIN is ambiguous. Certainty factors could be construed as representing either one (Buchanan and Shortliffe, 1984, Chap. 10), contributing no doubt to the semantic confusion of experts asked to provide these numbers. In light of the problems with validity indicated above, it cannot be concluded that MYCIN gives an adequate account of either of those concepts.

2.3.2 Other developments. Another well-known system, PROSPECTOR, incorporates elements of a Bayesian calculus, but deviates significantly from it in important respects, i.e., the treatment of AND and OR by MIN and MAX operators, and the concatenation of inferences across a series of rules (Duda et al., 1979). In the past two or three years, there has been a growing sense of dissatisfaction among developers of such systems with the ad hoc nature of the inference mechanisms thus far attempted, and an increasing interest in presumably more rigorous alternatives. For example, Gordon and Shortliffe (1984) have proposed that the next step for MYCIN is to replace certainty factors with Shafer's theory of belief functions. Some preliminary applications of belief functions (e.g., Lowrance and Garvey, 1983) have been proposed, and fuzzy logic now has a number of applications (cited in Zadeh, 1984a).

Unfortunately, such new departures may encounter difficulties comparable to those which faced MYCIN, unless careful consideration is given to conditions of validity involved in representing the appropriate concepts of uncertainty.

2.4 Bayesian Probabilities

2.4.1 Using probability theory for inexact reasoning. Probability theory has become central to modern scientific culture. As such, it is the obvious calculus to consider for handling inexactness in expert systems. Its supporters in this role date back to the early work on probabilistic information processing (see Edwards, 1966) and earlier; more recent contributors have been de Dombal (1973), in the field of medical decision making, and Schum (1980) in the intelligence field.

The application of probabilistic reasoning to rule-based expert systems is complex, but it can be illustrated with a simple example. Part of an expert system for image analysis could be a scene labeller, based on texture vectors. A rule in a system resembling PROSPECTOR might be:

```
IF (TEXTURE IS OF TYPE X)
     THEN (OBJECT IS A BUILDING) (LR = 2.3),
```

where LR quantifies the impact of the evidence (the texture) on the hypothesis (that the object is a building). LR is a likelihood ratio, i.e., the probability of finding a texture of type X given that the object is a building divided by the probability of that texture given that it is not a building. Satisfaction of the antecedent of this rule would lead to a process of Bayesian updating, in which the impact of the new evidence is combined with the prior odds of the hypothesis being true. Suppose H is the hypothesis that the object is a building. Then Bayes' Theorem gives, in odds-likelihood form,

$$\frac{\Pr[H|D]}{\Pr[\overline{H}|D]} = \frac{\Pr[D|H]}{\Pr[D|\overline{H}]} \cdot \frac{\Pr[H]}{\Pr[\overline{H}]}$$

where D is the data that the texture is of type X, and $\overline{H}$ is the hypothesis that some other interpretation for the object is appropriate. To carry out a simple analysis of this kind, three assessments are required, namely Pr[D|H], $\Pr[D|\overline{H}]$ and Pr[H], i.e., the likelihoods and the prior probability.

Information for understanding aerial photographs may come not only from the image itself, but also from other facts that are known about the world. So the prior belief about H might itself be derived from a probabilistic analysis. Suppose, for example, that our view of how likely an object is to be a building is affected by the existence of intelligence reports of some recent construction activity in the area. Call the existence of construction activities A, and its absence $\overline{A}$. Then we might write

$$\Pr[H] = \Pr[H|A]\Pr[A] + \Pr[H|\overline{A}]\Pr[\overline{A}].$$

Our estimation of the reliability of the reports is captured in Pr[A], and we can now think about how likely H is in the light of A or $\overline{A}$ separately.

Work on Bayesian approaches to inference has advanced from a simple one-step application of Bayes' rule to the elaboration in recent research of rather complex structures capable of capturing a wide diversity of human inference tasks and prescriptive intuitions (e.g., Schum, 1979, 1981). Bayesian techniques, for example, are able to accommodate a number of different ways that items of evidence can be related to one another with respect to a hypothesis (Schum and Martin, 1980): e.g., they may be contradictory (reporting and denying the same event), corroboratively redundant (reporting the same event), cumulatively redundant (reporting different events which reduce one another's evidential impact), or nonredundant (reporting different events which enhance or do not change one another's evidential impact). In other, more complex cases of interdependence, Bayesian techniques capture the evidential impact of biases in an information source or non-independence of information source sensitivity with respect to what is being observed.

As might be expected, evaluation of Bayesian theory leads to results that largely are the reverse of those for MYCIN; it ranks high in validity, but low in feasibility.

2.4.2 **Feasibility: Quantity of inputs.** When one attempts to use Bayesian probability theory on real inference problems, one quickly becomes aware of the complexity of the task. This complexity led Shortliffe (apparently) to construct his calculus of certainty factors as an alternative (see Shortliffe, 1976, Section 3.2). Schum (1980, p. 207) ends his advocacy of the Bayesian approach with a negative note: "...now we have other problems. I believe nobody realized how many ingredients there would be and how complex the judgments about these ingredients would be even in apparently simple cases." In all but the most trivial cases, a proper Bayesian analysis requires a great many conditional probabilities to be assessed. Schum presents the analysis of a fairly simple legal trial involving 7 pieces of evidence (Salmon's pills) and shows that at least 27 probability judgments are needed, even if all reasonable independence conditions hold. As well as requiring a very large number of probability assessments, the relations between them are difficult to organize, and the coherence of the total set of assessments is often difficult to determine.

Two important lines of defense for Bayesians are (a) that simplifying assumptions can always be made, e.g., equal prior probabilities, conditional independence of events; and (b) that variables which one does not care to deal with may be "integrated out," i.e., the resulting probabilities are regarded as marginal ("averages") with respect to possible values of the ignored variables. Thus, a Bayesian model may be created which is as simple as one likes.

Unfortunately, however, the situation is not quite as clear cut as this. "Simplifying assumptions" must in some sense be judgments (e.g., that priors are roughly equal, that events are conditionally independent). Otherwise, one sacrifices the validity of the Bayesian approach. As one Bayesian (Lindley, 1984) has put it, the Bayesian argument shows you the things you have to think about; so, think about them. From the Bayesian point of view, an argument which omits these factors is simply spurious. In the case of "integrating out" certain variables, no formal problem presents itself, since from a theoretical point of view the results with and without such variables should be the same. In actual fact, however, the difference in plausibility of the overall analysis can be very great (as we shall note below, Section 2.4.5). Thus, although the required number of assessments may in fact be reduced by either of these means, the difficulty of the judgments required to do so may be considerable. Schum speaks of them as "exquisitely subtle".

A quite different approach, which we shall explore in greater detail below, is to regard simplifying strategies as assumptions whose validity is tested implicitly through their use in reasoning. If the outcome of using such assumptions is plausible, the burden of explicitly judging their validity is avoided.

A related tactic is to accept the Bayesian framework as, in principle, the correct way to handle uncertainty, and divert our research interests to approximations that are as close as possible to the Bayesian norm. Indeed, Shortliffe (1976, p. 164) originally saw certainty factors as a device in this direction. Shortliffe, however, did not explicitly derive his theory as a special case of the more general Bayesian model. Adams (1976) showed that assumptions necessary to derive Shortliffe's postulates in some cases do not exist, and in other cases are far more restrictive and implausible than the usual assumptions of equal priors and conditional independence. We shall return to this topic in the discussion of Shafer's theory (Section 2.5).

2.4.3 **Computational tractability.** There is no known, computationally tractable method for propagating uncertainties consistently through an arbitrary Bayesian network. Restrictions of some sort on the kind of model that is utilized are necessary. The only question (as in the previous discussion of inputs) is whether the restrictions will be plausible (i.e., define a meaningful, useful special case of Bayesian modeling) or ad hoc. PROSPECTOR adopted the latter approach. More recently, Pearl (1982) and Kim (1983) have explored the former. They show that independence assumptions make sense, and probabilities can be propagated by simple local computations, if the inferential network has (a) a causal interpretation, and (b) the form of a Chow tree (i.e., it lacks undirected cycles). Unfortunately, not all real problems will fit this special structure.

If validity is not to be sacrificed, computational tractability for a Bayesian system can be purchased only in special cases; and even then, only at the cost of complex and subtle judgments regarding interdependence among items of knowledge and the overall structure of the inferential argument. As we shall see, the situation is quite similar for Shaferian belief functions. For this reason, Shafer (1984a) has recently argued, the introduction of probability into expert systems appears to be inconsistent with the modularity of knowledge representations that up to now has been the most salient characteristic of such systems.

In Section 3.0 we shall return to some of these questions. We will propose that a careful use of qualitative reasoning, superimposed upon a probabilistic system, may reduce the requirement for experts (or users) to address issues of interdependence and model structure explicitly, and make such assessments easier when they are required, without undo compromise of validity.

2.4.4 **Validity: Axiomatic derivation.** Bayesian probability theory has a preeminent, though perhaps not conclusive, claim to validity among current proposals for the handling of uncertainty. De Finetti (1937/1964) showed that unless your beliefs conform to the rules of probability, a clever opponent could make you the victim of a "Dutch book," i.e., a set of gambles you would accept, but in which you lose regardless of the outcome of an uncertain state of affairs. More recently, Lindley (1982) has given a new derivation. Suppose that people are going to measure the uncertainty of events by some method, and we wish to know how good they are at doing so. If we devise a scoring system of any sort--as along as (a) the score is a joint function of the uncertainty measure and the event's truth or falsity, and (b) scores are additive across different events--then no matter what events actually occur, the best achievable score will always go to a form of Bayesian probability. Lindley concludes that "only probability is a sensible description of uncertainty."

A common objection to this sort of demonstration is that we are not in fact always (or usually) faced with a malicious adversary or, indeed, with a scoring system. But the point is not that we are, or should somehow presume that we are, always subjected to such peculiar circumstances. Even if we never encounter these conditions, other things being equal, a system which has the property of working well in them is more desirable (in all circumstances) than one which does not. In terms of Section 3.3, it is plausible than an adequate system of uncertainty would guard against a Dutch book. It is plausible that such a system would score high if we ever chose to score it.

The more fundamental objection, in our view, is that while probability theory has been shown uniquely to possess a desirable property, but has not been shown to be uniquely justified. Other systems of uncertainty may have desirable properties that probability theory lacks. (In particular, alternative theories might deal more adequately with different kinds of uncertainty, such as incompleteness of evidence or imprecision. In this regard, note that De Finetti's and Lindley's arguments do not apply to systems which provide more than a single measure of uncertainty for each event, such as the upper and lower measures in Shafer's theory, or fuzzy probabilities in Zadeh's.)

Nonetheless, it seems incontrovertible to us that the existence of foundational arguments such as those described is a strong plus for Bayesian theory.

2.4.5 **Plausibility of instances.** As noted, the thrust of Bayesian analysis is to improve, rather than to replicate ordinary thinking. Bayesians argue that if one's ordinary intuitions are probabilistically incoherent, they ought to be changed. We might expect, nevertheless, that these revisions of belief would typically lead to judgments that are regarded as more plausible after reflection. In other words, the plausibility of the axioms should outweigh the initial plausibility of an incoherent set of judgments. In some cases, this seems true, e.g., most people who understand an explanation of the "gambler's fallacy" seem to accept that it is a fallacy; in other cases, perhaps, it is not true (e.g., Slovic and Tversky, 1974).

There is another issue here which is, we feel, more important. Even if revised (hence, coherent) beliefs are more plausible than unrevised, incoherent ones, all the credit cannot go to Bayesian theory. The reason is, that the selection of a specific revision is not uniquely determined by the requirement of coherence. Consider, again, the example above of inferring the chance of H, i.e., that a particular object is a building, based on intelligence reports of construction activity, A. Bayesian theory tells us only that our assessment of Pr[H] should be the same as $\Pr[H|A]\Pr[A] + \Pr[H|\overline{A}]\Pr[\overline{A}]$, which is based on our assessments of $\Pr[H|A]$, $\Pr[A]$, and $\Pr[H|\overline{A}]$. The theory provides no guidance in the case where the two are not equal. Coherence by itself does not dictate that the result of an analysis is to be preferred to a direct judgment. We might choose to revise one or more of the assessments in the analysis, rather than to revise Pr[H].

This problem, which we may call the <u>incompleteness</u> of Bayesian theory, is exacerbated by the fact that in any problem there is more than one possible form of analysis. Many advocates and many critics of the Bayesian approach seem to imply that there is only one way a probabilistic analysis could be carried out and only one possible conclusion. To see that this is not the case, we return to the example of inferring H. Let B be intelligence information that a strong pressure group exists within the country our photograph represents, for the erection of barracks in that general area. Instead of, or in addition to, conditioning our assessment on A, as above, we could condition on B, namely

$$\Pr[H] = \Pr[H|B]\Pr[B] + \Pr[H|\overline{B}]\Pr[\overline{B}].$$

Yet again, we could condition jointly on A and B:

$$\Pr[H] = \Pr[H|AB]\Pr[AB] + \Pr[H|A\overline{B}]\Pr[A\overline{B}] + \Pr[H|\overline{A}B]\Pr[\overline{A}B] + \Pr[H|\overline{A}\overline{B}]\Pr[\overline{A}\overline{B}].$$

Still more choices are open to us: for example, we could assess Pr[AB] directly, and/or further analyze it as Pr[A|B]Pr[B], and/or as Pr[B|A]Pr[A].

The Bayesian theoretical attitude is straightforward, namely that it does not matter which of these forms of analysis we perform or which answer we select, since coherent probability assessors should derive the same number whichever method they choose. Theory, however, is of use because we are not ordinarily coherent in our assessments. An analysis may well give us a different estimate of Pr[H] than if we directly judged it; otherwise, we wouldn't bother with the analysis. Moreover, different analyses may well give us different answers; otherwise, we would have no cause for regarding some analyses as "better" than others.

An important assumption of Bayesian theory is that all analyses (by the same person) are based on the same evidence; they do not differ in the knowledge they draw upon. We would argue that this is, psychologically, not true. Different ways of formulating the same problem may well tap different internal stores of information. What is missing from the Bayesian framework is some notion of the quality of probability inputs, i.e., the amount of knowledge or completeness of evidence that underlies them. Several points can be made:

- Revision of probability judgments should be guided by a judgment of their quality, i.e., the amount of knowledge they represent.
- More than one analysis may be of value, if they bring different knowledge to bear on a problem (cf., Brown and Lindley, 1982).
- The application of Bayesian theory to a problem is not necessarily a linear process in which inputs are provided and conclusions computed. It is (or often should be) an iterative process, in which comparison of conclusions arrived at by different methods leads to revisions of inputs and assumptions, until overall consistency is achieved.

In ordinary statistical problem solving, perhaps, judgments of quality may safely remain implicit. But a major limitation in the automation of Bayesian theory within expert systems is the lack of an explicit measure of completeness of evidence, and a mechanism for its use in the revision of probability estimates.

This will be a major focus in our discussion of Shafer, in Section 2.5, and in the new developments to be described in Section 3.0.

2.4.6 **Semantics: Behavioral specification.** Bayesian theory provides a clear behavioral interpretation of probabilities in terms of preferences among bets. We can know what someone's probabilistic beliefs are by observing their actions under specified conditions. By contrast, a common complaint by Bayesians regarding other theories is the difficulty of knowing what the basic measures mean.

There are three different, but related, misunderstandings of this "operational definition." First, critics point out that betting may be an awkward and in some cases an impossible method for eliciting probabilities. It is often easier to ask for direct verbal judgments. There is a standard answer to this point by sophisticated Bayesians: Meaning need not be equated with evidence. Bayesians can use any method they like for estimating your probabilities, if there is a reasonable expectation that the result will match, or at least approximate, what they would have gotten had they used the betting paradigm.

This response hides a more subtle misunderstanding. It is still assumed that we can, at least in principle, always know what a person's probabilities are, simply by testing his preferences among bets. Since the operational definition specifies a situation where he must make a choice, it is implied that any person "has" probabilities waiting to be uncovered or "elicited". Is Bayesianism thus inevitable? This conception seems to be contradicted by the incoherence we typically find in people's unaided judgments, and which is amply documented in the experimental psychology literature (e.g., Kahneman, Slovic, and Tversky, 1982).

The sophisticated Bayesian was right, we suggest, in distinguishing meaning and evidence. But--sophisticated as he is--he has not absorbed the full implications of that distinction. Although he permits other kinds of evidence, he is still equating meaning with a particular observable operation. The problem, as pointed out by Quine (1953) and others in a more general critique of positivism, is that the selection of this rather than some other component of the theory as a "definition" is arbitrary. To return to our earlier example, suppose we equate

Pr[H] for a person X with X's betting behavior in regard to H. Then we determine in the same way his value for Pr[H|A], Pr[H|$\overline{A}$], and Pr[A]. Finally, we compute a new probability of H, Pr'[H], from the latter three values. Why shouldn't we define X's probability for H in terms of <u>this</u> operation, i.e., as Pr'[H]? One reply is that this operation requires a theoretical assumption viz., that X is coherent, to justify the computation of Pr'[H] from Pr[H|A], Pr[H|$\overline{A}$], and Pr[A]. But the earlier "operational definition" <u>could</u> be regarded as theoretical, too, since it is a theoretical hypothesis (i.e., that X acts so as to maximize subjectively expected utility) that enables us to derive X's probability for H from his preferences among gambles involving H. Conversely, we <u>could</u> regard the definition in terms of Pr'[H] as purely "behavioral", by ignoring the theoretical hypotheses implicit in our calculations.

It is far more natural to regard <u>all</u> these potential "definitions" simply as theoretical predictions. How then, without definitions, do we assess the probabilities and utilities required to derive the predictions? The answer is that testing a theory is, inevitably, a bootstapping operation, in which we use the theory, as if it were true, to estimate values for an interrelated set of parameters, then test for consistency of the results. If the results are consistent, the theory is confirmed; if not, it is disconfirmed. (For a general discussion see Glymore, 1980.) To the extent that people are probabilistically incoherent, therefore, probability theory is disconfirmed, and they cannot be regarded as "having" probabilities at all.

Have we overlooked the difference between descriptive and prescriptive theories? Perhaps "operational definitions" make sense for probabilities because they form part of a prescriptive theory. On the contrary, we suggest that there is a strong and important parallel between theory testing, as we just described it, and prescriptive analysis (as we saw it in Section 2.4.5). Just as in descriptive science, we assume the prescriptive theory to be true, use it to perform a set of interrelated analyses, and then test them for consistency. However, if we find inconsistency among alternative prescriptive analyses, or between an analysis and direct judgment, we do not (necessarily) drop the prescriptive theory; we may choose to revise the values in one or more analyses so as to <u>make</u> them consistent.

In so doing, we <u>construct</u> rather than discover or confirm a probability model for our beliefs.

The analogy between descriptive and prescriptive processes may be carried a step further by recalling our observations in Section 2.2.3. If the inconsistency of our judgments with respect to probability theory is great enough, and if coherence-producing revisions seem implausible, we may indeed decide to reject probability theory as a proper prescriptive guide.

What then is left of the Bayesian claim that operational definitions are required for clarity of concepts? The third and final misunderstanding we wish to address is the notion that because "operational definitions" are arbitrary, and do not guarantee the applicability or even the relevance of a prescriptive theory, that <u>behavioral specification</u> is of no use. In fact, it is quite critical: without it, there is no link, or else no clear link, between the prescriptive theory and action. With it, the prescriptive process described above, in which a coherent set of judgments is arrived at through successive iterations, also produces a clear set of implications for action. In expert system applications, such implications are typically the reason for developing the system. Moreover, such specifications may play a clarifying role for the decision maker in the process of iteratively arriving at an appropriate set of judgments. (We return to this point in Section 2.5.11 below.) The existence of such specifications must, therefore, be counted as a plus for the Bayesian theory.

2.4.7 **Naturalness of inputs.** Behavioral specification is not sufficient to guarantee the usefulness of an inference framework. A common objection to Bayesian theory urged by proponents of alternative views, is that the inputs it requires exceed, in various ways, the capabilities of the decision makers it is designed to aid. Two complaints of this type must, however, be carefully distinguished:

<u>Imprecision</u>: Bayesians assume that experts are capable of quantifying their uncertainties and values to an arbitrary degree of precision. But this is true of

no other known process of measurement. Experts may simply not know, to the required exactitude, what their beliefs or preferences are.

Incompleteness of evidence: The evidence may not justify the degree of confidence suggested by use of a single number to assess an uncertainty. Some assessments (e.g., the probability that the Soviets will invade Western Europe within the next year) are less well supported than others (e.g., the probability that a coin in my pocket will land heads if tossed). In the former cases, the available evidence may justify no more than a range of probabilities rather than a single number.

There is an important distinction between these two complaints: the first is consistent with the basic prescriptive adequacy of probability theory, but seeks to accommodate human shortcomings in the assessment task. In contrast, the second objection has a normative basis: probabilities themselves are inappropriate where evidence is incomplete. We shall explore these positions in more detail in our discussions of Zadeh and Shafer, respectively.

2.4.8 Concepts of uncertainty. Bayesian theory is clearly designed to capture the concept of chance, or uncertainty about facts. We argued in Section 2.4.5 that an important gap in Bayesian theory is the lack of a measure of completeness or quality of evidence, i.e., the lack of a distinction between firm probabilities (.5 as the probability of heads on a coin toss) and those based on guesswork (.5 as the probability of a Soviet invasion). Intuitively, the weight of evidence supporting some probability judgments is stronger that that supporting others. We argued that this concept in fact plays an important role in ordinary applications of probability theory, by guiding the choice among potential revisions of belief in the light of an analysis or set of analyses. We hope to demonstrate below (Section 3.0) that an explicit measure of this sort is critical for the control of reasoning in an expert system that intelligently handles uncertainty about facts.

To what extent could Bayesian theory itself be extended to cover the concept of completeness of evidence? Lindley et al. (1979) have recently attempted to formalize the intuitive notion that we are firmer about some probability assessments than others. The tool they introduce is a second-order probability distribution over possible values of the true first-order probability. The spread of the second-order distribution is a measure of the firmness of the original probabilities. Lindley et al. have described procedures for statistically aggregating inconsistent probabilistic analyses by means of such second-order judgments.

These efforts have failed, in our opinion, for a variety of reasons. **Feasibility:** The quantity and difficulty of required inputs is increased, rather than decreased, to the degree that one's evidence is incomplete. Computational intractability will certainly be increased as well. **Validity:** Axiomatic justifications and behavioral specifications which apply to first-order probabilities become much less convincing at higher levels, where, for example, gambles or scores which depend on one's own "true" probabilities, rather than actual events, lack plausibility. Face validity is dubious as well: e.g., if we attempt to measure the quality of our second-order probabilities in the same way, we are threatened with an infinite regress. Perhaps the most serious difficulty, however, is the implausibility of the inferences to which this model gives rise. In brief, the procedure for aggregating probabilistic analyses assumes that they disagree only because of "noise," or random error, in the assessment process; hence, it yields results which do not reflect the possibility that different analyses have drawn on different evidence. We suggest that from a psychological point of view, different analyses may tap different portions of our store of knowledge, even when performed by the same individual. These points are amplified in Cohen et al., 1984, and in a planned paper by Cohen and Lindley.

2.4.9 Summary. Bayesian probability theory is strong in the formal aspects of validity. Its logical foundations are perhaps uniquely compelling in application to the concept of chance. However, the input and computational burdens which it imposes, except when specialized models are adopted, are considerable. It has no adequate resources for representing the quality of an inferential argument, and requires an arbitrary degree of precision in numerical judgments. Even its validity, in a more informal sense, can be questioned. Bayesian theory, as it stands, implies that one's beliefs should be coherent but provides no guidance for choosing among alternative equally coherent analyses. Moreover, by assuming that all assessments are based on the same evidence, it closes off the most promising source of such guidance. We have argued that the application of Bayesian theory to a problem is not linear process in which conclusions are computed from inputs. It is (or often should be) an iterative bootstrapping process in which comparison of conclusions arrived at by different methods leads to revision of inputs and assumptions, until overall plausibility is maximized. This process of revising probability assessments should be guided by a judgment of their quality. A more satisfactory account of completeness of evidence is, therefore, essential.

2.5 Belief Functions

2.5.1 Nature of the theory. In the theory of belief functions introduced by Shafer (1976), Bayesian probabilities are replaced by a concept of evidential support. The contrast, according to Shafer (1981; Shafer and Tversky, 1983) is between the chance that a hypothesis is true, on the one hand, and the chance that the evidence means (or proves) that the hypothesis is true, on the other. Thus, we shift focus from truth of a hypothesis to the interpretation of the evidence. As a result, the system (a) is able to provide an explicit measure of quality of evidence, (b) is less prone to require a degree of definiteness in inputs that exceeds the knowledge of the expert, and (c) permits segmentation of reasoning into analyses that depend on independent bodies of evidence.

In Shafer's system, the support for a hypothesis and for its complement need not add to unity. For example, if a witness with poor eyesight reports the presence of enemy artillery at a specific location, there is a certain probability that his eyesight was adequate on the relevant occasion and a certain probability that it was not, hence, that the evidence is irrelevant. In the first case, the evidence proves the artillery is there. In ~~no~~ neither case could the evidence prove the artillery is not there.

To the extent that the sum of support for a hypothesis and its complement falls short of unity, there is "uncommitted" support, i.e., the evidence is incomplete. Evidential support for a hypothesis is a lower bound on the probability of its being true, since the hypothesis could be true even though our evidence fails to demonstrate it. The upper bound is given by supposing that all present evidence that is consistent with the truth of the hypothesis were in fact to prove it. The interval between lower and upper bounds, i.e., the range of permissable belief, thus reflects the incompleteness of evidence for that hypothesis. This concept is not captured by Bayesian probabilities.

In Shafer's calculus, support m($\cdot$) is allocated not to hypotheses, but to sets of hypotheses. Shafer allows us, therefore, to talk of the support we can place in any subset of the set of all hypotheses. In the case of three hypotheses, H_1, H_2 and H_3, for example, we could allocate support to H_1, H_2, H_3, {H_1 or H_2}, {H_1 or H_3}, {H_2 or H_3}, and {H_1 or H_2 or H_3}. As with probability, the total support across these subsets will sum to 1, and each support m($\cdot$) will be between 0 and 1. It is natural, then, to say that m($\cdot$) gives the probability that what the evidence means is that the truth lies somewhere in the indicated subset.

Suppose, for example, that we know in the case of three hypotheses that H_3 is false, but have no evidence to distinguish between H_1 and H_2. In that case, we would put m({H_1 or H_2}) = 1, and give zero support to all the other possible subsets. Alternatively, we may feel that the evidence either means that H_3 is true, or that {H_1 or H_3} is true, or that it is not telling us anything (i.e., {H_1 or H_2 or H_3} is true), and that the weight of evidence is just as strong with each possibility. In that case m(H_3) = m({H_1 or H_3}) = m({H_1 or H_2 or H_3}) = 1/3. In a Bayesian analysis, arbitrary decisions would have to be made about allocating probability within these subsets, requiring judgments that are unsupported by the evidence. *

This same device, of allocating support to subsets of hypotheses, enables us to represent the reliability of probability assessments. Suppose, for example, that the presence of texture X in an image region is associated with a building 70% of the time and with other labels 30% of the time, based on frequency data from a set of training photographs. If we are confident that an image now being analyzed is representative of the training set, we may have m(building) = .7 and m(other) = .3. But if there is reason to doubt the relevance of the frequency data to the present problem (e.g., due to geological or cultural differences between the two geographical areas), we may discount this support function by allocating s~~a~~ome per- *

centage of support to the universal set. For example, with a discount rate of 30%, we get m(building) = .49, m(other) = .21, and m ({building, other}) = .30. The latter reflects the chance that the frequency data is irrelevant.

Shafer's belief function Bel(·) summarizes the implications of the m(·) for a given subset of hypotheses. Bel(A) is defined as the total support for all subsets of hypotheses contained within A; in other words, Bel(A) is the probability that the evidence implies that the truth is in A. The plausibility function Pl(·) is the total support for all subsets which overlap with a given subset. Thus, Pl(A) equals $1\text{-Bel}(\overline{A})$; i.e., the probability that the evidence does not imply the truth to be in not-A. In one of the examples above, with

$$m(H_3) = m(\{H_1 \text{ or } H_3\}) = m(\{H_1 \text{ or } H_2 \text{ or } H_3\}) = 1/3,$$

we get:

$$\text{Bel}(H_3) = m(H_3) = 1/3;\ \text{Pl}(H_3) = 1\text{-Bel}(\{H_1 \text{ or } H_2\}) = 1$$

$$\text{Bel}(\{H_1 \text{ or } H_3\}) = m(H_3) + m(\{H_1 \text{ or } H_3\}) = 2/3;\ \text{Pl}(\{H_1 \text{ or } H_3\}) = 1\text{-Bel}(\{H_2\}) = 1.$$

2.5.2 Dempster's rule. Thus far, we have focused on the representation of uncertainty in Shafer's system. For it to be a useful calculus, we need a procedure for inferring degrees of belief in hypotheses in the light of more than one piece of evidence. This is accomplished in Shafer's theory by Dempster's rule. The essential intuition is simply that the "meaning" of the combination of two pieces of evidence is the intersection, or common element, of the two subsets constituting their separate meanings. For example, if evidence E_1 proves $\{H_1$ or $H_2\}$, and evidence E_2 proves $\{H_2$ or $H_3\}$, then the combination $E_1 + E_2$ proves H_2. Since the two pieces of evidence are assumed to be independent, the probability of any given combination of meanings is the product of their separate probabilities.

Let X be a set of hypotheses $H_1, H_2, \ldots, H_n$, and write 2^X for the power set of X, that is, the set of all subsets of X. Thus, a member of 2^X will be a subset of hypotheses, such as $\{H_2, H_5, H_7\}$, H_3, or $\{H_1, H_2, H_3, H_4\}$, etc. Then if $m_1(A)$ is the support given to A by one piece of evidence, and $m_2(A)$ is the support given by a second piece of evidence, Dempster's rule is that the support that should be given to A by the two pieces of evidence is:

$$m_{12}(A) = \frac{\displaystyle\sum_{A_1 \cap A_2 = A} m_1(A_1)m_2(A_2)}{1 - \displaystyle\sum_{B_1 \cap B_2 = \emptyset} m_1(B_1)m_2(B_2)}.$$

*

The numerator here is the sum of the products of support for all pairs of subsets A_1, A_2 whose intersection is precisely A. The denominator is a normalizing factor which ensures that $m_{12}(\cdot)$ sums to 1, by eliminating support for impossible combinations.

Consider, for example, the following two support functions:

Table 2-1

	$m_1(\cdot)$	$m_2(\cdot)$	$m_{12}(\cdot)$
H_1	0.2	0.1	0.344
H_2	0.1	0.3	0.250
H_3	0.3	0	0.172
H_1H_2	0.1	0.3	0.125
H_1H_3	0.2	0	0.063
H_2H_3	0	0.1	0.016
$H_1H_2H_3$	0.1	0.2	0.031

In the third column, we have used Dempster's rule to compute $m_{12}(\cdot)$. For example

$$m_{12}(H_1H_2) = \frac{m_1(H_1H_2)m_2(H_1H_2)+m_1(H_1H_2)m_2(H_1H_2H_3)+m_1(H_1H_2H_3)m_2(H_1H_2)}{1-C}$$

where $C = m_1(H_1)[m_2(H_2) + m_2(H_3) + m_2(H_2H_3)] + m_1(H_2)[m_2(H_1) + m_2(H_3) + m_2(H_1H_3)]$

$+ m_1(H_3)[m_2(H_1) + m_2(H_2) + m_2(H_1H_2)] + m_1(H_1H_2)m_2(H_3) + m_1(H_1H_3)m_2(H_2)$

$+ m_1(H_2H_3)m_2(H_1)$

and so $$m_{12}(H_1H_2) = \frac{0.1\text{x}0.3+0.1\text{x}0.2+0.1\text{x}0.3}{1 - 0.36} = 0.125.$$

Let us now examine the performance, or at least the potential, of Shafer's theory within our evaluation framework.

2.5.3 **Feasibility: Quantity of inputs.** One of the main difficulties standing in the way of a Bayesian analysis is its complexity. At first sight the Shaferian approach seems simpler, since complicated independence judgments and conditional probability assessments appear not to be required. This appearance is illusory. Support functions must be assessed over not just the hypothesis set, but over the power set of the hypothesis set. With 10 hypotheses, for example, the support distribution has 1,023 elements. For both Bayesian and Shaferian models, the required number of assessments or judgments increases exponentially with the number of events or hypotheses. To see the parallel, compare the Bayesian rule:

$$\Pr[A \text{ or } B] = \Pr[A] + \Pr[B] - \Pr[A]\Pr[B|A]$$

with Shafer's rule:

$$\mathrm{Bel}(\{A \text{ or } B\}) = m(A) + m(B) + m(\{A \text{ or } B\}).$$

In each case, to get an uncertainty measure for a disjunction (i.e., a member of 2^X), we must make one assessment in addition to the measures already assessed for the elements. For Bayesians, the extra assessment is a conditional probability $\Pr[B|A]$; for Shaferians it is the direct evidential support $m(\{A \text{ or } B\})$.

A Shaferian response to this, in parallel with the Bayesian response (Section 2.4.2), is that specialized models may be developed that require far fewer assessments. In fact, the belief function framework admits a variety of interesting special cases: e.g.,

- simple support functions: all support goes either to some one individual hypothesis or to the universal set X, i.e., either the evidence is reliable and pinpoints the answer or it is totally untrustworthy;

- discounted probabilistic support functions: all support goes to individual hypotheses (as in a standard probability distribution), with some additional support possibly going to the universal set X (reflecting a judgment of the quality of the evidence for the probability distribution);

- consonant support functions: all support goes to a nested series of subsets of hypotheses; i.e., the evidence points in a certain direction but is unclear how far we should go;

- hierarchical support functions: the evidence supports subsets of hypotheses that can be arranged in a tree.

Here again, however, (as in the Bayesian case) complex and difficult judgments must be made to determine that a particular specialized model is applicable, before savings in quantity of assessments can be realized.

The problem for Shaferians may even be deeper. The applicability of Dempster's rule to two bits of evidence E_1 and E_2 is not automatic. It requires rather careful and difficult consideration of a whole set of independence assumptions. We shall return to this point in our discussion of the validity of Shafer's theory (Section 2.5.5).

2.5.4 **Computational tractability.** Here again the story is parallel to the Bayesian case. The employment of unrestricted belief function models would involve prohibitive computation. As a result, Gordon and Shortliffe (1984) propose to modify Dempster's rule to simplify computation in MYCIN. Shafer (1984a) has argued in response that ad hoc modifications of this sort might be avoided by a control strategy that intelligently exploits the structure of restricted belief function models, such as the hierarchical structure proposed for MYCIN. Here as in the Bayesian case, feasibility is purchased only in special cases, and, evidently, at the cost of complex and subtle judgments regarding the structure of the overall argument.

2.5.5 **Validity: Semantics.** Shafer argues that the requirement for a behavioral specification of probabilities is irrelevant. People bet in a certain way because of their beliefs and preferences; observing their own betting behavior will not help them to assess those beliefs. Shafer thus urges a shift from the positivist to a more cognitive orientation. He argues that uncertainty is quantified on the basis of an analogy between one's problem and a "canonical example". In Bayesian modeling, we assess the probability of an event by comparing its likelihood with the likelihood of a frequency-based event, such as a random drawing from an urn. Thus, for Shafer, to say that the Bayesian probability of an event is x is to say that it is "like" the chance of drawing a white ball from an urn with a proportion of white balls equal to x. Similarly, to say that your Shaferian belief in a proposition is y, is to compare it to canonical examples of the type we shall explore in Section 2.5.6, where the reliability of an evidential source is determined by chance.

Unfortunately, Shafer's position is weakened by two considerations: First, his canonical examples, as we shall see below, are far more complex and less obviously useable, even from a cognitive point of view, than the Bayesian examples. Second, behavioral specification probably plays a cognitive role in clarifying the sense of a canonical example. For example, what does it mean to say that my uncertainty about whether an object is a building is "like" my uncertainty about drawing from an urn? In what respects must they be similar? Many people will find it illuminating when told it means that I would bet at equal stakes on either event.

A major strength of Shafer's theory, nevertheless, is the naturalness of the input format it imposes:

- Assessments need go no further than the evidence justifies. As we have seen, "ignorance" is naturally represented by assigning support to a subset of hypotheses, with no further commitment to an allocation within the subset. A Bayesian must decide among quite definite and distinct, but equally arbitrary, allocations of probability.

- Weight or completeness of evidence is quite intuitively represented as the degree to which the sum of belief for a hypothesis and its complement falls short of unity.

- Assessments may be based on distinct, separable bodies of evidence, rather than requiring--as in Bayesian theory--that all assessments be based on all the evidence.

2.5.6 **Face validity.** Belief function theory possesses no deep axiomatic justification comparable to the de Finetti and Lindley arguments for Bayesian theory. Not coincidentally, however, Shafer has offered a view of model "validation" which contrasts sharply with the axiomatic approach. On Shafer's view (1981; Shafer and Tversky, 1983), theories of inference are tools which can be used to help us construct (rather than elicit or discover) a set of probabilities. The justification for applying a particular tool to a particular problem is that we see an analogy between that problem and the canonical example underlying the theory. For example, to the extent that the Bayesian theory has anything to contribute, it is by establishing a persuasive analogy between your problem and a situation, like drawing balls from an urn, where the truth is generated by known chances.

Bayesian analogies of this sort, according to Shafer, will usually be imperfect, because in the canonical example we know the rules of the game that determine how the truth is generated (e.g., the composition of the urn and the procedure for drawing a ball). In real problems, there are nearly always many aspects of the situation where comparable rules cannot be given without making numerous assumptions. When these assumptions become very extensive, it may be better to switch to a simpler kind of model, which is more plausible despite not giving a complete picture of how the truth is generated. Such simpler models can be based on canonical examples in which the meaning of the evidence rather than the truth is generated by known chances.

We comment on Shafer's position at two levels: First, how convincing is his concept of validity? Second, how plausible or useful are the canonical examples underlying belief functions?

2.5.7 Concept of validity. For Shafer, validity reduces to face validity and plausibility of instances. His argument for this position, however, contains some confusion. Shafer mistakenly assumes that the adoption of an axiomatic framework implies a belief in pre-existing rather than constructed probabilities. Thus, Shafer (1984a) speaks derisively of assessment in the Bayesian context as "pretending" that one already has probabilistically coherent beliefs and preferences, and then, somehow, "trying to figure out what they are."

Our own view is that Shafer is correct to regard probability frameworks as tools for the construction, rather than discovery, of probabilities. But he is wrong in supposing that the axiomatic derivation of a framework detracts from this role--as long as we understand, as argued in Section 2.2.3, that axiomatic derivation is only *one* argument in favor of a given framework. If taken seriously, Shafer's argument would declare as "non-constructive" any set of prior constraints on the way uncertainty is represented or manipulated; thus, it applies as strongly against belief functions and Dempster's rule as to Bayesian probabilities. The solution in our view is not to drop constraints, but to drop the view that any particular set of constraints is inevitable. Thus, probability assessment as we understand it (Section 2.4.5) is an iterative and constructive process, in which a tentative framework (e.g., Bayesian or Shaferian) is adopted, assessments are made within the framework, checked for consistency, and revised; if the overall result is unnatural or implausible, the framework itself may be rejected or revised. In other words, "pretending" that a framework is correct is a legitimate strategy in uncertainty assessment; indeed, it is the only *possible* strategy. A framework is of use as a tool precisely because it *does* impose (tentative) constraints on the assessments that are produced. It challenges the expert to actively shape a previously disorganized and perhaps even unverbalized set of beliefs. It serves as a medium or language in which the expert "thinks" about uncertainty and in which he expresses those thoughts. A supposedly "neutral" framework, that imposed no format or structure, beyond that already present, would not help the expert in the process of construction and could not advance his or our understanding of his beliefs. (See Cohen, Mavor, and Kidd, 1984, for a more general argument in the context of knowledge engineering.)

In sum, Shafer's argument for a constructive process of probability assessment is correct. But he appears to have drawn two unnecessary conclusions: (1) It in no way contradicts the added plausibility that may be lent to a framework by the existence of an axiomatic derivation; and (2) it should not blind us to the importance of the iterative strategy of tentatively adopting a framework and testing its implications.

2.5.8 Shafer's canonical example. As noted above, when we apply a belief function analysis, we "pretend" that the meaning of the evidence is generated by known chances. In order to evaluate Shafer's theory in terms of face validity, we must examine this analogy more closely. In particular, we must focus on the independence assumptions embodied in the canonical example which are required to license an application of Dempster's rule. It turns out that these assumptions are the primary constraints imposed by Shafer's theory on the process of evaluating evidence; hence, they are its main contribution to the "construction" of probability judgments. They have also been the major source of controvery between Shafer and Bayesians. Early critics of Shafer's work (e.g., Williams, 1978) complained about the obscurity of Shafer's notion of "independent evidence." In a recent paper, however, Shafer (in press) has clarified this concept considerably.

Shafer's interpretation of belief functions involves two sets of hypotheses (or "frames") as shown in Figure 2-3. One frame, S, is a set of background hypotheses which concern the state of the process that produced the evidence at hand. For example, if the evidence E_1 is a witness's testimony that he saw artillery in a certain location, the frame S may simply be the two possibilities {the witness is reliable, the witness is not reliable}. The other frame, T, contains the hypotheses of primary interest, e.g., {the artillery is present, the artillery is not present}. To get a belief function, we only need (i) a probability distribution over S, i.e., standard probabilities P_1 and P_2, for the reliability and unreliability of the witness; and (ii) a mapping from S to T based on the content of the evidence. Since the evidence is the witness's report of artillery, reliability in S maps onto {the artillery is present} in T; unreliability in S maps onto the set {the artillery is present, the artillery is not present} in T. Support m(A) for a subset A in T is just the probability for hypotheses in S that map only onto A. (We have referred to this, somewhat loosely, as the probability that the evidence "means" A). Bel(A) for a subset A in T is the sum of the probabilities for hypotheses in S that map onto subsets of T that are contained in A. Thus, in our example, Bel(artillery is present) = P_1; Bel({present, not present}) = $P_1 + P_2$.

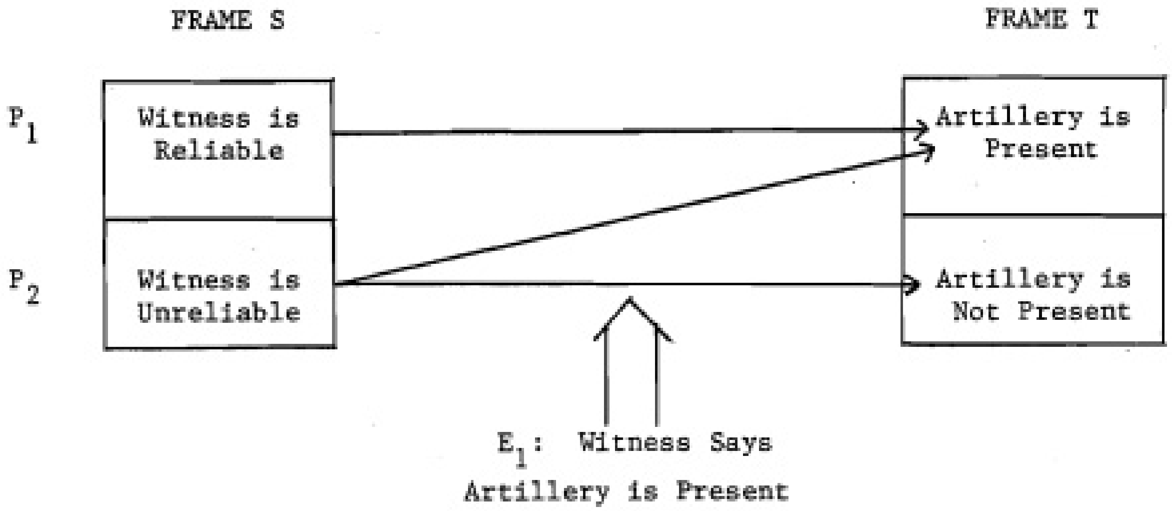


Illustration of Canonical Example
For Belief Functions

Figure 2-3

Suppose we now receive a second piece of evidence, E_2, which is the testimony of a second witness that he saw artillery in the same vicinity. We define a new belief function for this witness by specifying a frame S_2 with the elements {the second witness is reliable, the second witness is unreliable}, and by assessing probabilities P_1' and P_2' over S_2. What is our new overall belief in the elements of T? Renaming ~~Naming~~ S as S_1, Figure 2-4 shows a new frame, S_1xS_2, which results from combining elements of S_1 and S_2. Each cell has a probability which is the product of the probabilities of the elements from S_1 and S_2; and each cell is mapped onto a subset of hypotheses in T, based on knowledge of E_1 and E_2. According to this mapping (as shown by the labels in the cells), support for the artillery being present equals the chance that either witness 1 or witness 2 is reliable, i.e., $P_1P_1' + P_1P_2' + P_2P_1'$. This is the result given by Dempster's rule.

What if the report of the second witness contradicts, rather than confirms, the first? That is, E_2 is a report that artillery is not present in the specified location. In that case, the new frame, S_1xS_2, appears as in Figure 2-5. The only change is in the mapping of the cells to subsets in T--a change required by the change in E_2. It turns out, however, that the cell corresponding to both witnesses being reliable does not map to any subset in T. Since E_1 and E_2 are contradictory, both cannot be true. Thus, we use our knowledge of E_1 and E_2 to prune out impossible cells in S_1xS_2. According to the mapping, support for artillery being present equals the chance that witness 1 is reliable and witness 2 is unreliable, i.e., $P_1P_2'/(1-P_1P_1')$, normalizing to remove the impossible case. Once again, this is the result of applying Dempster's rule.

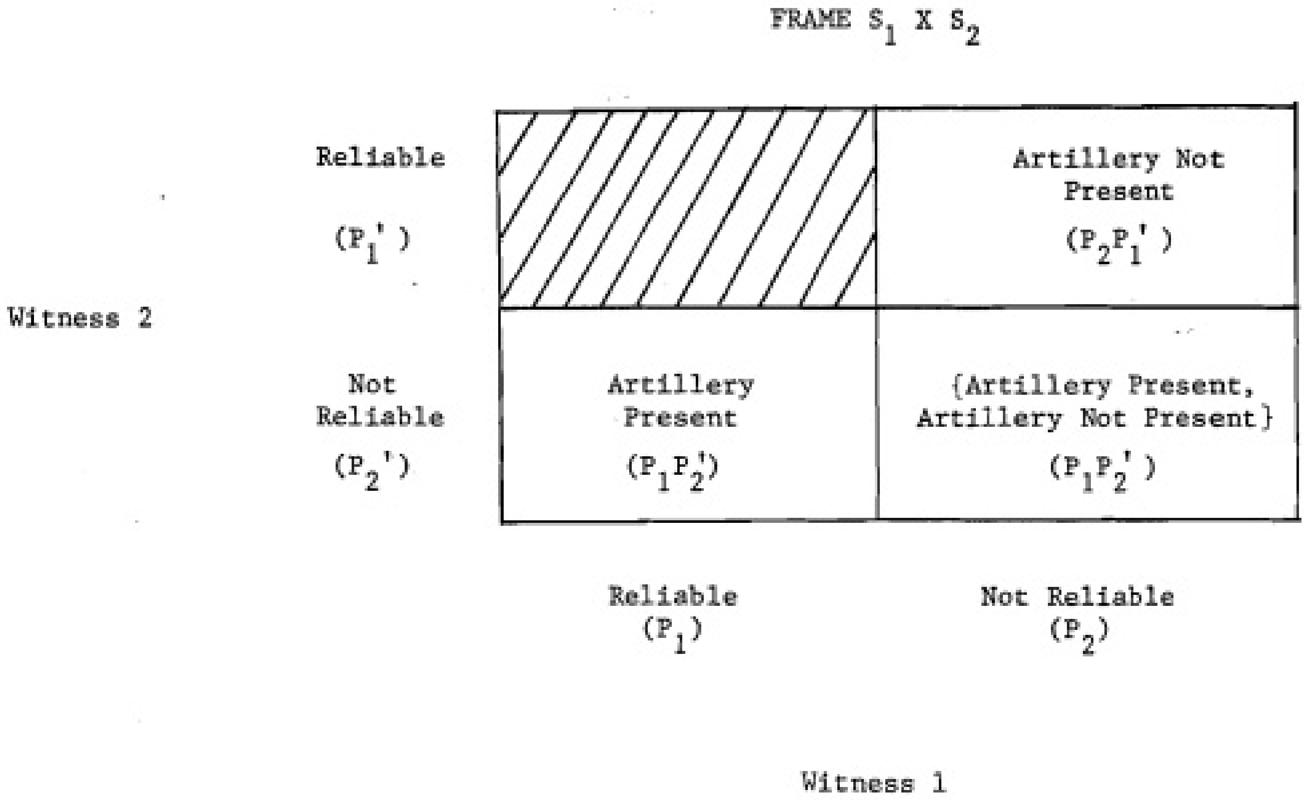
In many of Shafer's discussions, he appears to argue that Dempster's rule is justified in situations which "resemble" this canonical example, because it is the correct rule for the example (just as Bayesian rules are correct for the case of drawing balls from an urn). But what makes it correct? Even these simple examples may seem too complex for such a direct appeal to intuition. A recent paper by Shafer (in press) contains a more extensive discussion of the preconditions of Dempster's rule. We can use Dempster's rule, he says, only if the following judgments are made:

FRAME S_1 X S_2

Witness 2	Reliable (P_1')	Artillery Present (P_1P_1')	Artillery Present (P_2P_1')
	Not Reliable (P_2')	Artillery Present (P_1P_2')	{Artillery Present, Artillery Not Present} (P_2P_2')
		Reliable (P_1)	Not Reliable (P_2)
		Witness 1	

Canonical Example For Combination of Concurring Evidence

Figure 2-4



FRAME S_1 X S_2

		Reliable (P_1)	Not Reliable (P_2)
Witness 2	Reliable (P_1')	▨	Artillery Not Present (P_2P_1')
	Not Reliable (P_2')	Artillery Present (P_1P_2')	{Artillery Present, Artillery Not Present} (P_1P_2')

Witness 1

Canonical Example For Combination
of Conflicting Evidence

Figure 2-5

(a) Before consideration of the mapping to T, any hypothesis in S_1 is compatible with any hypothesis in S_2 (so S_1xS_2 can be defined as a new frame).

(b) Probabilities for elements of S_1 are independent of elements in S_2 (e.g., we do not alter our estimate of the reliability of one witness based on the reliability or unreliability of the other witness).

(c) If we could draw a conclusion about the truth of a subset in T by knowing that a certain combination of hypotheses from S_1 and S_2 was the case, then we could have drawn the same conclusion by knowing that either one or the other of the hypotheses (from S_1 or S_2) was the case. (In the example of concurring witnesses, we can conclude that artillery is present if both witnesses are reliable; but all we needed was one or the other to be reliable).

(d) The evidence we use for assessing S_1 and S_2 tells us nothing more directly about T. (All the work of reasoning about T is transferred to reasoning about S.)

Having enumerated these assumptions, we must remark that our original question about the rationale for Dempster's rule remains unanswered. It has not been demonstrated in any way that Dempster's rule "follows from" these preconditions. Perhaps Shafer means simply that when these particular conditions are met, Dempster's rule will appear more plausible or natural.

Note, however, that the canonical situation described by these conditions includes a chance model: Because of assumptions (a) and (b), the probability for a component of S_1xS_2 is simply the product of the probabilities assigned to the components of S_1 and S_2. It is tempting, therefore, to view the belief function model as a special case of a Bayesian analysis, defined by the restrictions outlined in (a) - (d). In that case, Dempster's rule should be justifiable from (a) - (d) by the rules of probability theory. Moreover, Shafer's model would then inherit the axiomatic justification of the Bayesian model in the special circumstances where it applied.

2.5.9 A Bayesian foundation for belief functions? To see how this might work, consider the simple case of Figure 2-3, with H = the artillery is present, $\overline{H}$ = the artillery is not present, R = the first witness is reliable, and $\overline{R}$ = the first witness is not reliable. It follows from probability theory that:

$$\mathrm{Pr}(H) = \mathrm{Pr}(H|R)\mathrm{Pr}(R_{\cancel{1}}) + \mathrm{Pr}(H|\overline{R})\mathrm{Pr}(\overline{R}).$$

Following Shafer's definitions, we interpret m(H) as Pr(R) and m(H or $\overline{H}$) as Pr($\overline{R}$). In addition, from our knowledge of E_1 (i.e., the mapping from S_1 to T which it establishes), and using (d), we know that Pr(H|R) = 1; if the witness is reliable, then the artillery is present. Hence, we may write

$$\mathrm{Pr}(H) = m(H) + \mathrm{Pr}(H|\overline{R})\ m(H \text{ or } \overline{H})$$

and this gives

$$\mathrm{Bel}(H) = m(H) \leq \mathrm{Pr}(H) \leq m(H){+}m(H \text{ or } \overline{H}) = \mathrm{Pl}(H),$$

where Bel(H) and Pl(H) are Shafer's belief and plausibility functions. It appears, then, that the belief function analysis is simply an incomplete Bayesian analysis. Our uncertainty about Pr(H) is due to our failure, in the belief function approach, to specify Pr(H|$\overline{R}$), i.e., the chance of the hypothesis being true despite the fact that the present evidence is unreliable. This is just another way of saying that Shafer is interested in the proof of the hypothesis, not its truth. If Pr(H|$\overline{R}$) = 0, Pr(H) = Bel(H); and if Pr(H|$\overline{R}$) = 1, Pr(H) = Pl(H). Thus, Bel(H) and Pl(H) give lower and upper bounds for the Bayesian probability.

Let us now see how Dempster's rule works within this Bayesian interpretation. Let R_1 and R_2 refer to the reliability of the first and second witness, respectively, and take the case where E_1 and E_2 agree. A Bayesian probability Pr(·|·), is a function of two arguments, the event and the evidence. Presumably, therefore, in using Dempster's rule, the probability to be bounded is Pr(H|E_1,E_2). Let us for the moment, however, ignore this consideration and use Pr(H). (Note that in the case of one piece of evidence, we likewise used Pr(H) instead of Pr(H|E_1).) By probability theory, we have

$$\mathrm{Pr}(H) = \mathrm{Pr}(H|R_1 \text{ or } R_2)\mathrm{Pr}(R_1 \text{ or } R_2) + \mathrm{Pr}(H|\overline{R_1 \text{ or } R_2})\mathrm{Pr}(\overline{R_1 \text{ or } R_2}).$$

Substituting based on conditions (a) and (b), we have

$$\Pr(H) = \Pr(H|R_1 \text{ or } R_2)[\Pr(R_1)+\Pr(R_2)-\Pr(R_1)\Pr(R_2)] + \Pr(H|\overline{R}_1\overline{R}_2)\Pr(\overline{R}_1)\Pr(\overline{R}_2).$$

By Dempster's rule,

$$m_{12}(H) = \Pr(R_1) + \Pr(R_2) - \Pr(R_1)\Pr(R_2)$$

* and by Shafer's definitions,

$$m_{12}(H \text{ or } \overline{H}) = \Pr(\overline{R}_1)\Pr(\overline{R}_2).$$

Using (c) and (d) and the mapping from $S_1 x S_2$ to T, $\Pr(H|R_1 \text{ or } R_2) = 1$. Therefore,

*

$$\Pr(H) = m_{12}(H) + \Pr(H|\overline{R}_1\overline{R}_2)m_{12}(H \text{ or } \overline{H}).$$

It follows that

$$\text{Bel}_{12}(H) = m_{12}(H) \leq \Pr(H) \leq m_{12}(H) + m_{12}(H \text{ or } H) = \text{Pl}_{12}(H).$$

Thus, Bel(H) and Pl(H), when computed by Dempster's rule, continue to give upper and lower bounds for Pr(H). (Note, however, that Bel(·) and Pl(·) are not bounds on what the future probability could be, given further evidence. They are bounds on Pr(·) implied by our present evidence.) A similar demonstration can be given for the case where E_1 and E_2 conflict. This approach can be generalized to the case where support is assigned to arbitrary subsets of hypotheses by regarding "reliability" as a set of separately assessed skills involved in discriminating subsets of hypotheses from their complements.

The problem, of course, is that we have not justified Dempster's rule as a bound on the Bayesian probability, $\Pr(H|E_1E_2)$. When we conditionalize on the evidence, as we certainly must in a Bayesian analysis, $\Pr(R_1 \text{ or } R_2)$ is replaced by

$$\Pr(R_1 \text{ or } R_2|E_1E_2) = \Pr(R_1|E_1E_2) + \Pr(R_2|E_1E_2) - \Pr(R_1|E_1E_2)\Pr(R_2|E_1E_2R_1).$$

This brings out a curious and critical feature of Shafer's theory. He is asking us to assess the reliability of a witness (or more generally, the status of an evidentiary process) without taking into account our knowledge of what the witness said. In Shafer's canonical example, knowledge of the evidence enters in only for the mapping from S to T, after all the probability work has been done on S. In a Bayesian analysis, on the other hand, the credibility of a witness can be shown to depend both on what is said and on its prior probability, i.e., our original tendency to think it true. If a witness says something which is independently believable, our estimate of his reliability increases. More importantly, perhaps, the credibility of one witness can, in a Bayesian analysis, be increased by corroboration of a second witness, and decreased by contradiction.

Assumption (b) is plausible only in light of this restriction. The strict Bayesian version of (b) is

$$Pr(R_2|E_1E_2R_1) = Pr(R_2|E_1E_2).$$

Note that E_1R_1 implies H, i.e., if witness 1 is reliable and says H, H is true. But we would expect, quite generally, that $Pr(R_2|E_2H) > Pr(R_2|E_1E_2)$, i.e., learning for a fact that what the witness said is true increases his credibility more than corroboration by a second witness. On the other hand, if we are assessing a witness's reliability prior to (or without consideration of) his testimony, it does make sense to require that his reliability be independent of the reliability of another witness. We thereby preclude shared uncertainties (e.g., a conspiracy) in the two evidential processes being combined.

A group of Swedish researchers, whose work is summarized and extended in Freeling and Sahlin (1983), and Freeling (1983), has explored issues such as this. Like Shafer, they focus on the reliability of the evidence, rather than the truth of the hypothesis, i.e., they reject the traditional Bayesian effort to model the chance of a hypothesis when the evidence is unreliable. But unlike Shafer, they analyze reliability in the light of the evidence, as Pr(R|E) rather than Pr(R). In effect, this is an effort to give a proper Bayesian account of the notion of quality or completeness of evidence, rather than truth. (As such, it is an alternative to the idea of second-order probabilities discussed in Section 2.4.8) The upshot of this research is that if m(H) is equated with Pr(R|E), Dempster's rule cannot in general be justified. Depending on the character of the belief functions being combined, and the kinds of conditional dependence assumed in the Bayesian analysis, Dempster's rule may be correct, a good approximation, or entirely off the mark in comparison to the "proper" Bayesian rule of combination.

While it fails to fully validate Dempster's rule, the Swedish work also lacks most, if not all, of the virtues of the belief function representation. In terms of feasibility, formulations which conditionalize on the evidence become extremely complex even for the simplest examples. The Swedish group has made little progress in deriving rules for the combination of evidence involving the full range of cases to which Dempster's rule applies, in particular, where varying degrees of support are assigned to arbitrary subsets of hypotheses. Moreover, the ~~requirement to assess~~ pervasive role of prior probabilities is incompatible with the segmentation of evidence into independent arguments which is a significant virtue of ~~vital for the naturalness of inputs in~~ Shafer's system.

Shafer (in press) explicitly rejects the attempt to provide any sort of Bayesian foundation for belief functions. Arguments based on Dempster's rule "have their own logic"--based on the appropriate canonical examples and an intuitive conviction that the appropriate conditions of independence are satisfied. As noted above, Shafer's appeal to intuition has not entirely succeeded in making that "logic" clear. We propose, however, that it can be clarified. In opposition to both Shafer and the Bayesians, we would argue the merits of the pseudo-Bayesian analysis of Bel(·) and Pl(·) as bounds on Pr(·), which we illustrated in this section. It fails to derive Dempster's rule as a <u>special case</u> of probability theory. Nonetheless, it clarifies the relationship of Dempster's rule to the canonical example, by an argument that <u>resembles</u> a valid Bayesian argument in most respects. Moreover, the dissimilarity can be crisply and clearly stated: the argument concerning reliability is conducted without consideration of the content of the evidence. The latter can be regarded as an explicit decision, justified by enormous gains in the simplicity and power of the calculus. This is not equivalent, however, to a fixed <u>belief</u> that the content of evidence is irrelevant. In an iterative, bootstrapping system, we can guard against the pitfalls of that assumption by continually reexamining it as an analysis proceeds. In Section 3.0 we explore the design of a system in which the function of recalibrating sources of evidence in light of corroboration or conflict is assigned to a process of qualitative reasoning.

2.5.10 Role of the assumptions in constructing an analysis. Conditions (b) and (c) play an important role as constraints in the construction of a belief function analysis. Violation requires reassessment of the overall structure of an analysis, redefining frames for either S or T or both (cf., Shafer, 1984a). (c) says that elements from both witnesses' testimony must not be required in order to construct a chain of reasoning that gets us to T. For example, if one witness said p and the other said $p \rightarrow q$ we would need to assume both were reliable to infer q. Therefore, these two statements must be counted as parts of a single evidential argument. In this sense, Dempster's rule combines self-contained "arguments" rather than "bits" of evidence. And application of the rule presupposes a more global process of reasoning addressed to problem structuring.

(b) and (c) represent a limitation on Dempster's rule in a second sense: Once our evidence has been segmented into independent arguments, we can combine it by Dempster's rule, but that rule tells us nothing about how two dependent pieces of evidence should be combined within a self-contained argument. For example, if we know "most C^3 installations are large rectangular buildings" and "most large buildings are near a road," what can we say about the chance that an object, known to be a C^3 installation, is near a road? Clearly, in any expert system application, Dempster's rule must be supplemented by other forms of inference. Interestingly, in a recent paper, Shafer (1984a) himself suggested that expert systems will have to make provision for dependent evidence, and that the full range of Bayesian operations can be applied on probabilities for the background frame, S. This is a departure from the position that only Dempster's rule is appropriate for combining evidence in the belief function context.

(*) separate lines

We have now noted three different ways in which an expert system application of Shafer's system might need to be supplemented:

- recalibration of sources of evidence in terms of the content of the evidence,
- reframing evidence and hypotheses to achieve independence of arguments, and
- reasoning about dependent evidence within an argument.

We may refer to this set of issues as the <u>incompleteness</u> of Dempster's rule, in analogy to the incompleteness of Bayesian theory discussed in Section 2.4.5. The system of qualitative reasoning proposed in Section 3.0 addresses all three.

2.5.11 Plausibility of instances: <u>Conflict of evidence</u>. To what extent does belief function theory yield inferences which are intuitive and plausible in specific applications? A topic of special concern in this regard is conflict of evidence. Zadeh (1984b) recently raised an example of the following sort. Suppose we have two experts who we believe to be very reliable and who produce conflicting judgments. For example, there are three possible interpretations of an object x in a specified location: H_1--x is a field; H_2--x is a forest; H_3--x is a building. Analyst A, using photographic evidence, assigns .99 support to H_1 and .01 to H_2; analyst B, using independent human intelligence information, assigns .99 support to H_3 and .01 to H_2. We have the following two support functions, and may combine them by Dempster's rule, as shown in Figure 2-6:

Table 2-2

	$m_A(\cdot)$	$m_B(\cdot)$	$m_{AB}(\cdot)$
H_1	0.99	0	0
H_2	0.01	0.01	1.00
H_3	0	0.99	0

The counterintuitive result, according to Zadeh, is that exclusive support is now assigned to H_2, a hypothesis that neither expert regarded as likely. Moreover, the result is independent of the probabilities assigned to H_1 or H_3.

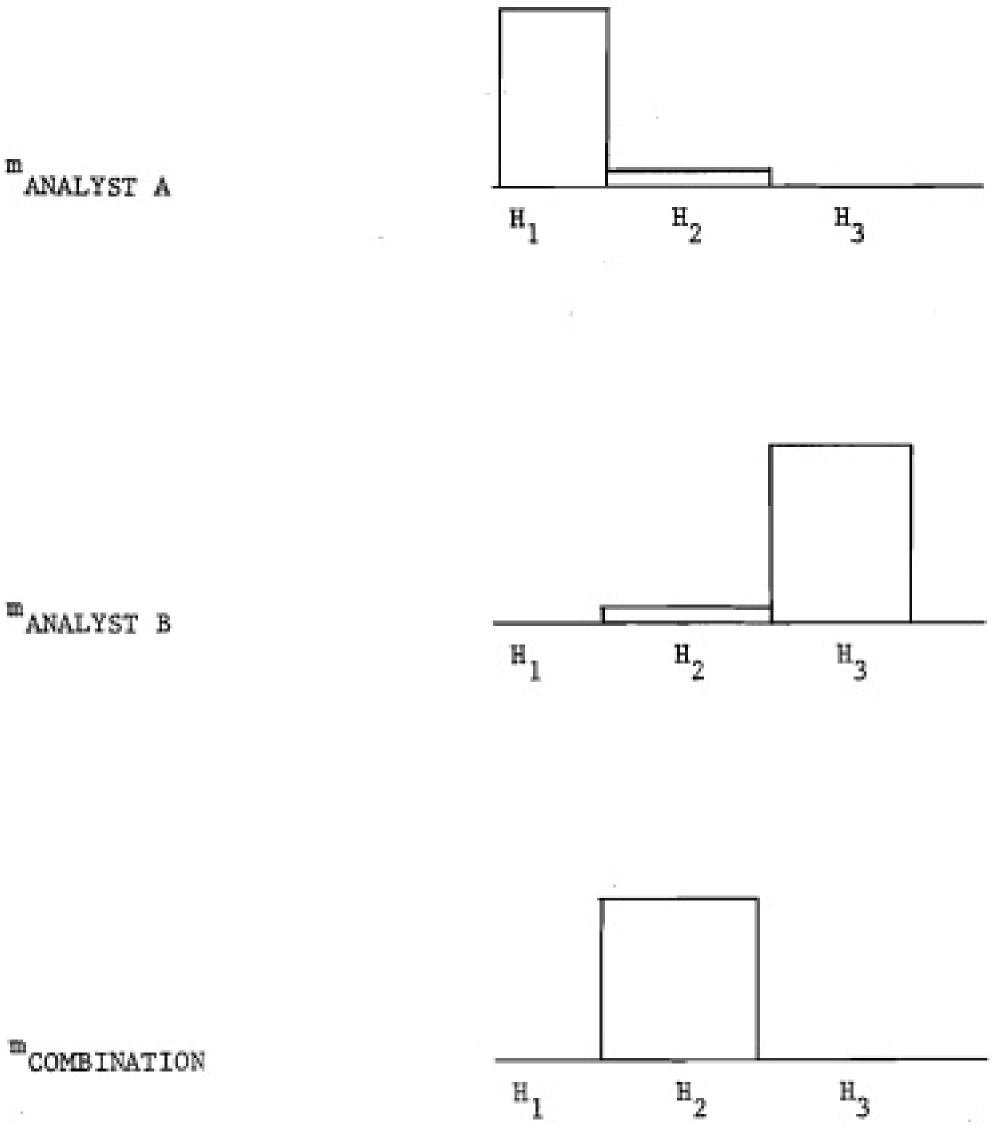


Figure 2-6. Support Functions to Illustrate Combination of Conflicting Evidence by Dempster's Rule

Shafer's response (in press) is cogent, but ultimately, we feel, off the mark. If we really regard these experts as perfectly reliable, Shafer says, the argument as stated is correct. After all, A says that H_3 is impossible, and B rules out H_1: that leaves H_2 as the only remaining possibility. (It is important to note that exactly the same result would be obtained in Bayesian updating, if we interpret the m(·) as likelihoods of the evidence given the hypothesis~~.~~) ~~and assume that prior probabilities for the three hypotheses are equal.)~~ On the other hand, Shafer argues that experts are seldom in fact perfectly reliable. A more reasonable procedure would be to "discount" the belief functions supplied by the experts to reflect our degree of doubt in the reliability of their reports. In discounting, we reduce each degree of support by a fixed percentage, and allocate the remainder to the universal set $\{H_1,H_2,H_3\}$. The result of applying Dempster's rule will now be a belief function that assigns support to all three hypotheses.

Let us examine this response in a bit more detail. Recalling that we regard these experts as highly reliable (though not perfect), suppose we discount A's belief function by 1% and B's by 2%. The result is the following, as depicted in Figure 2-7:

Table 2-3

	$m_A(\cdot)$	$m_B(\cdot)$	$m_{AB}(\cdot)$
H_1	0.9801	0	.656
H_2	0.0099	0.0098	.013
H_3	0	0.9702	.325
$\{H_1,H_2,H_3\}$	0.01	0.02	.007

We now have a "bimodal" belief function, with the preponderance of support going to H_1 and H_3. This appears, at first look, to be an intuitively plausible result: it reflects our feeling, which we represented in the form of discount rates, that A _or_ B (or both) could possibly be unreliable. But let us look a little more closely.

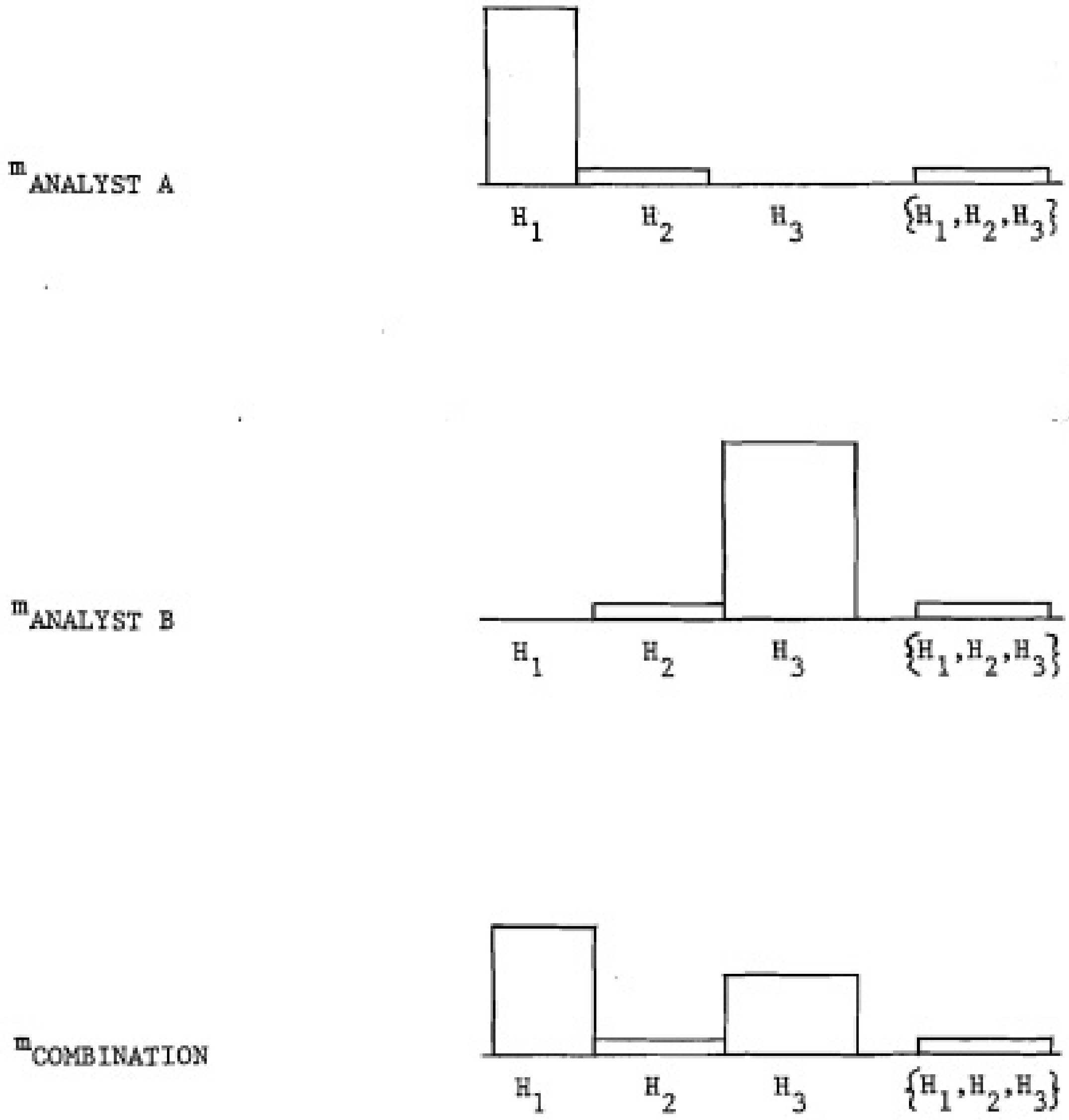


Figure 2-7. Support Functions to Illustrate Combination of Conflicting Evidence with Discounting

The first thing to note is what a vast difference a small amount of discounting makes. In Table 2-2, after combination by Dempster's rule, there was exclusive support for H_2. In Table 2-3, final support for H_2 is only slightly greater than 1%. The second thing to notice is the large discrepancy between $m_{AB}(H_1)$ and $m_{AB}(H_2)$. Although we did in fact discount B at twice the rate as A, the actual numbers (2% and 1%, respectively) and the difference between them were very small. It is by no means clear that the resulting difference in support for H_1 and H_3 is intuitively plausible. More to the point, the sensitivity of the result for all three hypotheses to very small differences in discount rates is disturbing. Finally, to dramatize the sensitivity even further, note that if support for (H_1,H_2,H_3) were 0 for both experts, and if A assigned 0 support to $H_{\cancel{3}}^{2}$, and B assigned 0 support to $H_{\cancel{1}}^{2}$, these very small changes render Dempster's rule indeterminate.

Perhaps the problem is that our original assessment of the reliability of the experts was mistaken. Suppose then we discount A by 29% and B by 30%. We now get:

Table 2-4

	$m_A(\cdot)$	$m_B(\cdot)$	$m_{AB}(\cdot)$
H_1	.7029	0	.4243
H_2	.0071	.007	.0085
H_3	0	.693	.4044
(H_1,H_2,H_3)	.29	.30	.1751

Support for H_1 and H_2 after combination is now roughly equal, certainly a more intuitive result. Then should we have discounted A and B more in the first place? According to Shafer, presumably, this is indeed the case; the fault is not in the theory, but in the initial allocation of support. The example, however, highlights a deeper problem. As we noted in Section 2.5.5, reliability is to be assessed <u>as if</u> we had no knowledge of the evidence actually provided. Thus, we are apparently not permitted to *use* the conflict between A and B as a clue regarding their capabilities or as a guide to the appropriate amount of discounting. We return to this issue very shortly.

Zadeh himself objects to the procedure in Dempster's rule of normalizing support measures to eliminate impossible combinations. But we think this objection is mistaken. Normalization is in fact the *only* way in Shafer's theory (albeit quite indirect) that our knowledge of the evidence enters into the assessment of reliability. It accomplishes a sort of *de facto* discounting as a function of conflict of evidence. Note in the earlier example of Figure 2-5 that the reliability of witness 1, after combining his testimony with the conflicting evidence of wit ness 2, is $(P_1P_2'/(1-P_1P_1'))$. This is less than P_1, the original assessment of witness 1's reliability.

Although normalization is in itself not problematic, nevertheless, it is not a complete or adequate solution to the problem of conflict. First, because there is no lasting effect on later problems, i.e., we have not truly updated our estimate, P_1, of A's reliability in the light of his conflict with B. Second, there is no procedure for exploring potential *reasons* for the conflict. A closer examination of (a) the factors that determined our original reliability estimates, (b) our assumptions regarding independence of the two arguments, and (c) the internal structure of the arguments employed by A and B, might lead to a revision in beliefs and assumptions that permanently improves our knowledge base.

We argue, then, that the revision of reliability estimates is only one possible result of an iterative, constructive process of problem solving prompted by conflict of evidence. (We also have the options of reframing evidence and hypotheses to reflect revised judgments of independence and of revising specific beliefs internal to the conflicting arguments. These are the alternatives outlined at the conclusion of Section 2.5.10). Therefore, such revisions must be justified by considerations which, once discovered, carry weight independent of the conflict of evidence that led to their discovery. Ideally, these newly discovered factors could be regarded as sufficient to justify revisions in reliability estimates independently of E_1 and E_2. (Referring to these factors as F, we would have

$Pr(R_1|E_1E_2F) = Pr(R_1|F)$.) This justifies the reassessment of reliabilities in the light of the evidence in the Shafer-Dempster system, and is the method implemented in the system to be described in Section 3.0.

2.5.12 What is "conflict of evidence"? So far, we have taken for granted the notion of conflicting evidence, and that in some cases at least special steps are justified in dealing with it. But it is by no means obvious what "conflict" is, or why steps outside the normal calculus of uncertainty should be required to handle it. Conflict of evidence does not appear, on the surface, to be the same as incoherence. The formal constraints of Bayesian theory dictate, as we saw in Section 2.4.5, that multiple probabilistic analyses should agree with one another and with direct judgment. Similar coherence constraints can be derived for Shafer's theory from the requirement that uncertainty on S be measured by a probability. But it is implicit that these analyses are, or should be, based on the same evidence. There appears to be no corresponding guarantee or prescription that arguments based on different evidence should arrive at the same or similar conclusions. Dempster's rule is designed explicitly to combine arguments based on independent evidence; hence, there are no direct constraints on the extent to which those arguments must agree (except that there be at least one pair of meanings from the two arguments whose intersection is non-empty).

Nevertheless, we propose that the resolution of conflict in a belief function analysis be construed as a desire for coherence. The missing element, which is responsible for the incoherence, is a judgment, often implicit, regarding the overall structure which the final belief representation is expected to have. Such judgments are based on one's knowledge about reasoning in a particular problem domain. "Conflicting evidence" is evidence whose combination produces a structure that violates such a prior expectation. Thus, the definition of "conflict" will vary from one problem domain to another. The locus of conflict is not, strictly speaking, between the two sources of evidence, but between both of them, on one side, and a structural expectation regarding the outcome of the argument, on the other. When a conflict of this sort occurs, in an iterative, constructive context, the decision maker has a choice of either revising the expectation or else making one or more of the three kinds of changes we discussed above (revising discount rates, frames, or steps in an argument).

If belief functions are probabilistic with discounting (i.e., assign support only to single hypotheses and to the universal set), then it is often plausible to require that hypotheses which receive very little support from either of two arguments not receive predominant support in the combined analysis. This was the basis of the adjustment of discount rates in the above example (and also seems to underlie the use of discounting in Shafer, 1982). Note that an analogous requirement is recommended for Bayesian analysis by deGroot (1982).

Other possible structural expectations regarding the form of a belief function model include that it be consonant or hierarchical. In these cases, support is assigned only to nested subsets of hypotheses or to subsets that form a tree, respectively. Neither of these properties is necessarily preserved through combination by Dempster's rule. Yet, as we noted in Section 2.5.3 above, such structural constraints may (a) be quite plausible for particular problem domains (cf., Gordon and Shortliffe, 1984, on medical diagnosis), and (b) be required to reduce the computational intractability of a Dempster-Shafer model. Thus, once again, a higher-order process of qualitative reasoning may be necessary to explore revisions in beliefs and assumptions, in order to handle "conflict" and to ensure the applicability and plausibility of a Dempster-Shafer calculus (see Section 3.0 below). ⊛

An important by-product of requiring consonance should be noted. One potential criticism of Shafer's theory is that it lacks a concept of the <u>acceptance</u> of a hypothesis once it achieves a sufficient degree of evidential support (e.g., Levi, 1983; L.J. Cohen, 1977). A precondition of acceptance--and what makes it a useful concept in some contexts--is that it should yield a logically consistent and complete story. Neither is true if a threshold or cutoff for acceptance is defined on Bel(·) in Shafer's system. Both a hypothesis and its complement could have positive support, and thus conceivably both could be accepted, yielding a contradiction. Moreover, two propositions, p and q, might be accepted but their conjunction, p&q, rejected. Both of these problems disappear in a consonant belief function: Since a hypothesis and its complement are not nested, they can not both receive support; and it can be shown that Bel(p&q) = MIN(Bel(p),Bel(q)) and thus that a conjunction is at least as credible as either of its conjuncts.

In all these cases, there is a tension between the desirability or plausibility of depicting the state of evidence "as it is," conflicts and all, and attempting to produce a resolution or reconciliation within the framework of some plausible or desirable global requirement. We claim that this tension is at the heart of any truly intelligent and flexible reasoning with probabilistic systems.

2.5.13 Summary. Shafer's theory provides a natural representation of quality of evidence and relaxes the assessment requirement to the extent that the evidence is incomplete. Like Bayesian theory, however, belief function models impose inordinate input and computational demands unless specialized models are adopted. The validity of Shaferian theory has not been clearly established, although it may be illuminated by a partial Bayesian derivation. A major difference is that Shafer's theory does not permit reassessment of the quality of an information source in terms of what that source says; the credibility of one witness cannot be increased by corroboration of a second witness or decreased by contradiction. In belief function theory, the outcome of combining the information from two conflicting data sources can vary dramatically, depending on our assessment of their credibility. Yet we cannot use the two sources to crosscheck one another. We argue that this gap in Shafer's theory requires that it be supplemented by a process of qualitative reasoning that reexamines sources of evidence as an analysis proceeds, and recalibrates them in the light of corroboration or conflict. The same process might supplement Shafer's theory in other ways: by reframing evidence and hypotheses to establish independence of evidential arguments, and by revising inferential steps which are internal to such arguments.

2.6 Fuzzy Set Theory

2.6.1 Nature of the theory. Since L.A. Zadeh advanced fuzzy set theory in 1965, an enormous amount of interest, and a very large literature, has been generated. Most of this interest has been theoretical, concerned with the mathematical implications of the theory, but there have been a number of attempts to apply the theory to practical problems. This is in ~~line~~ the spirit of with Zadeh's original reason for introducing the concept. ~~He argued that much systems analysis was inadequate because its requirements were too precise.~~ He felt that our intuitive understanding of concepts and, more interestingly, our reasoning about those concepts, were typically imprecise, yet most systems analysis ~~(especially with computers)~~ required precisi~~fication.~~on. To resolve this paradox, he introduced the now well-known concept of the fuzzy set--a set with imprecise boundaries. The essential element is the membership function $\mu_A(x)$ which represents the degree to which an element x belongs to some set A. If $\mu_A(x) = 1$ then x ~~indisputably~~ fully belongs to A, while if $\mu_A(x) = 0$, x does not belong to A. An intermediate value, such as $\mu_A(x) = 0.6$, indicates that x belongs to the set to some degree. Fuzzy sets are thus a precise tool for representing and manipulating imprecise notions.

Application of fuzzy set theory involves: first, the representation of ordinary imprecise concepts by fuzzy sets; second, the use of a calculus to construct other fuzzy sets which represent~~ing~~ the output variables ~~in an~~ of the analysis; and third, reinterpretation of the results in ordinary imprecise language (see L.A. Zadeh, 1975). The first and last steps are crucial if the flavor of the fuzzy theory is to be fully captured. The core idea is to construct a calculus for the formal (i.e., precise) manipulation of imprecise concepts, which takes in imprecise inputs and puts out imprecise outputs.

2.6.2 Applications of fuzzy set theory to inference. The theory of fuzzy sets can be applied in many ways, in the sense that wherever a mathematical relationship exists, it can be fuzzified. Thus, there are many possibilities for using the fuzzy calculus in conjunction with other inference theories. Alternatively, it can be applied directly to ordinary imprecise reasoning (by experts or non-experts) in natural language. We will introduce some of the formalism of fuzzy set theory by examples of these two types.

2.6.3 Fuzzy implication. Suppose further a rule for an image interpreter could be written:

> "If the texture is rough, and the illumination is good, then the object is a forest."

Suppose we have two facts:
"The texture of a region is very rough."
"The illumination is not very good."

To express this rule using fuzzy set theory, we need to define the input fuzzy sets. The first will be $\mu_R(t)$, which measures the extent to which a particular texture-vector t can be said to belong to the set of 'rough' texture vectors. The second will be $\mu_G(i)$, the extent to which an illumination level, i, can be said to be 'good.' The third will be $\mu_F(x)$ describing the 'forest'-ness of the object: x is some variable which gives a precise categorization of each object and $\mu_F(x)$ will be a fuzzy-set on the variable x.

The first manipulation will be to represent $\mu_{RG}(t,i)$, the extent to which an image with texture-vector t and illumination level i can be said to be both "rough" and "good." Zadeh's calculus suggests that this is the minimum of the two membership functions:

$$\mu_{RG}(t,i) = \min(\mu_R(t),\mu_G(i)).$$

Implication in fuzzy set theory is defined as a relation. Thus, "if U is F, then V is G," where F and G are fuzzy sets on the variables u and v underlying U and V, is described by the relation

$$\mu_{V/U}(u,v) = \min(1,\mu_2(v) + 1- \mu_1(u))$$

using an obvious notation. This may be interpreted as the extent to which a particular value of U implies a particular value of V.

The next step is to combine the rule with a statement about the fact described in its antecedent. In fuzzy implication, not only may be the concepts involved be fuzzy, but the match between a fact and the antecedent of a rule may be a matter of degree as well. Thus, we may have a rule stating "If U is F then V is G," but
* an input stating that "U is F*", w~~j~~here F ~~amd~~ and F* are not the same. Zadeh defines this as

$$\mu_Y(v) = \max_u(\min(\mu_{F*}(u),\mu_{V/U}(u,v))).$$

where Y is the fuzzy set that results from combining F* and V/U. Thus, in our example, suppose $\mu'(t,i)$ is ~~a~~ another fuzzy set on the variables for texture and illumination, t and i. $\mu'(t,i)$ may reflect an input to the effect that ~~the~~ texture of the region is "very rough" and the illumination is "not very good." We find that

$$\mu_Y(x) = \max_{t,i}(\min(\mu'(t,i),\min(1,1-\min({}_R(t),\mu_G(i))+ {}_F(x))))$$

is the induced fuzzy set on the categorization variable, x. $\mu_Y(x)$ is a quantitative measure of the possibility that the object is a forest given the fuzzy evidence regarding roughness and illumination and the fuzzy implication rule. The output may now be translated into an imprecise natural language expression (e.g., "very possibly a forest") corresponding to $\mu_Y(x)$.

2.6.4 Fuzzy probabilities. Uncertainty about facts (i.e., chance) was not mentioned above; we just talked about imprecision. Zadeh stresses that the two concepts are distinct, and that fuzzy set theory should only be used to describe imprecision. If we are imprecise our uncertainties, however, then a role exists for describing that imprecision with fuzzy sets. Watson et al. (1979) and Zadeh (1981) discuss this idea in the context of decision analysis, but it can clearly be applied to any use of Bayesian probability theory, or belief function theory.

The basic tool for fuzzifying a calculus is Zadeh's extension principle, which enables us to compute the fuzzy set membership function for a variable when it is a function of variables whose fuzzy set membership functions are known. Let $Y = F(X_1,X_2,\ldots,X_n)$. Then $\mu_Y(y) = \max[\min(\mu_{X_1}(x_1), \mu_{X_2}(x_2),\ldots, \mu_{X_3}(x_3))$ where $\mu_Y(y)$ is the extent to which a value y belongs to the set of possible numbers for the output variable.

Suppose a scene labeling procedure leads to a probability p that an object should be classified as a building. Imagine we have a loss function which gives unit loss if misclassification occurs, and zero loss if not. Then the expected loss from classifying the object as a building is

$$1 \times (1-p) + 0 \times p = 1-p$$

while the expected loss from classifying the object as 'not a building' is

$$1 \times p + 0 \times (1-p) = p.$$

Clearly, we minimize expected loss by categorizing it as a building if $p>1/2$. Now suppose that we are imprecise about p to the extent that we can only describe a fuzzy set $\mu(p)$ about possible values of p. Fuzzy sets for the expected loss in the two cases (actually $\mu(1-p)$ and $\mu(p)$) can be produced using Zadeh's extension principle. But what conclusions can we draw? Freeling (1980) discusses this in some detail, suggesting several alternatives approaches. As we might expect, when results are fuzzy, the analyzis may not indicate any particular decision regarding classification.

As with the Bayesian analysis, there are some non-trivial problems in attempting to apply fuzzy set theory to inference in expert systems.

2.6.5 Feasibility. We criticized both Bayesian theory and belief function theory on the grounds that the analysis involved in practical problems can be quite complex. This will also be true of fuzzy set theory. The fact that functions of variables have to be handled in computations makes the analysis difficult to handle numerically. Nonetheless, there are indications that the max-min operations are numerically easier than the sum-product operations of the other theories. It would be wrong, however, to assert that the use of fuzzy set theory removes all of the difficulties caused by complexity in the other two theories examined here.

2.6.6 **Validity**. For a theory which has had an enormous literature, there is still a considerable discussion amongst scholars on the justification and interpretation of the theory.

2.6.7 **Semantics**: Where do the numbers come from? This question is raised by

most people when they first study fuzzy set theory. There are no standard procedures to be applied in every case; anything plausible would seem to do. In particular, there are neither behavioral specifications nor canonical examples of the kind Shafer claims to be important. Zadeh would argue that a theory of imprecision should not need precise inputs, so that we should not bother too much over the exact nature of the imput membership functions. If that is the case, then answers should not be very sensitive to input membership functions.
In many applications, this is not the case, and indeed, sometime~~s~~ answers are sensitive to just one point on a membership function.

as a result of the max-min connectives

What is the meaning of the output? Paralleling the uncertainty relationship between human perceptions of imprecision and the calculus of fuzzy sets is the reverse relationship: once we have computed an output fuzzy set, what do we do with it? We briefly discussed the possibility of linguistic interpretation above. This does not appear to have been a satisfactorily implemented approach, ~~although~~ in part because people differ in the conclusions they draw from the same natural language statement.

In the light of these difficulties, it is not surprising that efforts should be made to assimilate fuzzy sets to some other framework of uncertainty, such as the Bayesian or Shaferian. It is difficult to do this in a natural way, however, due to the difference between imprecision and uncertainty about facts. For example, suppose Analyst A refers to an object x as "long", after having measured x exactly. There is no doubt as to x's actual length, and although A may regard x as long only to a certain degree, he is not uncertain whether or not x is long. What fact then could A be uncertain of? We add three caveats: (i) if A tells a second Analyst B that x is long, then B may be uncertain regarding x's actual length; (ii) if A had only glanced at x, rather than measuring it, he might be uncertain (as well as imprecise) about x's actual length; (iii) we may in fact be uncertain as to whether a random English speaker would call the object "long". Nevertheless, the most natural approach is to treat this kind of uncertainty as the degree to which x (or an object of x's length) is long, rather than the chance that x is long. Put another way, these degrees are part of the meaning (denotation) of "long", and not (necessarily) a result of uncertainty about what "long" means or about the actual length of an object.

Nonetheless, it may be worthwhile exploring ways to represent imprecision in terms of other frameworks. For example, a consonant Shaferian support function (Section 2.5.3 above) obeys a calculus that closely approximates Zadeh's possibility theory. Consonant support functions seem appropriate for representing imprecision in the implications of evidence (it points to a set of nested regions where the truth could lie). And they have the advantage of a somewhat more secure normative foundation (Sections 2.5.5 - 2.5.11 above). Thus, the possibility of translating between natural language expressions and support functions might be worth exploring, despite some cost in naturalness.

2.6.8 **Inference:** What are the appropriate connectives? In terms of either axiomatic justification or face validity, the procedures Zadeh recommends for combining his membership functions are not unique. For example, Zadeh argues that the degree to which an element belongs to a set A_1 and another set A_2 should be computed by

$$\mu_{A_1 \cap A_2}(x) = \min(\mu_{A_1}(x), {}_{A_2}(x)).$$

This is clearly consistent with the requirement that if both sets are crisp (i.e., only takes the values 0 or 1), set membership should obey the usual rules (i.e., x $\mu A_1 \cap A_2$ if and only if $x \varepsilon A_1$ and $x \varepsilon A_2$). Note however, that this is not the only connective rule with this property. For example, the family of connectives

$$\min(\mu_{A_1}(x)\mu_{A_2}^{1-\alpha}(x),\ \mu_{A_2}(x)\mu_{A_1}^{1-\alpha}(x)), \qquad 0 \leq \alpha \leq 1.$$

all have this property, where $1-\alpha$ is a power to which the membership function is raised. Zadeh chooses $\alpha = 1$; the choice of $\alpha = 0$ gives the Bayesian rule for the probability of a conjunction (namely $\mu_{A_1}(x) \cdot \mu_{A_2}(x)$). There are many other possible definitions (see Dubois and Prade, 1984).

Similarly, disjunction, negation and implication all have alternative representations, and the choice of the forms usually employed is arguable. So far as we are aware, very little research has been carried out on the implications of using different connectives on the results of a fuzzy analysis. There is, therefore, some arbitrariness in the connectives chosen by Zadeh--an arbitrariness which pervades the theory.

2.6.9 **Plausibility of instances**: The main strength of Zadeh's theory is in its ability to produce instances of reasoning that are acceptable on a case by case basis. In this regard, it has a richness and scope that no other theory even attempts to capture. In particular, it is the only theory that attempts to formalize the combination of considerations based on similarity (e.g., the closeness of F* to F in the above example) with more traditional considerations in inference (e.g., traditional logic or probability). In this largely uncharted domain, the (present) absence of deep normative foundations may be no disgrace.

Nonetheless, there may be cases where fuzzy logic gives implausible (or non-useful) answers. Fuzziness is concerned with what is possible, rather than what is probable. Zadeh sees a possibility distribution as being an upper bound on a probability distribution. Articulating the possible may be important, but if many options are possible, it does not help in our search for what is probable. In practice, this point is expressed by the tendency for fuzzy sets to produce rather bland answers, giving high values of the membership function for large sets of variables. One can see some applications when this is not an obstacle to understanding, if some important options are seen to have very low or zero possibility. In general, it does present a difficulty.

2.6.10 Summary. Fuzzy logic is a highly flexible and versatile tool for handling imprecision. It may be applied directly to reasoning with verbal expressions or, at a higher level, to reasoning with a numerical calculus like probability theory. Unfortunately, the meaning of fuzzy measures is not always clear; and the rules for manipulating them seem to lack any deeper justification than the plausibility of the answer in a specific application.

2.7 Non-Monotonic Reasoning

In this section we turn to a quite different approach to reasoning under conditions of uncertainty. Although non-monotonic reasoning emerges directly from the tradition of non-numerical reasoning in artificial intelligence, it is designed to address problems of incomplete information. The basic ideas of non-monotonic reasoning were first applied by Stallman and Sussman (1977) in a system for electronic circuit analysis. Since then, theoretical work has been associated with Doyle (1979), McDermott and Doyle (1980), Reiter (1980), McCarthy (1980), and others.

2.7.1 Nature of the theory. Traditional, axiomatic formal systems are monotonic, in the following sense: beginning with an initial set of premises, the number of provable statements or theorems of the system increases monotonically in time as new axioms or premises are added on.

In contrast, the content of practical structures of argument and belief may diminish as well as increase. New data may compel an analyst to challenge and reject previously derived conclusions. Such systems are non-monotonic in time. Humans become skilled at merging conflicting data into existing arguments or beliefs so as to regain consistency while minimally disrupting the established systems. Non-monotonic logic is the name associated with a set of formal and computer-based systems designed to incorporate new, conflicting data into systems of belief based on incomplete information.

2.7.2 Dependency-directed backtracking is a key concept in implementing non-monotonic systems. As data and constraints are added to a non-monotonic system, they are treated as valid until a contradiction is found. Traditional systems, in the face of a contradiction, must backtrack past the data that was added immediately prior to the contradiction, searching for a new path that is contradiction-free. Many dead-ends are likely to be encountered in an exhaustive search of this type before a consistent total set of beliefs is found. In a non-monotonic system, only those beliefs which actually contributed to the contradiction need to be considered.

Dependencies among statements in a non-monotonic system (Doyle, 1979) are represented (primarily) by data structures called support lists. A support list justification for a statement has the form

Statement #   statement   (SL <*in*list> <*out*list>).

Such a justification is a valid reason for belief in the statement if every statement in its *in*list is believed, and every statement in its *out*list is not believed. For present purposes, we can distinguish three kinds of justification in these terms:

(1) A *premise* justification has an empty *in*list and an empty *out*list; i.e., (SL()()). Thus, nothing else needs to be demonstrated, or not to be demonstrated, to ensure acceptance of a statement with such a justification. Observational data (or unquestioned general principles) might be treated in this way. For example,

N-1   Object has texture of type x   (SL()())

is automatically regarded as IN.

(2) A *monotonic* justification has a non-empty *in*list, but an empty *out*list. For example,

N-2   Object is a building   (SL(Object has texture of type x) ())

is a monotonic justification. Note that it corresponds to the example discussed in Section 2.4: This type of node simply states that *if* certain other facts are believed (e.g., texture is type x), *then* the relevant statement should be accepted (e.g., the object is a building). N-1's being IN, in conjunction with this justification for N-2, is sufficient to cause N-2 to be IN.

(3) If only monotonic justifications exist, no statements can be retracted. Hence, they are appropriate only if all possible evidence is explicitly stated in

the inlists corresponding to various possible conclusions. In other words, we must resolve not to accept any statement until we possess all the information regarding its truth or falsity that we ever intend to regard as relevant. In this example, N-2 would make sense only if texture was the sole clue relevant to classifying an object as a building. More typically, we cannot afford to be this conservative. We may wish to accept a statement provisionally, to act "as if" it were true, and to use it in subsequent reasoning, based on only a subset of the possible observations. The appropriate means for doing so is via a non-monotonic justification, i.e., a support list whose outlist is non-empty. Statements with non-monotonic justifications are called assumptions. The inlist states the conditions (if any) under which it is desirable to assume the truth of the statement; the outlist states the conditions under which the assumption would have to be rejected. Thus, to continue the example, a more appropriate version of N-2 might be:

```
N-2'   Object is a building    (SL(Object has texture of type x)
                               (Object is far from road))
```

In other words, if we know the texture of the object to be x, we can assume the object is a building as long as we have not proven that it is far from the road. Thus, N-1's being IN, in conjunction with this justification for N-2', is still sufficient to cause (provisional) acceptance of the statement that the object is a building. The assumption is appropriate even if we have as yet collected no data at all regarding the object's distance from a road. But suppose we now collect such data and as a result add the following premise to our system:

```
N-3                    Object is far from road               (SL()()).
```

N-3's being IN is now sufficient to cause N-2' to go OUT.

The latter is an extremely simple example of dependency-directed backtracking. Let us spell out the steps in a bit more detail. N-2' and N-3 being jointly IN is detected by the system as a contradiction. The system then sets up a CONTRADICTION node with N-2' and N-3 in its inlist:

N-4 CONTRADICTION (SL(N-2' N-3)()).

N-4 states a "local constraint" governing the relationship of N-2' and N-3: they cannot both be IN. Note, however, that N-4 is IN only so long as N-2' and N-3 are IN. The system now searches for the set S of assumptions (i.e., statements with non-empty outlists) which are responsible for the CONTRADICTION node N-4; in other words, S contains the assumptions whose being IN has caused N-2' and N-3 to be IN. The system then sets up a NOGOOD node as a permanently IN record of the inconsistency of S. This node has the form:

Statement # NOGOOD S (CP(CONTRADICTION)(S)())

where CP is a conditional-proof type of justification. Essentially, the NOGOOD node is justified by the relationship between S and the CONTRADICTION, independently of whether the CONTRADICTION happens to be IN or not. In our example, there is only one assumption responsible for N-4's being IN, and that is N-2' itself. Thus, we get the following:

N-5 NOGOOD N-2' (CP(N-4)(N-2')()).

In this case, the CP justification is valid (and N-5 is IN) because N-4 is IN whenever N-2' is IN.

The next step is crucial in more complex examples. The system selects a "culprit" C from the members of S, i.e., it identifies some one assumption among those collectively responsible for the problem and decides to deny that assumption. To do so, it further selects some member O of the outlist of the culprit. It then sets up a support list justification for O. This justification says, in effect, that if you want to keep all the other assumptions in S (except C), and if you have not proven any of the other grounds for retracting C, then you should believe O. (The inlist of this justification contains all the assumptions in S, except C, together with the NOGOOD node; the outlist contains all the members of the outlist of C exno other – adequately justified reason for rebutting C.

cept O.) The result is that O is (provisionally) treated as IN; C is retracted; and the CONTRADICTION node goes OUT. Of course, O is only an assumption; later contradictions may lead to its retraction and to the use of some other member of the <u>out</u>list of C, or else to the restoration of C and the denial of some other assumption in S.

Although in our example this process is trivial, it does illustrate another important aspect of the truth maintenance system. In our example, as noted, dependency-directed backtracking must select N-2' as the "culprit" for denial. Since N-3 is the only member of its <u>out</u>list, N-3 receives a new justification. It now appears as

N-3'	Object is far from road	(SL()()) (SL(N-5)()).

It appears that N-3' can be justified <u>either</u> as a premise (data) <u>or</u> an assumption required to resolve the inconsistency represented by N-5. This, however, is wrong. The second justification is circular, since it was N-3 that led to the inconsistency in the first place. Doyle's Truth Maintenance System guards against circular justifications of this sort, by designating certain justifications as "well-founded" and others as not.

We now turn to a somewhat more detailed example.

2.7.3 <u>Example of informal non-monotonic reasoning</u>. An image analyst is shown two images taken from a platform directly above the object of interest, a rectangular structure on the deck of a vessel. The images are taken at different times of day. The sun angles and the height of the platform above the vessel are known, and the analyst is tasked to measure the object and make some inferences about its structure. The images are shown below:

Image #1

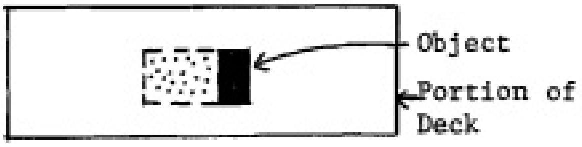
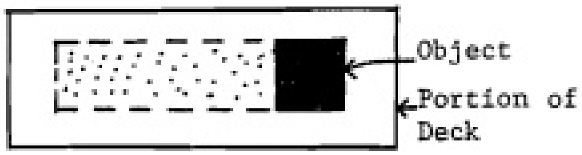


Image #2

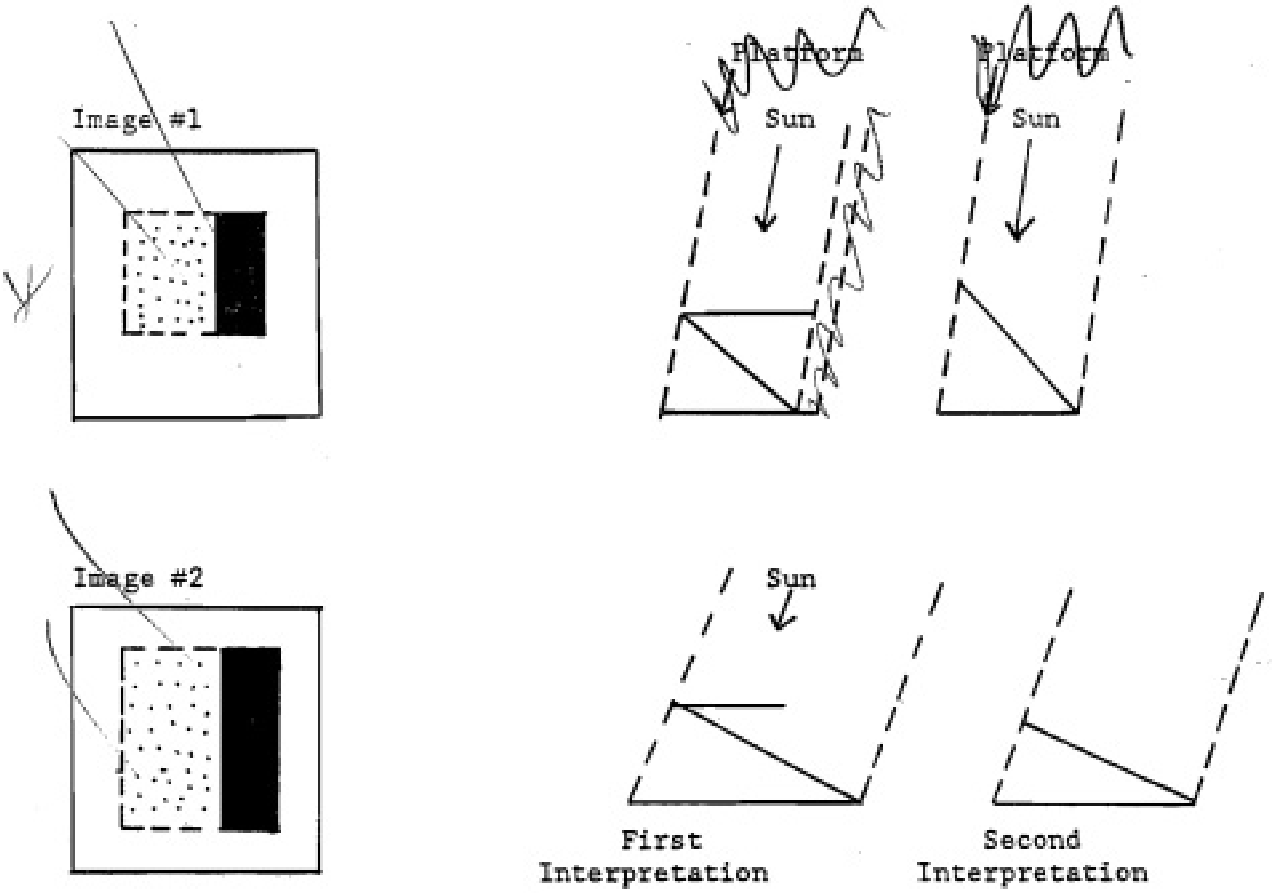


A question of particular interest is whether the dark "object" is a hole in the deck through which the dark interior of the vessel's hold is seen, or a solid structure on or above the deck.

The analyst might reason quickly as follows:

"The object is uniform in reflectance, therefore, probably planar. It casts a shadow, therefore, must be an opaque structure elevated above the deck. From the distance between the left-hand edge of the shadow and left-hand edge of the object, I can measure the height of the object above the deck."

"There's a problem with this simple model. The shadow in the second image is much longer than the object. Therefore, either the object is a planar structure attached to the deck at some angle, or if it is a horizontal planar structure it must be supported by some other structure, invisible to me, that contributes to the shadow." The analyst might proceed to sketch several configurations that are consistent with the data:



The analyst has quickly noted and resolved two inconsistencies: First, the existence of the shadow doesn't jibe with the theory that the dark object is an aperture in the deck, so this hypothesis is ruled out. Second, the size of the shadow in the second image doesn't fit the theory that the object is a horizontal plane suspended above the deck; this is ruled out and ~~rephased~~ replaced with the "leaning wall" and "planar support" hypotheses, as illustrated.

2.7.4 Application of a non-monotonic system. We will next illustrate how this argument would be treated in a non-monotonic reasoning system. We assume that object recognition and feature extraction have been performed, either by an analyst or by a machine, and that these data have been represented in computer-compatible form. The image-processing system or analyst will have recognized objects and shadows and will have measured the distances from object to shadow boundaries. A set of plausible hypotheses (flat object on surface; aperture in deck; tilted object) will have been formulated and recorded as statements. The resulting data set is as follows:

Statement #	Statement	State		Support List	
		IN	OUT	In	Out
1	Object is aperture in deck.	X		5,7	2,3,4 6,8,9
2	Flat object lying flat on deck.		X	5,7	1,3,4 6,8,9
3	Flat, horizontal object ~~supported~~ suspended above deck.		X	6,8	1,2,4 5,7,9
4	Flat object, tilted at angle to deck.		X	6,9	1,2,3 5,7,8
5	At sun angle θ_1, object is uniformly bright, casts no shadow.		X		
6	At sun angle θ_1, object is uniformly bright, casts a shadow of dimension less than object.	X			
7	At sun angle θ_2, object is uniformly bright, casts no shadow.		X		
8	At sun angle θ_2, object is uniformly bright, casts a shadow smaller than object.		X		
9	At sun angle θ_2, object is uniformly bright, casts a shadow larger than object.	X			

As in our earlier discussion, a statement is IN or OUT at any given time depending on whether or not it is justified based on evidence currently available. The justification for a statement being IN or OUT is based in turn on certain other statements being IN or OUT. The support of a given statement is the set of statements required to be IN or OUT for that statement to be IN. Thus, the statements and the justification relationships form a tangled network. The set of IN statements grows and shrinks in a non-monotonic fashion as new evidence changes the states of particular statements, and as the effects of these changes propagate through the network. (The set of *justifications*, however, grows monotonically.)

For example, the support list of statement 1 is (SL(5,7)(2,3,4,6,8,9)). To see how the system deals with conflicts between data and observations, let us assume the analyst starts by assigning IN as the state of statement 1. The observation data states are:

5,7	OUT	(Object *does* cast a shadow)
6	IN	(At sun angle θ_1, object casts a small shadow)
8	OUT	
9	IN	(At sun angle θ_2, object casts a large shadow)

The non-monotonic system checks the network for consistency among the states and support sets, notes an inconsistency, and introduces a new conflict assertion:

Statement #	Statement	State		Support List	
		IN	OUT	In	Out
10	CONTRADICTION	X		1,6,9	5,7

The system proceeds to resolve this conflict by changing statement states; observation data is challenged only as a last resort. For efficiency, the system may attempt first to achieve consistency with a subset of the observation data, since this is potentially a large data base. In our example, the system works initially with the (5,6) observation data, and subsequently considers the (7,8,9) data. Initial consistency is achieved by setting statements 1 and 2 to OUT and statement 3 to IN, retaining statement 4 in the OUT state. Statement 10, CONTRADICTION, reverts to the OUT state (although the system retains a permanent trace of this conflict "proof" for subsequent possible activation.)

Since statements 7,8,9 are not being considered at this moment, statement 3 IN is consistent with the data (5 OUT, 6 IN).

Next, the system broadens its scope to consider a larger piece of the data base.

A new CONTRADICTION statement is generated:

11	CONTRADICTION	X	3,9	8

To resolve this conflict the system considers new state settings. Resetting statement 1 to IN is disallowed by the trace of the previous conflict. The correct solution setting statement 3 to OUT and statement 4 to IN achieves consistency.

The scenario sketched above illustrates the truth maintenance feature to be found in deductive retrieval systems, such as DUCK (McDermott, 1983). Non-monotonic reasoning is very much, however, an active area of AI research, with open questions remaining both in feasibility and validity.

2.7.5 **Feasibility.** Dependency directed backtracking is a species of discrete relaxation (like Walz filtering, as described in Cohen and Feigenbaum, 1982). It seeks a consistent allocation of truth values across a set of statements, by utilizing local consistency constraints between pairs of statements, rather than by exhaustive search through the space of all possibilities. Thus, a high level of computational efficiency can be achieved.

To make this efficiency possible, however, in non-monotonic systems, the traces of proofs are retained, even though the premises utilized by the proof, and the statement that was proved, may (temporarily) be judged invalid or OUT. Therefore, if the premises become valid or IN at some later time, the work of rediscovering the proof need not be repeated. The justifications consume space in memory, and the tradeoff is therefore made between memory storage and the processing overhead of regenerating proofs on the fly.

2.7.6 **Face validity.** Implementations of non-monotonic reasoning revise beliefs so as to arrive at a consistent overall system of beliefs in the face of a contradiction. But they provide only a very limited capability for deciding among alternative possible revisions. The selection of an assumption as the "culprit," and the selection of a member of its _out_list to be assumed as true, are both highly arbitrary. Some control information is implicit in the ordering of nodes in the outlist of statement 5; i.e., if 5 is to be rejected, the system will then assume the truth of members ~~of numbers~~ in the outlist in the order shown. But (a) this is insufficient to remove all ambiguities, and (b) it makes control information implicit rather than explicit, hence, difficult to evaluate or modify.

2.7.7 Plausibility of instances: Conflicting evidence. An often voiced criticism of non-monotonic reasoning is that uncertainty calculi (e.g., Bayesian, Shaferian, or fuzzy) can do the same job better. In the example of Section 2.7.4, for example, our initial state of belief, before consideration of either image, could be represented as a belief function assigning some support to statement 1 and some support to {1,2,3,4}. The evidence represented by (5 OUT, 6 IN) could be construed as lending some support to node 3 and some to {3, 4}. The second bit of evidence (7,8 OUT; 9 IN) could be construed as assigning exclusive support to node 4. Combination by Dempster's rule leaves node 4 as the only viable hypothesis. The belief function analysis appears to be more general, since it accommodates sources of information which conflict to varying degrees, and provides a measure of degree of belief in various possible conclusions.

Although we are convinced of the value of numerical representations of uncertainty, we will argue that this objection misses the mark in two ways. It overlooks an important role of non-monotonic reasoning (1) in drawing implications for the validity of one argument or line of reasoning from another, even where they are independent, and (2) in reasoning about the application of the uncertainty calculus itself.

The basic idea of (1) is the following: Suppose we have two independent lines of reasoning, A and B, with regard to the same sets of hypotheses. Each line of reasoning depends on certain data and certain assumptions, as illustrated in Figure 2-8. In Argument A, the impact of Data 1 and Data 2 depends on the acceptance of Assumption 1; for Argument B, the impact of Data 3 and Data 4 depends on Assumption 2.

What happens when A and B support conflicting hypotheses? In a non-monotonic

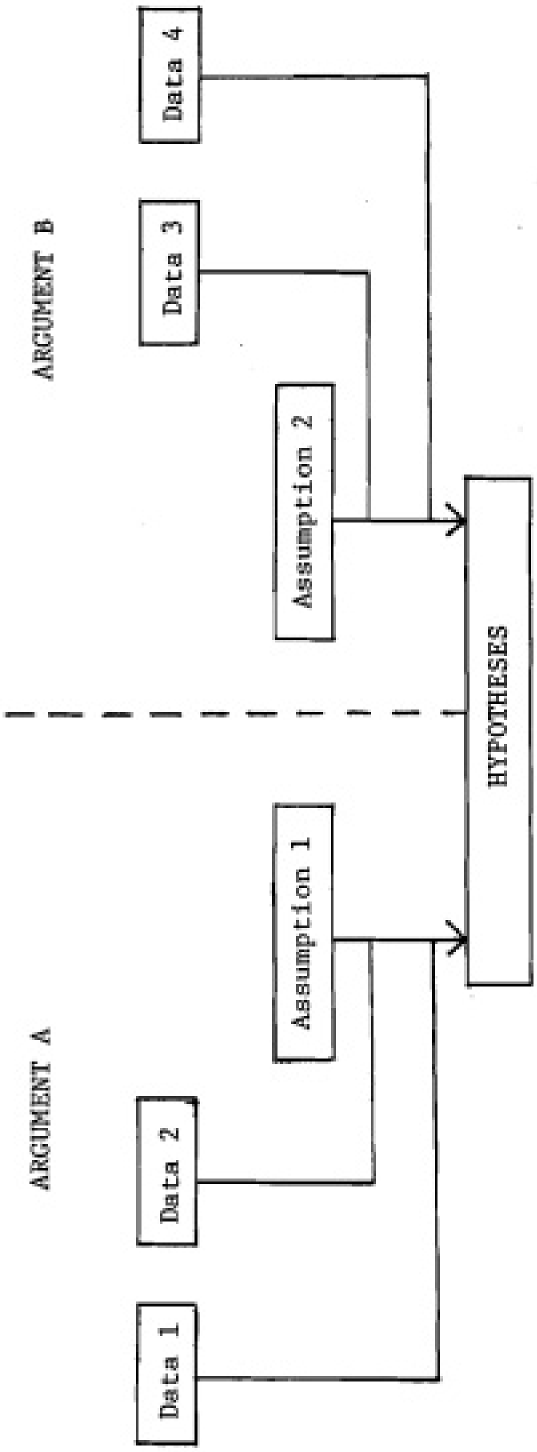

ARGUMENT A
ARGUMENT B
Data 1
Data 2
Assumption 1
Data 3
Data 4
Assumption 2
HYPOTHESES

Figure 2-8

system, the set of assumptions that contributed to the contradiction are identified and declared inconsistent (as a set). Then a selected member of this set is rejected, by producing a justification (itself an assumption) for a member of its *out*list. As a result, at least one of the two arguments fails (or has a different conclusion), and consistency is restored.

The key point here is that conflict between A and B causes the system to *reach inside* each of the arguments. Conflict resolution is a process of reasoning about knowledge: what are the weakest links in each line of reasoning? where would revision accomplish the most?

It will be worthwhile to illustrate this process by a modification of our example. Imagine (somewhat fancifully) that we are less sure about reported observations of large shadows than about small ones, due to possible large-scale non-uniformities in the reflectance of the deck. Then we make the following changes to the initial state of belief:

Statement #	Statement	State		Support List	
		IN	OUT	In	Out
9'	At sun angle θ_2, object is uniformly bright, casts a shadow larger than object	X		11,12	
11	At sun angle θ_2, object is uniformly bright, *appears* to cast a shadow larger than object	X			
12	Surface of deck has uniform reflectance	X			13
13	Surface of deck has non-uniform reflectance		X	No justification	

We see that 9', unlike 9, is not a premise; it is inferred from 11 and 12--i.e., the *appearance* that the shadow is large (11) plus the *assumption*, in effect, that this appearance is not deceiving (12). Statement 12 is a "default assumption:" its acceptance depends only on the absence of evidence to the contrary. At the start of reasoning, 12 is declared IN, since statement 13, that the deck has non-uniform reflectance, has no justification. As a result, all inferences based on the two images proceed exactly as described above.

Now suppose we receive some new, independent evidence. For example, an intelligence report from Agent Y, who is inside the country which owns the ship, says that plans were made to place a device Z on the deck at the precise spot in question--and we know that such a device would appear as a flat horizontal object supported above the deck. This evidence, if reliable, supports statement 3, and is inconsistent with the other hypotheses. We now add nodes corresponding to this evidence, and add a new justification for statement 3 to represent its potential impact:

Statement #	Statement	State		Support List a		Support List b	
		IN	OUT	In	Out	In	Out
3'	Flat horizontal object ~~supported~~ suspended above deck			6,8	1,2,4 5,7,9	14	1,2,4
14	Device Z is present	X		15,16			
15	Device Z is reported present by Agent Y	X					
16	Agent Y is reliable	X			17		
17	Agent Y is not reliable		X	No justification			

We also add 14 to the outlists of statements 1, 2, and 4. A premise, statement 15, describes our new evidence. But, here too, we have explicitly represented an assumption (16) which is required to make the evidence useful. Since the reliability of Agent Y (16) is a default assumption, the system infers that device Z is in fact present as reported (14 IN). (14 IN) leads to (3' IN, 1,2,4 OUT), which is a contradiction of our previous conclusion.

Dependency-directed backtracking will resolve the conflict by revising one of the assumptions that produced it. It may assume that the surface of the deck must, after all, have non-uniform reflectance, (12 OUT, 13 IN), hence, 3' is to be accepted. Or it may assume that Agent Y must be unreliable, (16 OUT, 17 IN), hence, 4 is to be accepted. As noted above, a clear inadequacy of the system described by Doyle (1979) is the lack of some measure of the firmness of an assumption upon which to base this choice. Nonetheless, the important point is that conflict of evidence leads to inferences regarding the acceptability of beliefs (12 and 16) which are internal to each of the conflicting arguments.

Consider, on the other hand, how an uncertainty calculus such as Shafer's would handle this problem. We examined the issue of conflict resolution, in the context of belief function theory, in some detail in Section 2.5.6. There we found that, depending on the degree of conflict, and on the existence and degree of discounting for the two arguments, we could have: (a) an indeterminate result (if there is no non-empty intersection between possible meanings of the two arguments), (b) exclusive support for hypotheses in the intersection of meanings (if there is no discounting), or (c) strong support for each of the two conflicting conclusions). None of these alternatives examines the sources of the conflict and seeks insights regarding its causes. Adjustments of discount rates in the light of conflict are likely, moreover, to be invalid in the absence of some exploration of reasons for the adjustment.

Of course, a belief function analysis can examine the contents of two arguments. To do so, however, it must enormously complicate the frame T (see Section 2.5.5). In other words, the original set of hypotheses {1,2,3,4} must be replaced by a much larger set which also includes the assumptions: {1,2,3,4} x {12,13} x {16,17}. Then evidential support must be assessed, for each of the two conflicting arguments, on the subsets of this expanded set. The price we pay for this strategy, however, is enormous: in quantity of inputs and computational tractability, but also in the naturalness of inputs. It is not likely to be very clear, for example, what bearing our evidence for or against the reliability of Agent Y would have on our beliefs regarding the reflectance of the deck; and similarly, vice versa. The reason, of course, is that the link is highly indirect and is discovered only by means of the conflict in conclusions which the two sets of beliefs engender. The truth maintenance system represents this connection in a quite natural way.

Nonetheless, non-monotonic systems as presently constituted are inadequate in a number of ways. Problems are chiefly attributable to their exactness, on two levels. For example, non-monotonic systems provide a way of reasoning with incomplete information, i.e., by adopting assumptions, tracing their consequences, and revising them if they lead to an inconsistency. But they provide no measure of the degree of incompleteness in the support for a belief, and no concept of degree of conflict. As we have already noted, a measure of this sort seems essential in selecting among alternative possible revisions.

On a second level, the statements whose truth or falsity is adjudicated are themselves exact. However, there is no reason why similar principles of qualitative reasoning might not be applied to probabilistic or imprecise constraints and data. The need for such a "meta-reasoning" capability is the chief conclusion of our comments in earlier discussions of Bayesian and Shaferian calculi. In our view, non-monotonic logic may have its most convincing application at a higher level, in controlling the application of an uncertainty calculus itself. Assumptions of more than one sort--about the quality of uncertainty assessments, about the independence of evidential arguments, and about the validity of steps in an argument--are inescapable in the application of such a calculus. Most of these assumptions are not easily represented in the language of the calculus itself. Hence, non-monotonic reasoning may be the appropriate tool for keeping track of assumptions and revising them when they lead to anomalous results. As such, it may be the key to a truly "intelligent" or flexible application of those models. It is to this possibility that we turn in Section 3.0.

2.7.8 Summary. Non-monotonic logic is a computationally efficient method for reasoning with incomplete information, i.e., for adopting assumptions and revising them in the face of conflicting data. Statements are associated not with numerical indices of uncertainty, as in the other theories we have examined, but with reasons. Certain statements (called assumptions) may be accepted in the absense of positive support, as long as certain other beliefs have not been disproven. Non-monotonic logic provides a natural method for revising beliefs within indepen-

dent lines of reasoning when they lead to conflicting conclusions. Unfortunately, validity is diminished by the arbitrariness of its procedures for selecting among alternative possible belief revisions. We argue that the most useful application of non-monotonic reasoning may be as a control process for the application of an uncertainty calculus.

3.0 THE NON-MONOTONIC PROBABILIST: AN APPLICATION OF BELIEF FUNCTIONS, FUZZY LOGIC, AND NON-MONOTONIC REASONING

3.1 Contrast Between Probabilistic and Qualitative Approaches to Conflict Resolution

The attempt to introduce non-"ad hoc" probabilistic reasoning into expert systems has led to a variety of dilemmas. Probabilistic analysis, as practiced by statisticians, typically requires extensive judgments regarding interdependencies among hypotheses and data, and regarding the appropriateness of various alternative models. The application of such models to real problems is typically an iterative process, in which the plausibility of the results confirms or disconfirms the validity of judgments and assumptions made in building the model. All these features seem to conflict with the modularity of knowledge representations associated with expert systems. In a recent paper, for example, Glenn Shafer (1984a) has concluded pessimistically

> ...that the expert systems we see using probability in the near future are not likely to have the flexibility and judgmental capacity that we associate with genuine intelligence. Instead, these systems will continue to leave the work of genuine intelligence to their designers and users. Their designers will have to design the forms of probability argument for the particular problem, and their users will have to supply the probability judgments.

The present work addresses this problem in the context of conflict resolution. Probabilistic and qualitative approaches to reasoning offer quite different conceptions of what it is for two lines of argument, or two pieces of evidence, to conflict. From the Bayesian point of view, for example, divergence can be regarded as stochastic; it is comparable to the chance occurrence of errors, or "noise," in a process of measurement. Extreme divergence of results is unlikely, but is in fact expected to occur a small percentage of the time. From the qualitative point of view, however, divergence is a result of faulty knowledge; that is, conflicting results are taken as evidence that one or more assumptions or forms of argument that led to the conflict are mistaken.

These two conceptions of conflict lead to quite different rationales for the process of combining evidence or lines of reasoning. From the Bayesian point of view, the process is akin to that in which independent errors in repeated measurements tend to cancel one another out. From the qualitative point of view, the object is to improve the overall truth of a system of beliefs--to explicitly identify potentially erroneous steps in the argument and to change them.

- This contrast with qualitative approaches does not apply merely to Bayesian theory. In Shafer's probabilistic conception, for example, the divergence of two arguments is simply attributed to the fact that they are based on different, independent bodies of evidence. The object ~~direct~~ of combining evidence is, in essence, to tally support for the alternatives conclusions, not a true "reconciliation".

Shortcomings in both probabilistic and qualitative points of view are, in part, complementary. An objection to both Bayesian and Shaferian systems of probability, for example, is that they take no formal account of the iterative process--of tentatively adopting a model and a set of assessments, testing its implications, and revising--which is essential to the efficient and valid application of such theories. Moreover, they provide no coherent criterion for the provisional "acceptance" of a conclusion as true. Use of conflict as a stimulus for the restructuring of probability models or revision of probabilistic inputs may lead to such a criterion. On the other hand, qualitative systems of reasoning, such as Doyle and McDermott's non-monotonic logic, do not accommodate degrees of belief or degrees of conflict, and suffer from an arbitrariness in the process of selecting beliefs for revision in the face of a conflict. Numerical indices of uncertainty may be of use both for communication with users and for purposes of control in reasoning.

3.2 Functional Outline of a Proposed System: The Non-Monotonic Probabilist

These considerations suggest the design of a system that regards conflict as jointly knowledge-based and stochastic. It would reduce conflict by a process of non-monotonic reasoning prior to statistical aggregation by probabilistic rules; i.e., non-monotonic processes would operate on and modify the assumptions and judgments embodied in a rule-based belief function model. At the same time, however, the non-monotonic processes would be guided by measures of completeness of support provided by the belief function calculus. Each model--non-monotonic and probabilistic--thus in a sense embeds the other.

The justification for such a system, and the motivation behind its basic functions, have been argued in Section 2.0. Our purpose in this subsection is to pull these threads together in a high-level conceptual outline of a Non-Monontonic Probabilist (NMP) System. Further details are given in Section 3.3, which discusses the role of the system in human-computer interaction, and in Section 3.4, which discusses fuzzy measures required to implement the system's functions. Appendix A shows how certain features of this system could be applied to illustrative problems of image understanding.

3.2.1 Rule-based belief function module. The core of the probabilistic model is a set of production rules. The action components of the rules assign Shaferian support measures to subsets of hypotheses. For example,

R.1 If a region has texture of type x,

then

	m(·)
S.1: Region is a field	.98
S.2: Region is a forest	.01
S.3: Region is a building	0
S.4: {S.1,S.2,S.3}	.01

R.2 If an intelligence agent reports
presence of a building in a region,

then

	m(·)
S.1: Region is a field	0
S.2: Region is a forest	.01
S.3: Region is a building	.98
S.4: {S.1,S.2,S.3}	.01

Current knowledge about the problem domain is maintained in a database, which includes statements about subsets of hypotheses, such as S.1-S.4 above, together with their current degrees of belief. When the antecedent of a rule appears in the database, the rule is triggered, and the support it assigns is combined by Dempster's rule with the existing support for the relevant subsets of hypotheses. Support is attenuated if the antecedent of a rule is only partially established.

In this model, inference may be either forward-chaining or backward-chaining; an image understanding system could involve either or both. Note, however, that a simple forward-chaining model could capture many critical features of both "bottom-up" and "top-down" reasoning. In bottom-up processing, degrees of belief for labels of a region are assigned when image data from that region trigger a rule, such as R.1. above. Shaferian template matching, described in Section A.3.5., falls under this heading. In top-down processing, on the other hand, rules regarding the assignment of labels to a region may be triggered by extraneous knowledge, as in R.2. Section A.2.6. describes a different use of extraneous knowledge involving relations among regions. In that example, the classification of certain regions as roads reduces the support for classifying any distant region as a building.

These examples strongly suggest an iterative, forward-chaining processing strategy for image understanding. First, belief functions are computed for all regions based on (bottom-up) image data and non-relational extraneous knowledge. Then the belief functions established in this way are used to trigger a second set of rules involving relational extraneous knowledge.

Where forward-chaining inference proves inadequate is in the use of the rule-base, together with partial results, to prioritize the need for new information. This will be an essential aspect of the non-monotonic processes to be described. We believe, therefore, that an effective image-understanding system will utilize backward, as well as forward-chaining inference.

The use of belief functions (rather than, say, Bayesian probabilities) provides the advantages discussed in Section 2.5 above. There is a natural representation of incompleteness of evidence as the support assigned to the universal set (S.4 in the above example); this will play a critical role in the control of non-monotonic reasoning. And support need not be assigned arbitrarily when appropriate evidence is missing. In image analyses, as in medical diagnosis (Gordon and Shortliffe, 1984), we might expect a hierarchical structure of support for hypotheses: e.g., one bit of evidence establishes that a region is a building; a second bit establishes the kind of building it is; etc. Belief functions are a highly natural tool for capturing such a structure. As a final note, we remark that specialized belief function models of this sort may be required to ensure computational feasibility (Section 2.5.3 above).

3.2.2. Non-monotonic reasoning as an embedding context. In the NMP system, both rules and statements are assumptions, whose acceptance or use depends on the failure to disprove certain other beliefs (cf., Section 2.7 above). Those other beliefs are the reasons for the rule or the statement. Such beliefs include:

(1) Model characteristics (e.g., linearity, normality, consonance, etc.) used in generating the support measures associated with a rule,

(2) the representativeness of frequency samples or expert experiences used in generating such support measures,

(3) the independence or non-independence of different items of evidence, and

(4) the occurrence or non-occurrence of facts or events which could affect belief in a statement by triggering some rule, but for which there is (as yet) no direct evidence.

(5) premises in a "proof" of the statement

(For discussion of these factors in the belief function context, see Section 3.2.5.10 above.) Beliefs of types (1), (2), and (3) are among the suppositions required for application of a rule. Beliefs of type (4) are presupposed by the current assignment of degrees of belief to declarative statements. In addition, of course, belief in a statement depends on the validity of the rules applied in deriving it, hence, indirectly, on suppositions of types (1), (2), and (3).

Measures of credibility for both rules and statements are mathematically derived from the degree of their dependence on suppositions of this type. For example, the "discount rate" for a rule's support function (in R.1 above, this is the support for the universal set, m({S.1, S.2, S.3}) = m(S.4) = .01) will depend on the nature of the suppositions in categories (1), (2), and (3). This reflects the possibility that the evidence summarized in the rule is in fact irrelevant; e.g., because the set of photos used as a training set was from a different geographical or cultural area.

The credibility of a statement, in turn, will be a joint function of its discount rate (computed by Dempster's rule from the support functions applied in deriving it) and the suppositions of type (4). Thus, if R.1 and R.2 are both triggered with regard to a particular region, the resulting support function by Dempster's rule is:

	$m_{R.1,\ R.2}(\cdot)$
S.1 Region is a field	.49
S.2 Region is a forest	.015
S.3 Region is a building	.49
S.4 {S.2, S.2, S.3}	.005

The discount rate, m(S.4), is reduced to .005. However, the credibility of the support assignments to S.1, S.2, and S.3 also depends on the existence or non-existence of other rules in the rule base (e.g., the rules concerning distance from roads) which, <u>if</u> they were to be triggered, would significantly change the support measures.

A state of conflict exists when a significant degree of belief is assigned by statements in the data base both to a subset of hypotheses and to its complement. Conflict triggers a process of dependency-directed backtracking, in which one or more of the suppositions listed above may be revised: e.g., the structure of a model may be altered; the presumed relevance of frequency data or probabilistic expert assessments to the current problem may be adjusted; the problem may be reframed so as to merge dependent arguments; or the occurrence of relevant facts or events upon which beliefs depend may be hypothesized. Adaptive learning in such a system could, therefore, involve revision of belief not only about the oc-

[illegible] subject to shared premises (or only disagree as necessary for multi-level premise [illegible]) currence of external facts or events, but about the validity of inferential procedures in its own rule base.

In our example, $m_{R.1,\ R.2}(\cdot)$ appears to present a conflict; thus, the system will explore potential revisions in R.1 and in R.2. In doing so, it will try to reject suppositions upon which R.1 and R.2 depend. For example, (a) it may question the relevance of the training set used to derive R.1; (b) it may question the competence or trustworthiness of the agent in R.2; (c) it may try "reframing" the problem, e.g., the region may be partitioned into smaller regions or merged with other neighboring regions. (The latter might occur by adjustment of parameters in a low-level segmentation procedure.) Finally, (d) the system might look for evidence supporting (as yet unconfirmed) events or facts that would significantly change the assigned support function (e.g., discovery that the region is distant from a road would reduce support for S.3).

3.2.3 Belief functions as a controlling context for non-monotonic reasoning. How will the system choose among these alternative tactics for conflict resolution? More fundamentally, since conflict within a belief function is not typically an all-or-nothing matter (like logical contradiction), how will the system determine when conflict exists? In the Non-Monotonic Probabilist, the control of dependency-directed backtracking is determined (a) by a domain-specific definition of conflict for belief functions, and (b) by the relative standing, in terms of credibility, of statements, rules, and the beliefs upon which they depend. The actual mechanisms are implemented using a set of fuzzy measures described below in Section 3.4.

Conflict is domain-specific (or even problem-specific) in several senses: (1) The type of conflict which the system is designed to address can be specified explicitly, and easily modified. For example, conflict may be regarded as significant support for a hypothesis and its complement (as above); but it might also include, for example, the assignment of strong support to a single hypothesis based on two support functions neither of which assigns significant support to that hypothesis. (This case is illustrated in Section 2.5.11) (2) Conflict is a matter of degree; and the "significance" of any given degree of conflict is represented by a single parameter which is easily modified. (3) Conflict resolution is not simply "triggered" when the significance of conflict exceeds some threshold. Conflict resolution is subject to a graded control process, in which the significance or seriousness of the conflict is continually compared with the credibility of the beliefs contributing to the conflict. Conflict resolution stops when the seriousness of the conflict drops below the degree of "revisability" of the relevant suppositions. In effect, then, any diagnosis of "significant conflict" can be overruled by strong independent plausibility of the contributing beliefs. The result is a system of beliefs which, in an intuitive sense, maximizes _global_ plausibility.

The selection of beliefs for revision in the face of conflict is a non-random process. It is guided by measures which capture the extent to which critical evidence for a particular belief is at present incomplete or unreliable. Independent confirmation for hypothesized revisions is then sought either from image data, the store of extraneous knowledge, or the user.

When a conflict occurs, the system locates chains of reasoning that (a) contributed strongly to the conflict and (b) have weak, or relatively unsupported, starting points. In our example, these are a variety of candidates. R.1 is a strong contributer to the conflict, since its discount rate is quite low. The system would search among the reasons for R.1-- e.g., a list of purported similarities and dissimilarities between the current image and the training set -- for those which have the least evidential basis. For example, in constructing the support function of R.1, we may have supposed (without really knowing for sure) that weapons facility construction procedures in the target region resemble those in our country. If this belief were to be revised, the newly posited dissimilarity would inflate the discount rate for R.1's support function, and the conflict with R.2 would be decreased. Alternative chains of reasoning involving R.1 and R.2 lead to other possible revisions, e.g., in the reliability of the agent referred to by R.2, or in the segmentation of the relevant region. The choice of a revision would depend on a measure that reflects the potential benefit in terms of conflict reduction, and the potential cost, in terms of evidential restraints on possible revisions. Whatever revision is chosen, additional information regarding the revision may then be sought: by more extended or more sensitive processing of the image, by a more inclusive search for relevant extraneous knowledge, or by directly querying the user of the system.

A different sort of example involves the chain of reasoning that goes from the statement S.3 (that the region is a building) to *its* reasons. The validity of the support function assigned to S.3 ($m_{R.1,R.2}(\cdot)$) presupposes that *other* potentially relevant rules have *not* been triggered. In particular, if the relevant region were found to be distant from all roads, support for S.3 would decline; yet it may be that no data has as yet been obtained regarding the presence *or* absence of roads in neighboring regions. One avenue for belief revision, then, is to posit the absence of roads in the vicinity. Through a backwards chaining inference, this posit could direct further processing of the image in the relevant regions, in a search for evidence of roads.

As in "standard" non-monotonic reasoning, revisions in belief are retained by the system until new conflicts involving those beliefs are discovered. At that point, the revision will be undone--unless additional information has in the meantime provided an independent basis for its retention.

3.3 The Non-Monotonic Probabilist as an Interactive System

In many applications, an image-understanding system will be required to function interactively with a human user. The appropriate allocation of effort between the analyst and the computer can, however, vary drastically as a function of such variables as time pressure, workload, the importance of the task, and the need for "judgment" not incorporated in the automated system.

Under conditions of low time stress and with relatively high-level, unstructured tasks, the appropriate allocation mode might involve predominant human control of the problem-solving process. The computer's role (as explored in Cohen et al., 1982) might be to monitor the user's behavior and to prompt when the user's actions are likely (in the computer's opinion) to be significantly suboptimal. The user would determine the degree of suboptimality that justifies a prompt.

By contrast, under high time stress and workload or in relatively "mechanical", structured tasks, the appropriate allocation mode might involve a predominant role for the computer. In this case (explored in Chinnis, Cohen, and Bresnick, 1984) the computer might monitor its own problem-solving activity and prompt the human when conditions appear that suggest value in a potential human contribution.

An important feature of the Non-Monotonic Probabilistic system is that it can provide, if desired, a framework for collaborative problem solving between the user and the system in either of these two modes.

The system described in Section 3.2 already contains an implicit "executive" function for human-computer task allocation under conditions of high workload. Control may be shared between user and computer in the following ways: (a) Users may specify their own definition of the type and degree of conflict among items of evidence that will trigger belief revision. (b) Based on this user-defined objective, and on an assessment of limitations and conflict in its own knowledge, the system will direct user attention to areas where his contribution can be most valuable. Beliefs which are subject to revision are labeled according to whether or not users are a potential source of information. When an appropriately labeled belief is selected for possible revision by dependency-directed backtracking, the user will, if he desires, be queried. (c) Users may then adjust support assessments and add and delete support list elements, to reflect their on-the-spot knowledge.

The advantages of this framework in a high workload and highly uncertain task environment are considerable: (i) Users will <u>not</u> be bothered by the need to provide inputs when default assumptions are adequate; (ii) when anomalies do occur, the system does take advantage of potential user contributions; (iii) the system reduces user workload by generating promising options (i.e., potential revisions which would restore consistency) for consideration by the user; (iv) imprecise linguistic inputs could be accepted; and (v) ultimate control over the objectives of the reasoning process, its outcome, and his own degree of participation is left in the hands of the user.

For high-level tasks, where the human has a predominant role, some fairly straightforward elaborations of the basic conflict resolution mechanism are required. The computer could develop hypotheses regarding the user's beliefs and assumptions and their degree of suboptimality by observing the user's performance (e.g., manual labeling of image regions) and working the problem itself in parallel. Discrepancies between user and computer solutions would be treated as conflicts, triggering a process of (hypothetical) belief revision. The computer would identify the least disruptive changes in its own beliefs required to make them consistent with the human's conclusions. The resulting set of beliefs is attributed, heuristically, to the human. If these beliefs exceed a certain criterion of implausibility (according to the computer), the user would be prompted. Moreover, the system would display the assumptions which it has inferred to be involved in the user's solution, and the reasons for their implausibility according to the computer model. The user may then weigh the computer's arguments against his own. The user himself will control the frequency with which he receives such advice, by determining the criterion of implausibility required to trigger a prompt.

3.4 Fuzzy Measures

Fuzzy variables have a variety of potential roles in this system:

- in the description of facts or events (e.g., "rough" or "smooth" textures);
- in the assessment of numerical measures of support (e.g., "about .30"); and
- in the system's internal processes of reasoning.

In this section, we focus on the third of these roles, briefly outlining a set of (tentative) measures corresponding to the concepts described in Section 3.2.

In a certain sense (aŝ discussed in Section 2.6 above), these measures are ad hoc.

However, they provide an extremely flexible tool for duplicating, in a continuous rather than discrete fashion, some of the concepts used in "standard" non-monotonic reasoning. They enable us to avoid an elaborate calculus, like second-order probabilities, which would seem gratuitous, and indeed equally ad hoc, for this purpose. They provice a graded process of high-level control through a reasonably plausible and simple set of definitions.

3.4.1 Conflict. A simple measure of degree of conflict in a belief function is the following. Let A be a subset of hypotheses and $\overline{A}$ its complement. If $Q = \{A,\overline{A}\}$, then

net conflict [illegible] function like - as a [illegible] and want a [illegible] → its compl.

(1) $$\mu_{\text{conflict}}(Q) = 2 \min[\text{Bel}(A),\text{Bel}(\overline{A})].$$

This can be justified in two ways. From the fuzzy logic point-of-view, we might regard it as the membership function for the intersection of belief in A and belief in $\overline{A}$, i.e., a contradiction. Multiplication by two normalizes the measure, so that maximum $\mu_{\text{conflict}}(Q)=1$ is achieved when $\text{Bel}(A) = \text{Bel}(\overline{A}) = .5$. Secondly, note that is it equivalent to the following expression:

x x

$$\left\{ 1 - \left\{ \frac{|\text{Bel}(A)-\text{Bel}(\overline{A})|}{\text{Bel}(A)+\text{Bel}(\overline{A})} \right] \right. (\text{Bel}(A)+\text{Bel}(\overline{A})) = 2\text{Bel}(\overline{A})$$

when we assume, without loss of generality, that $\text{Bel}(A)\geq\text{Bel}(\overline{A})$. This expression intuitively captures the notion of conflict in a belief function: the first bracketed expression is the relative similarity of the degrees of belief in A and $\overline{A}$; the larger this is, the greater the conflict. The second bracketed expression is the total committed belief; to the extent that the belief function is "discounted" by assigning support to the universal set $\{A,\overline{A}\}$, we regard the conflict as reduced. (x In short, the maximum, Bel(A), doesn't matter since increasing it (with $\text{Bel}(\overline{A})$ constant) has two opposing effects: it increases the difference between Bel(A) and $\text{Bel}(\overline{A})$, but also increases the total committed belief.

Conflict resolution is prompted, however, by "significant" conflict, and the degree of significance required may be a variable function of the problem domain. A simple, though somewhat ad hoc, way to accomplish this is to define

$$\mu_{\text{signif. conflict}}(Q) = \mu_{\text{conflict}}{}^{\gamma}(Q)$$

where γ is a power to which $\mu_{\text{conflict}}(Q)$ is raised. Increasing γ has the effect of requiring higher degrees of conflict to achieve "significance". comparable degrees of membership

[illegible] 3.4.2 Support lists. Each rule and each statement is associated with a set of potential reasons, in the form of a support list. However, in place of a discrete classification (inlist vs. outlist) we substitute a "fuzzy membership function," i.e., a continuum from in to out. Moreover, strictly speaking, it is the current support assignment to a statement, rather than the statement itself, which has reasons or which serves as a reason. We will de~~v~~note the support assignment to statement A by underlining, $\underline{A}$. relay on context X X

Location of a statement $\underline{S}$ on the support list continuum for a second statement A or a rule R depends on only two things: (a) the presence of S on the list of possible reasons for A or R, and (b) the amount of support for the universal set $\{S,\overline{S}\}$. In particular, where S is a possible reason for A, ✱

$$\text{(2a)}\qquad \begin{aligned} \mu_{\underline{\text{out}}\text{-}\underline{A}}(\underline{S}) &= m(S,\overline{S}) \\ \mu_{\underline{\text{in}}\text{-}\underline{A}}(\underline{S}) &= 1\text{-}m(S,\overline{S}) = \text{Bel}(S)+\text{Bel}(\overline{S}) \end{aligned}$$

where in and out hereafter refer to the inlist and outlist membership functions respectively (not to the statement S's being accepted or believed as IN or OUT). Correspondingly, when a rule R is a possible reason for $\underline{A}$,

$$\text{(2b)}\qquad \begin{aligned} \mu_{\underline{\text{out}}\text{-}\underline{A}}(R) &= m_R(A,\overline{A}) \\ \mu_{\underline{\text{in}}\text{-}\underline{A}}(R) &= 1\text{-}m_R(A,\overline{A}) \end{aligned}$$

where $m_R(\cdot)$ is the support function assigned by R.

(*)

relevance of that reason in this particular application, i.e.,

These measures capture a very simple intuition. They place the potential reasons for A (or R) in an order corresponding to the reliability or completeness of evidence presently underlying ~~each~~ the reason. To the extent that confidence in A or use of R depends upon reasons with high $\mu_{\underline{out}}$, they rely on unproven (but not disproven) suppositions. (We argue that this is inevitable in any probabilistic analysis.)

What determines the content of the list of possible reasons? For a statement A, it contains (a) the rules in the system which have a support assignment for A in the consequent, and (b) the statements which occur in the antecedents of those rules. The possible reasons for a rule are less well-defined. They may include a list of potential similarities (or absences of potential dissimilarities) between the target application of the system and the exemplars upon which it was trained. They may also include specifications of model assumptions used to generate support assignments. Finally, they include (in practice) assertions of independence of the evidence summarized by the rule from evidence utilized in all other rules of the system.

Equation (2) may be elaborated in two respects. First, it might be desirable (though a bit *ad hoc*) to fuzzify the membership of a statement S in the list of possible reasons, i.e., S may only "resemble" some member of that list S*. In that case,

$$\mu_{\underline{out}\text{-}\underline{A}}(\underline{S}) = \min[\sup(S \cap S^*), m(S,\overline{S})] \tag{2a'}$$

(?) —

$$\mu_{\underline{in}\text{-}\underline{A}}(\underline{S}) = \min[\sup(S \cap S^*), 1-m(S,\overline{S})]$$

* where $\sup(S \cap S^*) = \sup_u(\mu_S(u) \wedge \mu_{S*}(u))$, with $\wedge$ referring to min. The latter is a measure of the intersection of two fuzzy sets S and S*; the outer min in (2') reflects the conjunctive requirement for $\mu_{\underline{out}\text{-}\underline{A}}(\cdot)$.

A second elaboration of (2) is perhaps more substantive. It involves the observations (a) that a statement S can have no impact, as a reason, on another statement A unless there is a rule linking them (with S in the antecedent and a support assignment for A in the consequent), and (b) that a rule R can have no impact on A without the (at least partial) satisfaction of its antecedent by a statement. Thus, we must take members of the support list for a statement A to be pairs of statements and rules $(\underline{S}_i, R_i)$, rather than statements and rules separately. Ignoring the complications of (2'), we get:

$$\begin{aligned}\mu_{\underline{out}\text{-}\underline{A}}(\underline{S},R) &= \min[\mu_{\underline{out}\text{-}\underline{A}}(\underline{S}),\mu_{\underline{out}\text{-}\underline{A}}(R)] \\ &= \min[m(S,\overline{S}),m_R(A,\overline{A})] \\ \mu_{\underline{in}\text{-}\underline{A}}(\underline{S},R) &= 1-\mu_{out\text{-}A}(\underline{S},R).\end{aligned} \tag{2"}$$

*

3.4.3 Assumptions. A statement or a rule is an assumption to the degree that its acceptance or use depends on what is possible, rather than on what is supported by evidence. The following is a simple measure of that concept:

$$\sum_{(S,R)} \mu_{\underline{out}\text{-}\underline{A}}(\underline{S},R)$$

$$\mu_{assumption}(\underline{A}) = \frac{\sum_{(S,R)} \mu_{\underline{out}\text{-}\underline{A}}(\underline{S},R)}{\underset{\text{all possible support}}{n}} \tag{3}$$

* *

~~where n is the~~ total number of statement-~~rule pairs~~ in the support list for A. lower this → ~~$\mu_{assumption}(\underline{A})$~~ is simply the (fuzzy) proportion of $\underline{A}$'s reasons which are out, i.e., unsupported by evidence. *

$\mu_{assumption}(A)$ is – —

3.4.4. Foundations. One requirement of dependency-directed backtracking is the ability to find statements or rules which have an impact, as reasons, on a given statement or rule. A statement-rule pair $(\underline{S},R)$ in fact has an impact on the support assignment to a statement A to the extent that S or its complement is believed (thus, triggering the corresponding rule) and to the extent that R assigns a non-discounted support function. Other pairs of statements and rules, however, may have an indirect effect on $\underline{A}$ by having an impact on $\underline{S}$ or R. All these pairs are, to a degree, part of the "foundations" of $\underline{A}$. We measure this as follows:

$$\mu_{foundations\text{-}\underline{A}}(\underline{S}_n,R_n) = \min_{1 \le i < n}[(\mu_{\underline{in}\text{-}\underline{S}_{i-1}}(\underline{S}_i,R))] \tag{4}$$

where $\underline{S}_o = \underline{A}$. In effect, the min function says that the chain of impact linking (S_n,R_n) to A <u>via</u> $(S_{n-1},R_{n-1})\ldots(S_1,R_1)$ is only as strong as its weakest link.

To what extent is a statement $\underline{S}$ by itself (or a rule R by itself) part of the foundations of $\underline{A}$? Here, we get:

$$(5) \qquad \mu_{\text{foundations-}\underline{A}}(\underline{S}_n) = \sup_R[\mu_{\text{foundations-}\underline{A}}(\underline{S}_n,R)],$$

i.e., $\underline{S}_n$'s impact is equal to the impact of the most effective chain to which it belongs. Similarly,

$$\mu_{\text{foundations-}\underline{A}}(R) = \sup_{\underline{S}}[\mu_{\text{foundations-}\underline{A}}(\underline{S},R)].$$

3.4.5 <u>Suppositions</u>. Suppositions are <u>assumptions</u> with an <u>impact</u>. More precisely, the statements and rules which $\underline{A}$ requires us to "suppose" are (a) in the foundations of $\underline{A}$, and (b) assumptions in their own right. The degree to which a statement $\underline{S}$ (or a rule R) is a supposition of $\underline{A}$ is given by the following:

*

$$(6) \qquad \mu_{\text{supposition-}\underline{A}}(\underline{S}) = \min[\mu_{\text{foundations-}\underline{A}}(\underline{S}),\mu_{\text{assumption}}(\underline{S})].$$

3.4.6 <u>Dependency-directed</u> <u>backtracking</u>. There are a variety of ways that these measures, or other similar ones, might be used to direct backtracking and belief revision. Here we give one, quite tentative, approach. Suppose that $Q = \{A,\overline{A}\}$ has a high degree of conflict. The strategy is simply to select the maximal supposition for A as the "culprit" C, and then to "negate" C by revising the maximal member of C's <u>out</u>list. More precisely, we select a rule or statement C such that

* ?

$$\max_{C'}[\mu_{\text{supposition-}\underline{A}}(C')] = \mu_{\text{supposition-}\underline{A}}(C).$$

Then we select a statement-rule pair ($\underline{S}$,R) for revision such that

$$\max_{\underline{S}', \underset{\text{[illegible]}}{\uparrow} R'} [\mu_{\underline{out}\text{-}C}(\underline{S}',R')] = \mu_{\underline{out}\text{-}C}(\underline{S},R). \qquad \ast$$

Finally, $\underline{S}$ or R may be revised, depending on which has the least evidential support, i.e., $\max[m(S,\overline{S}),m_R(C,\overline{C})]$.

3.4.7 Conflict as the control over revision. No revisions in fact take place unless the degree of conflict is serious enough to justify them. This involves a simple comparison between the measure of significance of the conflict and a measure of the "resistance" to revision for our best available candidate. Thus, if

$$\mu_{\text{signif. conflict}}(Q) \geq \mu_{\underline{in}\text{-}C}(\underline{S},R),$$

$\underline{S}$ or R may be revised; otherwise, not.

3.5 Conclusion

How does NMP relate in general to currently existing AI software tools? Tools for building expert systems now exist which provide for quantitative reasoning about uncertainty (e.g., EMYCIN). Other systems permit qualitative reasoning about and revision of assumptions (e.g., DUCK). NMP is a superset of these capabilities. Our description of it has dwelled on its capability of combining aspects of both: i.e., qualitative reasoning about a quantitative model, and quantitative measures to guide that reasoning. But note that each extreme can be achieved in NMP itself as a special case. If no assumptions are associated with rules or statements, we get a pure system for probabilistic inference (like EMYCIN or PROSPECTOR, with a Shaferian belief function calculus). On the other hand, if all belief functions were to allocate full support between some single hypothesis and the universal set, we get a pure non-monotonic system (like DUCK).

The problems with these extremes, as we pointed out in Section 3.1, are complementary. Pure probabilistic systems never learn anything new about their probabilistic beliefs and assumptions from the experience of applying them. Pure non-monotonic systems do learn, but they have an arbitrariness and an all-or-none quality about the new beliefs they acquire. Our argument, quite simply, is that both capabilities are needed, and that satisfactory systems will, in general, require their combination.

4.0 SUMMARY AND PROSPECTS

4.1 The Requirement for a Non-Monotonic Probabilist

The development of efficient and accurate devices for automated feature extraction from photographic images has been hampered by a variety of methodological obstacles. Utilization of general knowledge--about physics, geometry, geography, and culture--is critical in the face of noisy, ambiguous, and incomplete data. But the relevant expert system technologies are often difficult to integrate with bottom-up procedures that utilize very different modes of representation and reasoning. More significantly, both expert system and image processing technologies have depended on ad hoc devices for inference and for handling uncertainty, with consequences that are in many cases seriously suboptimal.

In imagery, and in virtually all problem domains where expert system technology might be introduced, there is a need for explicit and valid quantitative modeling of uncertainty; at the same time, there is a need for a metastructure of qualitative reasoning in which the assumptions utilized in the probability model are reassessed and revised in the course of the argument. These are the dual requirements addressed by the Non-Monotonic Probabilist (NMP) described in Section 3.0 above.

NMP will be a general-purpose AI tool, like PROLOG, LOGLISP, OPS5, DUCK, or EMYCIN. Currently existing AI system-building tools either neglect uncertainty altogether (PROLOG, LOGLISP, OPS5), utilize assumptions but provide no explicit probabilistic measures (DUCK), or incorporate ad hoc calculi with no provision for qualitative reasoning about their application (EMYCIN and related systems). NMP will be designed to fill this void. It will serve as an expert system building tool, which accommodates uncertainty both at the level of probabilistic reasoning and at the level of qualitative testing and revising of assumptions.

At the same time, NMP's design can be tailored so that it is optimal for image understanding applications. NMP could be capable of embedding within powerful image processing configurations, to produce systems that perform specialized image understanding tasks.

4.2 Main Results

Sections 2.0 and 3.0 have established the requirement for a system such as NMP and developed its technical foundations. Here we will simply summarize the main arguments and describe the basic technical concepts that enter into the NMP high-level design.

The NMP system (described in section 3.0) blends technology from Shaferian belief functions, non-monotonic reasoning, and fuzzy logic, as well as more traditional features of expert system technology. Shaferian belief functions (Section 2.5) have been chosen as the basic measure of uncertainty, rather than Bayesian probabilities, for several reasons: they do not require definiteness of inputs beyond what the evidence suggests; they provide an explicit representation of the quality of an inferential argument; and they permit "modular" probabilistic analyses based on only subsets of the evidence. Shafer's system permits a variety of useful specialized models for representing evidence. One of these special cases is (very nearly) Bayesian probability theory itself; Shaferian belief functions can represent chance as Bayesian probabilities do, but permit a simple assessment of the _quality_ or reliability of those probabilities as well.

Unfortunately, Bayesian theory is not exactly captured within Shafer's system; the latter does not permit recalibration of the reliability of an information source in the light of what that source says, or in the light of conflict or corroboration by another source. (Bayesian theory does this only at the cost of enormous complexity.) To correct this flaw, we argued that belief functions--as an inference mechanism within expert systems--should be supplemented by a process of qualitative reasoning. That process would keep track of assumptions involved in a belief function model (e.g., concerning the reliability of an information source) and revise them when they lead to anomalies (e.g., conflict with other highly regarded information sources).

The same conclusion was arrived at by consideration of two other features of Shafer's system: the requirement that different bodies of evidence be independent in order to be combined by Shaferian rules, and the lack of any simple mechanism for assessing steps of reasoning _within_ an independent inferential argument. Once again, the solution we propose is a process of qualitative reasoning that tracks assumptions about the independence of two arguments or the internal structure of a reasoning process, and revises them when they contribute to anomalous results.

In concrete applications, such as image processing, these are by no means idle concerns. With noisy and incomplete data, no single form of analysis is free of error; and each relies on different aspects of the data and/or makes different analytical assumptions. Conflicting results, therefore, may be obtained from the application of multiple operators to a pixel array, or from combining extraneous information and expectations with the outcome of a bottom-up analysis. In these cases, the appropriate course of action is to reexamine the factors underlying our evaluation of reliability for the conflicting sources. In addition, their assumed independence might be questioned, for example, by revising the segmentation of the image. Alternatively, new analyses might be initiated to confirm the presence of patterns for which there is as yet no support, but which could account for the anomaly.

We argue that no application of a probabilistic framework is complete in itself. Whether Bayesian or Shaferian, assumptions of various types are always lurking in the background. Conflict among diverse analyses is what forces them into the open. To the extent that assumptions are explicitly tracked and reevaluated, conflict is a prompt for increasing the validity of our beliefs, rather than an occasion for ignoring part of the data or meaningless statistical compromise.

The Non-Monotonic Probabilist implements these requirements by providing a superstructure of non-monotonic reasoning around the application of a belief function model. Non-monotonic logic (Section 2.7) is a method of reasoning with incomplete information, in which assumptions may be adopted and subsequently revised when they lead to contradictory results. The traditional approach, however, has been exact both in the statements to which it applies and in its own control mechanisms. As a result, it fails to capture the important intuitive notion that support for hypotheses may be graded; and the selection among alternative equally consistent belief revisions is highly arbitrary. The NMP system advances beyond this, by applying non-monotonic logic to the application of an uncertainty calculus, and by utilizing measures derived from that calculus to direct the process of belief revision itself.

In the specification of measures suitable for the control of non-monotonic reasoning in NMP, fuzzy logic has been a valuable tool. It provides a precise calculus for vague or imprecise concepts (Section 2.6). It thus makes possible the redefinition, in continuous form, of concepts which occur discretely in traditional non-monotonic systems. In NMP, for example, "conflict" is a matter of degree, and so is the status of a statement or rule as an "assumption". As a result, NMP incorporates a graded control process for belief revision, in which assumptions are subject to retraction only so long as their resistence to revision is outweighed by the strength of the conflict.

An important additional feature of NMP is that it can provide a framework for collaborative problem solving between a user and the system. In a high volume image interpretation task, users will be free for other tasks as long as automatic processing based on default assumptions is adequate. But when anomalies appear, the user's potential contribution may be solicited. The user himself will control the degree of conflict that triggers a system prompt.

4.3 Next Steps

As noted above, NMP can be implemented as a general-purpose tool for constructing expert systems, and in addition, may be embedded it within an image-processing environment. That environment might contain a currently existing system that performs pixel-level operations such as filtering and smoothing, and which provides a preliminary segmentation and labeling of the image. NMP would serve as a higher-level tool for combining bottom-up results with general knowledge and intelligence information, and for resolving conflict. It would influence the operations of the lower-level processor by directing the resegmentation of the image, the recalibration of knowledge sources, and/or the implementation of a more sensitive search for specified patterns. And it would solicit the inputs of a human analyst when the degree and nature of the conflict, as specified by the user himself, call for it.

A variety of technical issues need to be addressed in the course of implementing NMP:

- Refinement and verification of fuzzy measures and algorithms for control of non-monotonic reasoning.

- Final design of basic system architecture: e.g., the mix of forward-chaining and backward chaining inference, control over sequences of iterative processing, and possible use of a blackboard to represent multiple levels of analysis.

- Specification of rules for combining dependent items of evidence within an independent inferential argument, based on Bayesian and/or fuzzy logic principles.

- Development of input routines permitting fuzzy specification of linguistic and numerical facts (e.g., "rough texture," "about 30% probability"). These may include fuzzy descriptions of interdependencies among items of evidence and hypotheses (e.g., "A strongly corroborates B"), and of degrees of permissable conflict among lines of reasoning.

- Design of outputs, consisting of displays of labels for image regions, together with uncertainty measures and explanations where appropriate.

Successful accomplishment of these goals would will yield a product of potential importance to organizations involved in image analysis and image understanding both in the Army and inside and outside of government. More generally, it would advance the state-of-the-art of expert system inferencing and provide a new, highly effective tool to support expert system technology.

APPENDIX A

A.0 APPLICATION OF ALTERNATIVE INFERENCE THEORIES TO PROBLEMS OF IMAGE UNDERSTANDING

A.1 Introduction

In this section we show how different inference theories may be applied to representative problems in image understanding. Our goal is both to extend the evaluation process of Section 2.0 through concrete examples, and to suggest some new ways that some standard problems may be attacked. We start, in Section A.2 with a discussion of how prior context information can be combined with data derived from the pixels. We show how a Bayesian approach, a fuzzy approach, and a Shaferian approach differ in their handling of the same problem. The same kind of arguments are used in Section A.3, where we discuss template matching, and in Section A.4, on relaxation and scene labeling.

A.2 Extraneous Information

A.2.1 Introduction--The problem context. In this section, we shall show how different theories of belief may be applied to a specific example. The problem we have chosen, as suggested by ETL, is in the area of feature extraction from aerial photographs. This is a very complex problem area, as is evidenced by the enormous literature on the subject (see e.g., Rosenfeld, 1983), or the large effort devoted to this, and closely related topics, by DARPA over the last twenty years. In spite of this effort, there appear to have been few attempts to construct an expert system (in the strict AI sense) to effect automatic feature identification from aerial photographs, let alone to use alternative inference schemes within such an expert system. One such system we have discovered in the literature (NEWSIP: Cambier et al., 1983) uses the inference scheme adopted by the PROSPECTOR expert system (Duda et al., 1977), which employs a mixture of ideas from probability theory and fuzzy set theory. NEWSIP is not designed, however, to deal specifically with the problem of forming a consensus of the evidence contained in the image with exogenous information about the geographical area being photographed.

A.2.2 The example. In order to illustrate both how inferences may be drawn from several different sources of information within an expert system and how different theories of belief modification may be used in doing so, we have constructed the following inference task.

> Task: An aerial photograph is available of a known area of countryside. It is known that a single road crosses the area, and that hither to there has been no evidence of any building in the area. The task is to determine if a building has been erected anywhere.

The normal way to handle this problem is to use edge and corner detectors, or texture measures, to segment the image into areas which are then classified into one of several possible categories. Any region classified in this way as a 'building' should be tentatively identified as such. There are now many sophisticated algorithms available to carry out this process automatically (see, for example, Crombie et al., 1982).

These methods do not, however, provide an explicit framework for combining information derived from the photograph with information from other sources. We shall suppose that we also have available the following information:

- In the area represented by the photograph, buildings are usually erected near roads.
- Buildings are not generally erected on boggy ground.
- Some information exists on how boggy the ground is for each point on the photograph.

Our task now is to construct part of an expert system, which will combine this information with that produced by the photograph to determine if a building exists at any point. In the next four sections we describe in detail how that might be achieved, using four different inference theories.

A.2.3 <u>Deterministic inference</u>. We shall assume that we have available a state-of-the-art segmentation algorithm which provides, for any pixel in the image, a set of classification probabilities, $\{p_i\}$. For each possible classification category, i, p_i is the probability that the pixel is indeed correctly classified as belonging to category i (or, more precisely, that the area of land corresponding to the pixel in question belongs to category i). What is of most interest to us is p_B, the probability that the true categorization should be 'building.' (Note, at this stage, that we shall assume that the segmentation algorithm involves appropriate relaxation procedures which relate the classification probabilities at a pixel to those at neighboring pixels.)

As with the other inference schemes that we shall discuss below, there are several possible ways to carry out a deterministic inference.* The following seems a reasonable scheme, however.

We must first convert the somewhat inexact information presented above into precise statements. Somehow, the information on bogginess must be converted into an assessment of whether a particular location can, or cannot, support a building. No degrees of partial truth will be allowed here. The truth value of:

A_1: the ground cannot support a building

will be either 0, false, or 1, true, for each pixel.

*We mean, by the title 'deterministic inference,' a scheme which not only gives an unambiguous answer to the question whether a building does or does not exist at a point, but also one which uses the clearcut implications of standard logic.

Similarly, the distance from the road at which a building becomes impossible must be determined, so that a truth value of 0 or 1 can be associated, for each pixel, with:

A_2: the point is too distant from the road for a building to be present.

The inference engine will now consist of the following rule:

```
IF      (($A_1$ is not true) and ($A_2$ is not true) and ($p_B$>1/2))
THEN    (a building is present)
ELSE    (a building is not present).
```

Writing H for the hypothesis 'a building is present,' this can be computed as

$$\theta(H) = \min(1-\theta((A_1), 1-\theta(A_2), \theta(p_B>1/2))$$

where (H) is the truth value of the hypothesis H and $\theta(p_B>1/2)=1$ if and only if $p_B>1/2$. In this framework $\theta(\text{not } H) = 1-\theta(H)$. This completes the construction of a procedure which will give an unambiguous answer on whether H is true or not.

A.2.4 Probabilistic inference. An obvious drawback to the deterministic inference scheme above is that it forces a somewhat arbitrary classification for locations in terms of their distance from the road, and their bogginess. It is more natural to think of distance and bogginess as being factors which might make a categorization of a pixel as 'building' more or less likely, rather than simply ruling some places out of consideration. A framework for doing this is provided by Bayesian updating.

The probability of H, in the light not only of the pixel data which led to p_B, but also the distance from the road, d, and bogginess of the ground, b, may be written, using Bayes' theorem, as

$$p(H|b,d,D) = \frac{f_1(b,d|H,D) \cdot p_B}{f_2(b,d|D)}$$

where D is all the relevant data provided by the photograph, f_1 is the probability density on b and d given D and the knowledge that H holds, and f_2 is the same density marginalized over (H, not-H). A similar relation holds for $\bar{H}$, the hypothesis that a building is not present. On dividing one relation by the other, we get that the posterior odds on H,

$$O(H|b,d,D) = \frac{p(H|b,d,D)}{p(\bar{H}|b,d,\ D)} = \frac{f_1(b,d|H,D)}{f_1(b,d|\bar{H},D)} \cdot O_B$$

where $O_B = \dfrac{p_B}{1-p_B}$,

the prior odds on a building being present based on the pixel data alone. Now knowledge of the pixel data D will not change our opinion of how likely any particular values of b and d are, once we know whether H holds or not. For example, if we were told that a building <u>was</u> present at a particular location, and asked our opinions on what b or d might be, then the availability of pixel information should not change that view, since it could only do so by affecting opinions about whether H held or not, about which no doubt existed. It follows that f_1 should not depend on D.

We thus obtain the formula

$$O(H|b,d,D) = L(b,d;H) \cdot O_B \qquad (A.1)$$

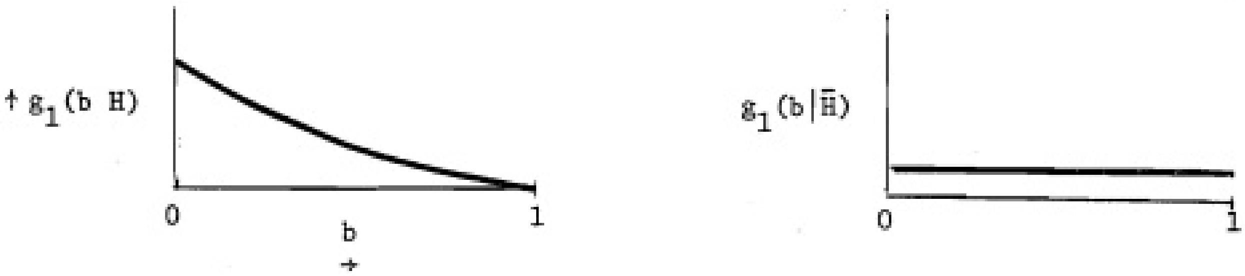
where L is the likelihood ratio for (b,d) in relation to the hypothesis H.

In the event that our views about b and d are independent, in the probabilistic sense, then we can write $f_1(b,d|\cdot)$ as the product of two densities $g_1(b|\cdot)$ and $g_2(d|\cdot)$, thus deriving

$$L(b,d;H) = L_1(b;H) \cdot L_2(d;H)$$

$$\text{where } L_1(b;H) = \frac{g_1(b|H)}{g_1(b|\bar{H})} \qquad \text{and } L_2(d;H) = \frac{g_2(d|H)}{g_2(d|\bar{H})} \quad .$$

The imprecise statement that 'Buildings are not generally erected on boggy ground' can now be represented in the likelihood ratio L_1. If bogginess b is measured on a (0,1) scale with 0 meaning 'not boggy at all,' and 1 measuring 'very boggy,' then the density g_1 will be of the form



The exact form would be determined by elicitation from experts. These curves are reflecting the fact that if a building is present, low bogginess is much more likely than high; whereas if a building is not present, the chance of any particular level of bogginess will just equal the general distribution of bogginess on land of the type analyzed (this distribution need not be flat as in our example). Similar curves for the distance measures would be elicited.

The <u>result</u> of this analysis will be to modify the initial classification probability p_B, according to formula A.1 above. The method of doing it, by multiplying the odds on H by the likelihood ratio L, captures extraneous information about the image under discussion. The effect will be to increase the odds on H for sites with low bogginess and near the road, and to decrease the odds elsewhere.

This probabilistic analysis ends, therefore, with a revised probability that the pixel and its surrounding area should be classified as 'building.' If a definitive answer is required at this stage, a classification could be adopted based on the deduced probability and on the relative costs of classifying a non-building as "building" or a building as "non-building".

A.2.5 Fuzzy inference. Since its inception in 1965, the calculus of fuzzy sets has been used in many different ways to represent imprecision. Zadeh (1983) has provided a good argument for a particular way in which the calculus could be used in the management of uncertainty in expert systems, and we follow his approach here. Zadeh sees a 'serious shortcoming of [existing expert systems in] that they are not capable of coming to grips with the pervasive fuzziness of information in the knowledge base, and, as a result, are mostly ad hoc in nature.' Zadeh's stress on the imprecision of the knowledge base (rather than its uncertainty) is certainly relevant to the example we are considering in this chapter. The statement 'buildings are not generally erected on boggy ground' is clearly imprecise, and in the previous two inferential methods, it had to be made precise before it could be included in the analysis. Fuzzy inference allows this imprecision to persist through the analysis. Zadeh also points out that implication may be imprecise. He handles this by his generalized modus ponens, which we can illustrate with the following example.

The proposition:

If a person is tall then he is heavy,

is represented by a fuzzy relation on variables u and v, describing height and weight respectively. If $\mu_H(v)$ is a fuzzy set describing the meaning of 'heavy', and $\mu_T(u)$ a fuzzy set describing what is meant by 'tall,' then

$$\mu_{T\rightarrow H}(u,v) = \min(1, 1-\mu_T(u)+\mu_H(v))$$

is the membership of the pair (u,v) in the set of (u,v) consistent with (if a person is tall, he is heavy).

This definition may seem somewhat arbitrary, but Zadeh supports it by its consistency with a definition found in Lukasiewicz's logic (see Zadeh, 1983, p. 208). He also calls it a conditional possibility distribution on v given u. To use this implication to say something about the heaviness of a person, given some fuzzy statement about his height (e.g., that he is "very tall"), we use

$$\mu_{(T\to H)\circ T'}(v) = \max_{u}(\min(\mu_T'(u),\ \mu_{T\to H}(u,v));$$

i.e., to find the degree to which a value v could describe the person's weight, we find the most possible height consistent with his being "very tall" (expressed by $\mu T'$) and with the rule that tall people are heavy, and use the height possibility there as the weight possibility measure.

To apply this to the present example, we will need to extend the notions. Instead of a single variable u, we will have two variables: b, the bogginess at a particular site, and d, its distance from the road; instead of v, we will have p, the probability that a building is present. The appropriate equation for $\mu(G\to P)^{o}D'(P)$, the possibility distribution over probabilities that a building is present, which we abbreviate as $\mu_{H|E}(P)$, is

$$\mu_{H|E}(p) = \max_{b,d}(\min(\mu_D'(b,d),\ \min(1,1-\mu_G(b,d)+\mu_P(p))))$$

where $\mu_G(b,d)$ is the possibility distribution for 'the ground is boggy and the location is far from the road,' and $\mu_P(p)$ is the possibility distribution for 'very unlikely.' $\mu_D'(b,d)$ is the representation of the information we have in a special case.

Of course, if we have crisp information about b,d (namely that they are equal to b_0, d_0, so that $\mu(b_0,d_0) = 1$, $\mu(b_1,d_1) = 0$, elsewhere),

then
$$\mu_{H|E}(p) = \min(1,1-\mu_G(b_0,d_0)+\mu_P(p)).$$

This makes a lot of sense: the possibility of a particular probability being true depends in this case only on the imprecision of the implication.

Suppose, by way of example, that we define a membership function for "very unlikely" as follows:

$$\begin{array}{lll} \mu_P(p) & = 1, & \text{for } p \le 0.05 \\ & = 1-\frac{p-0.05}{0.05}, & \text{for } 0.05 \le p \le 0.1 \\ & = 0, & \text{for } p \ge 0.1 \end{array}$$

This gives:

$$\begin{array}{lll} \mu_{H|E}(p) & = 1 & \text{for } p \le 0.1(1-\frac{\mu_G}{2}) \\ & = 3-\mu_G-\frac{p}{0.05} & \text{for } 0.1(1-\frac{\mu_G}{2}) \le p \le 0.1 \\ & = 1-\mu_G & \text{for } 0.1 \le p \end{array}$$

Thus, if $\mu_G = 1$, that is, the ground is clearly boggy and distant from the road, then a building is very unlikely ($\mu_{H|E}(p) = \mu_P(p)$). If, on the other hand $\mu_G = 0$, the ground is clearly not (boggy and distant from the road) then $\mu_{H|E}(p) = 1$, for all p: our evidence does not exclude any probabilities.

This extraneous information needs to be combined with evidence from the pixels. Let us suppose that this evidence can be expressed as another membership function $\mu_{Data}(p)$, for the possibility of a probability p that a building is present. Then combining these two sources of information we get

$$\mu_{Comb}(p) = \min(\mu_{Data}(p), \mu_{H|E}(p)).$$

This will have the effect of reducing the possibilities for probabilities which have low possibility, from the extraneous information, but leaving the others unchanged.

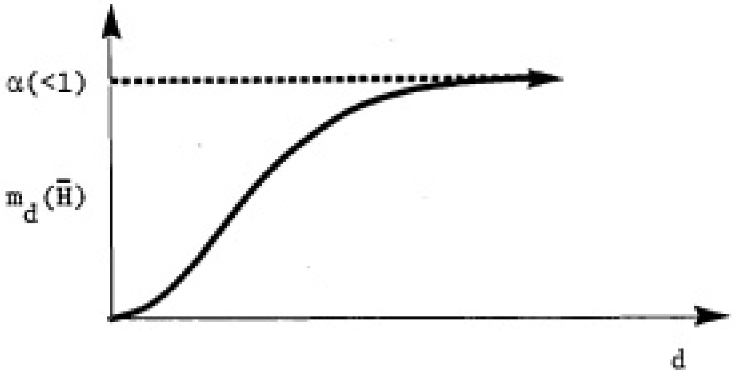
The output of this fuzzy analysis would not be a clearcut answer to the question whether a building is present, nor even a modified probability that it is present, as in the Bayesian case. Rather, it will be a fuzzy probability. This could be used in several ways; we could try linguistic interpretation, producing an output such as 'it is not very likely that a building is present;' we could attempt some sort of fuzzy maximum likelihood analysis; or we could construct a procedure to produce a fuzzy truth value for the hypothesis H. Different theoretical arguments could be produced to support each of these, but we recommend experimental use of a method such as this to explore the practical implications of the different schemes.

A.2.6 Dempster-Shafer inference. Dempster-Shafer theory is concerned with the combination of evidence, and the strength of support that it is proper to have in any subset of the set of hypotheses. In our example we have three pieces of evidence, the distance of a location from the road, the bogginess of the ground, and the evidence from the pixels, D. We shall start by seeing how to represent belief about H in the light of information on bogginess and distance, and how to combine these pieces of evidence.

We construct support functions $m_d(H)$, $m_d(\bar{H})$, $m_d(H \text{ and } \bar{H})$, representing the support given by distance from the road to the hypothesis, its negation and the union of these two hypotheses. In Shafer's theory, the total support allocated to each element of the power set of the set of hypotheses (i.e. each subset of the set of hypotheses) must sum to unity. In this case, since there are only two hypotheses (H and $\bar{H}$), the power set has just 3 elements (H, $\bar{H}$ and (H and $\bar{H}$)), and this requirement gives

$$m_d(H) + m_d(\bar{H}) + m_d(H \text{ and } \bar{H}) = 1.$$

The statement that buildings are usually near roads does not imply that any knowledge about d supports H; it is merely that large distance supports $\bar{H}$. So let us assign $m_d(H)=0$, $m_d(H \text{ and } \bar{H})=1-m_d(\bar{H})$, and $m_d(\bar{H})$ by a curve of the following type:



$m_d(\bar{H})$ can be interpreted as the probability that a distance d implies that $\bar{H}$ is true. It can, in principle, be elicted from an expert.

In a similar way we can construct a support measure $m_b(\cdot)$ based on the evidence of bogginess. Once again it will be very reasonable to ascribe $m_b(H)=0$, $m_b(H \text{ and } \bar{H})=1-m_b(\bar{H})$ and $m_b(\bar{H})$ by an empirical curve of the type above.

To combine evidence, Shafer recommends the use of Dempster's rule, which may be stated as follows. If $m_1(\cdot)$, $m_2(\cdot)$ are the support functions for two different pieces of information, then for any element x in the power set of the set of hypotheses, the support for x in the light of the two pieces of information is

$$m_{12}(x) = \frac{\sum_{y \cap z = x} m_1(y) m_2(z)}{1 - \sum_{y \cap z = \theta} m_1(y) m_2(z)}$$

where θ is the null set.

Using this rule, we see that the support function given both b and d is

$$m_{bd}(H) = 0$$

$$m_{bd}(\bar{H}) = m_b(\bar{H})m_d(\bar{H}) + m_b(\bar{H})(1-m_d(\bar{H})) + (1-m_b(\bar{H}))m_d(\bar{H}) = m_b(\bar{H})+m_d(\bar{H}) - m_b(\bar{H})m_d(\bar{H})$$

$$m_{bd}(H \text{ and } \bar{H}) = [1-m_b(\bar{H})][1-m_d(\bar{H})].$$

We must now combine this support function with a support function deriving from the photographic image. If p_B is the probability of classification as a building, derived from the segmentation algorithm, as in A.2.4 above, then it is reasonable to assign the following support function given the pixel information D.

$$m_D(H) = \alpha p_B$$

$$m_D(H) = \alpha(1-p_B)$$

$$m_D(H \text{ and } H) = 1-\alpha.$$

This reflects the insight that the credibility of the segmentation algorithm may not be total; some of the weight of support (in fact, $1-\alpha$) should be allocated to the complete set of hypotheses, H and $\bar{H}$.

Using Dempster's rule again, we get

$$m_{bdD}(H) = \frac{\alpha\, p_B(1-m_b(\bar{H}))(1-m_d(\bar{H}))}{1-\alpha p_B[m_b(\bar{H})+m_d(\bar{H})-m_b(\bar{H})m_d(\bar{H})]}$$

$$m_{bdD}(\bar{H}) = \frac{\alpha(1-p_B)+(1-\alpha)[m_b(\bar{H})+m_d(\bar{H})-m_b(\bar{H})m_d(\bar{H})]}{1-\alpha p_B[m_b(\bar{H})+m_d(\bar{H})-m_b(\bar{H})m_d(\bar{H})]}$$

$$m_{bdD}(H \text{ and } \bar{H}) = \frac{(1-\alpha)(1-m_b(\bar{H}))(1-m_d(\bar{H}))}{1-\alpha p_B[m_b(\bar{H})+m_d(\bar{H})-m_b(\bar{H})m_d(\bar{H})]}$$

As with the fuzzy version of this problem, there is no agreed procedure now for determining what to do with this support function. We are thinking of using these computations in an automatic feature extraction system, however, and so they must lead to action implications. One approach is parallel to the Bayesian one, with the introduction of a region of indeterminacy in which no answer is provided. Thus, a region is classified as a building if $m_{bdD}(H)$ exceeds some threshold γ and as a non-building if $m_{bdD}(\bar{H})$ exceeds a threshold $1 - \gamma$, where γ is determined by the relative costs of mislabeling a building or a non-building. In some cases, neither threshold will be crossed. An alternative approach, which does always give an answer, is to normalize the support for H and $\bar{H}$, i.e., $p(H) = \dfrac{m(H)}{m(H)+m(\bar{H})}$ and $p(\bar{H}) = 1-p(H)$, before testing against γ. This might be appropriate where the system is to suggest possible buildings for subsequent checking by a human interpreter.

A.3 Template Matching

A.3.1 Introduction. A common problem in analyzing aerial photographs is searching for a particular object, such as a building, in a set of photographs. One way to handle this is through template matching, where portions of the photograph are compared with one, or more, templates, each giving a representation of possible objects. The art of template matching is to construct an algorithm that computes a measure of fit in such a way that the object is properly identified when the measure of fit is good. This idea has been studied in the field of computer vision for many years (see, for example, Cheng et al., 1968). It can be applied either at the level of raw pixel data or at a higher level in which features or relational structures extracted from an image are matched with a stored pattern.

There are problems associated with template matching at the pixel level. First, the appearance of the object may well depend on the illumination, which may be unknown precisely. A partial solution is to normalize both the image and the template, by taking deviations from the mean at each point, before comparing. But in addition, the size and orientation of the object may well not be known in advance, so a great number of possible templates may need to be used in the search; and in certain cases, such as the search for a building, intrinsic qualities such as shape and surface reflectance may also be unknown.

On the other hand, even at the pixel level, template matching is very useful asa filtering technique, e.g., in heightening edges and corners (see Ballard and Brown, 1982). Moreover, some variant of it is usually required to identify the features that are used in a higher-order matching of relational structures. It is, therefore, a good problem for beginning our investigation of the application of belief theories to "bottom up" feature recognition in aerial photographs. In this section, we will first describe the standard approach to template matching, and then go on to show how Bayesian statistics, fuzzy set theory, and Shafer's belief function theory could be used, both to validate an ad hoc approach, and to give reasons for varying the standard approach in certain circumstances.

A.3.2 Standard template matching. Suppose we have an aerial photograph digitized so that it can be represented as a set $\{g(i,j)\}$ of pixel gray levels, where $i=1,\ldots,M$ and $j=1,\ldots,N$ index the pixels in the photograph. Let $t(k,l)$, $k=-m,-m+1,\ldots,0,\ldots,m-1,m$; $l=-n,-n+1,\ldots,0,\ldots,n-1,n$, be a template, that is, a set of gray levels for the ideal object. If the template is centered at (i_0,j_0); then for (k,l) within the template, the difference in gray level at (k,l) is $t(k,l)-g(i_0+k,j_0+l)$.

Clearly the template matches very well if this difference is very small in absolute terms for all (k,l) within the template (i.e. for $k\in[-m,m]$, $l\in[-n,n]$). We need a single measure of goodness-of-fit, for any center point i_0,j_0, to assess how well the template fits at that point. An obvious measure, much used in fitting problems, is the sum of the squared differences,

$$D(i_0,j_0) = \sum_{k=-m}^{m} \sum_{l=-n}^{n} (t(k,l)-g(i_0+k,j_0+l))^2.$$

Note that this is only defined if (i_0,j_0) is sufficiently far away from the boundary of the photograph for all the points to be within range; that is $m \leq i_0 \leq M-m,\ n \leq j_0 \leq N-n$.

The standard algorithm for template matching now seeks (i_0,j_0) to minimize this. Now we can write

$$D(i_0,j_0) = \sum_{k=-m}^{m} \sum_{l=-n}^{n} [t^2(k,l) - 2t(k,l)g(i_0+k,j_0+l) + g^2(i_0+k,j_0+l)].$$

The first term here is independent of (i_0,j_0) and so does not affect the best choice of (i_0,j_0). In some cases, the last term

$$G(i_0,j_0) = \sum_{k=-m}^{m} \sum_{l=-n}^{n} g^2(i_0+k,j_0+l)$$

does not change much with (i_0,j_0) either. If this is the case, then the best (i_0,j_0) is obtained by maximizing

$$C(i_0,j_0) = \sum_{k=-m}^{m} \sum_{k=-n}^{n} t(k,l)g(i_0+k,j_0+l),$$

the correlation of the template with the data. $C(i_0,j_0)$ is, in fact, the result of a finite filter applied to the image, and so in this case it is possible to view template matching as a special case of filtering. This is somewhat contrived, since G is not often constant enough to be neglected. Nonetheless, this is one justification for the selection of important classes of filters, such as edge and corner detectors, and the developments which we shall give in the next sections can be extended to the choice of such detectors.

A.3.3 Bayesian template matching

A.3.3.1 **Probability updating.** The goodness-of-fit measure $D(i_0,j_0)$ adopted in the last section was chosen in a rather arbitrary way. What is at root of interest to us is the probability that the data around the pixel (i_0,j_0) is _really_ a noisy representation of the template. In other words, we can establish the hypothesis

$$H(i_0,j_0):\quad g(i_0+k,j_0+l) = t(k,l) + \varepsilon(i_0,j_0;k,l)$$

where $\varepsilon(i_0,j_0;k,l)$ is an error term.

Then, if $p(i_0,j_0)$ is our prior probability that $H(i_0,j_0)$ holds (i.e., that the object _is_ in fact centered at (i_0,j_0)), Bayes' Theorem gives us

$$p\ (i_0,j_0) = \Pr[H(i_0,j_0)|\{g(i,j)\}] = \frac{f(\{g(i,j)\}|H(i_0,j_0))p(i_0,j_0)}{\sum\limits_{i',j'} f(\{g(i,j)\}|H(i',j'))p(i',j')}$$

where $f(\{g(i,j)\}|H(i_0,j_0))$ is the multivariate density for the $(2m+1)(2n+1)$ values of $g(i,j)$ within the template around (i_0,j_0), given that $H(i_0,j_0)$ holds. We have assumed that one instance of the object is to be found _somewhere_ in the image, so that the set of hypotheses $\{H(i,j)\}$ are mutually exclusive and exhaustive. In general, this will not be the case, and this will lead us to modify the denominator on the right hand side of the equation above. The conclusions of this analysis will not change, however, and so, to avoid inelegant algebra, we will work on the simpler case.

A.3.3.2 **Using loss functions.** We could, at this stage, take the posterior probability, $p_\pi(i_0,j_0)$, as our measure of goodness-of-fit, and identify the object at $(\bar{i},\bar{j})$ where $p_\pi(\bar{i},\bar{j}) = \max p_\pi(i,j)$. Alternatively, we can consider this as a decision problem, recognizing that what matters is the cost of identifying the object to be at (i_1,j_1), when it is, in fact, at (i_2,j_2). Let this cost be $L((i_1,j_1),(i_2,j_2))$. Then the expected cost of making the decision (i_1,j_1) is

$$\bar{L}(i_1,j_1) = \sum_{i_2,j_2} p_\pi(i_2,j_2)L((i_1,j_1),(i_2,j_2)).$$

The best choice of position is at i^*,j^*, where (regarding $L((i_1,j_1),(i_2,j_2))$ as a positive measure of cost)

$$\bar{L}(i^*,j^*) = \min_{i_1,j_1} \bar{L}(i_1,j_1).$$

Note that, in the special case that $L((i_1,j_1),(i_2,j_2)) = 0 \quad \text{if } i_1=j_1,\ i_2=j_2$
$\qquad = 1 \quad \text{elsewhere}$

$$\bar{L}(i_1,j_1) = 1-p_\pi(i_1,j_1)$$

In this case, where all errors are equally costly, $i^*=\bar{i}$, $j^*=\bar{j}$; the problem reduces to maximizing the posterior probability on $H(i,j)$.

Other loss functions will give different procedures, however. For example, suppose

$$\bar{L}((i_1,j_1),(i_2,j_2)) = (i_1-i_2)^2 + (j_1-j_2)^2$$

i.e., the misplacing becomes dramatically more important, the further away the object is placed from its true position. Then

$$L(i^*,j^*) = \min_{i_1,j_1} \Big[\sum_{i_2,j_2} p_\pi(i_2,j_2)((i_1-i_2)^2+(j_1-j_2)^2)\Big]$$

and i^*,j^* are given, to the nearest integer, by

$$i^* = \sum_{i_2,j_2} i_2 p_\pi(i_2,j_2); \; j^* = \sum_{i_2,j_2} j_2 p_\pi(i_2,j_2).$$

In this case, it is best to choose not the <u>most likely</u> location, but an average location, weighted according to probabilities.

A.3.3.3 Recovering the standard algorithm, and some modifications. To carry out the analysis in the previous section, we have, of course, to compute $p_\pi(i,j)$, and this involves the multivariate density $f(\{g(i,j)\}|H(i,j))$, which we have not yet discussed. In one special case, we can derive the simple formula given in Section A.3.2 above which is used in standard template matching.

Suppose $\varepsilon(i,j;k,l)$ has zero mean, is normally distributed, with a variance σ^2 which is independent of (k,l), and that all the error terms are independent.

Then $$f(\{g(i,j)\}|H(i,j)) = \prod_{k=-m}^{+m} \prod_{l=-n}^{+n} \frac{1}{\sqrt{2\pi\sigma}} \exp(-(g(i+k,j+l)-t(k,l))^2/2\sigma^2)$$

$$= (\sqrt{2\pi\sigma})^{(2m+1)(2n+1)} \exp(-\frac{D(i,j)}{2\sigma^2})$$

If, further, $p(i,j)$ is independent of (i,j) (i.e. our prior opinion is that the object is equally likely to be anywhere), then maximizing $p_\pi(i,j)$ is equivalent to minimizing $D(i,j)$.

So we conclude that if:

a) the loss involved in misplacing the object is constant,

b) we have a uniform prior distribution on location,

c) the noise on the image is normally distributed, unbiased, and has constant variance,

d) the noise on the image is uncorrelated,

we recover the standard algorithm - minimize D.

We have already seen, in Section A.3.3.2 above, that if a) does not hold, a different procedure results. The same is true if b), c) or d) are relaxed.

A.3.3.4 **Using prior information.** Suppose that we have prior belief that some lcations are more likely than others for the object, but that conditions a), c) and d) above still hold. Then we should identify the object at $(\bar{i},\bar{j})$, where $(\bar{i},\bar{j})$ maximizes over (i,j)

$$\exp(-D(i,j)/2\sigma^2)\cdot p(i,j).$$

As would be expected, this more or less rules out locations which are extremely unlikely (where $p(i,j)$ is near zero); more significantly, it shows precisely how the sum of squares should be offset to take account of prior opinion.

A.3.3.5 **Systematic error.** It is possible that there could be physical reasons for the error to have a systematic bias, but one that varies over the image. In other words, we could take

$$E(\varepsilon(i,j;k,l)) = \phi(i,j;k,l),$$

(thus changing part of condition c) in Section A.3.3.3). Keeping the other conditions constant, this leads us to want to minimize

$$\sum_{p=-m}^{+m} \sum_{l=-n}^{+n} \Big(g(i+k,j+l)-t(k,l)- \phi(i,j;k,l)\Big)^2$$

This provides another modification of the standard algorithm. We could also, of course, vary condition d), that the noise is uncorrelated to yield yet another modification of the standard algorithm.

A.3.3.6 **Summary**. It should be stressed that the problem we have looked at in this section is somewhat special. We have assumed that the object is to be found at one, and only one location in the image, and that any failure of the template to match is caused by noise. We have excluded the possibility that more than one, or zero, matches exist. The analysis could have been presented for the more general case, but at a cost of clarity in argument.

What we have shown, however, is how Bayesian Decision Theory may guide the choice of a template matching algorithm, taking into account:

(i) the possibly variable cost of a wrong identification,
(ii) the inclusion of prior probabilities on location,
(iii) the effect of correlated noise,
(iv) the effect of systematic bias.

A.3.4 Fuzzy template matching. The theory of fuzzy sets provides an alternative way of representing beliefs within a model. L.A. Zadeh, the originator of the concept of the fuzzy set, stresses that fuzzy sets should be used to handle imprecision, or what is possible, while probability theory should be used to handle uncertainty (see, for example, Zadeh, 1981, p. 70). While there are those who argue that because of imprecision, people are uncertain, and so where information is imprecise, it can be handled through probability theory, it is clear that fuzzy set theory is not a strict alternative to probability; it is, in a sense, a broader theory, saying less than probability theory, but still in keeping with the input information. For example, some values of a variable could be highly possible, but very improbable.

The goal of fuzzy template matching, then, should be to ask to what extent a particular template fits the observed data; the question will be, "How possible is it that what we are observing fits the template?" This question has been previously addressed by Kandel (1982). As is often the case in applications of fuzzy set theory, there are generally many different ways in which the calculus of the theory may be applied to a problem. We shall give two approaches, both of which differ markedly from Kandel's development.

We can first concentrate on the imprecision of our answer to the matching question. When a photo-interpreter analyzes a photograph, he is likely to respond initially with a statement such as: "There could be a building of the type I am looking for just there." This is an imprecise statement, of the kind produced by a fuzzy analysis. When such an analysis yields a result that the possibility of a data-set being derived from a given template is, say, 0.8, one interprets this numerical result by a statement such as that above. In the first instance, let us suppose that the template t(k,l) is precisely defined, but that the imprecision in our answer derives from the fact that the data image is, in essence, an imprecise representation of the template.

One way of looking at this imprecision is on a pixel-by-pixel basis. Comparing a pixel in the data with the corresponding pixel in the template, we can ask, "How possible is it that the gray level in the data is consistent with the gray level in the template?" We can express this as a membership function $\mu_{kl}(g(i+k,j+l),t(k,l))$ using the notation developed in the last section. The construction of this function we shall leave for a moment, but it clearly should depend both on the pixel gray level, g(i+k,j+l) and on the template gray level, t(k,l). We now argue that the degree to which the template fits the data, $\mu_F(i,j)$ is given by

$$\mu_F(i,j) = \min_{k,l} \left(\mu_{kl}(g(i+k,j+l),t(k,l)) \right)$$

This is the rule recommended by fuzzy set theory for finding the possibility for the conjunction of events. We can summarize it by the proverb that a chain is as strong as its weakest link; or observe that, if it is quite impossible for one pixel in the template to be represented by a particular gray level in the data ($\mu_{kl} = 0$), then indeed it is impossible for the template to match, no matter how good the fit is at other pixels. At least in this extreme case, the rule above makes a lot of sense. If, however, it is possible for any data gray level to result from any template gray level at each pixel, then $\mu_{kl} = 1$ for each pixel, and the rule above tells us nothing at all. It is in this sense that fuzzy set theory is bland.

It might be reasonable to suppose that the possibility of a match at a pixel could be given by a function of the form

$$\mu_{kl}(g,t) = 1 - \alpha(g-t)^2$$

So if the match was very good (g=t), the representation would be totally possible; but if the match was as bad as it could be (say, g=0 and t=1, supposing gray levels to be measured on a [0,1] scale), then the degree of possibility would be reduced to $1-\alpha$.

With this formula we would get

$$\mu_F(i,j) = \min_{k,l}\left(1 - \alpha(g(i+k,j+l) - t(k,l))^2\right)$$

Having defined the possibility of a match centered on pixel (i,j) by this formula, we could choose the best match as the point where $\mu_F(i,j)$ is biggest. But this would, to some extent, be contrary to the spirit of fuzzy set theory, where the goal is not to come up with a definitive, clear cut answer, but rather to lead to imprecise, yet informative statements about the problem. If installed in an automatic system, one could set a level of possibility (say 0.9) above which locations could be identified for further study either by human experts, or a more complex expert system.

The second way of using fuzzy set theory in this context is to recognize that the template itself should be imprecise. We are not looking for an exact image in the photograph, but rather for one that is *something like* some sort of norm. So we could specify in advance, for every possible set of gray levels in the image, the extent to which that *could be* the object we are looking for. This could be specified by a membership function

$$\mu_T(t(-m,-n),t(-m+1,-n),\ldots,t(+m,-n);t(-m,-n+1),\ldots,t(+m,-n+1);\ldots;\\ t(-m,+n),\ldots,t(+m,+n)) = \mu_T(\underline{t}),\text{ say.}$$

Setting aside for the moment the difficulty of how to specify a $(2m+1)(2n+1)$ dimensional membership function (even for $m=n=1$ this is a 9-dimensional function), we now see how simple it is to compute the possibility is that the data centered at (i,j) represents the object.

Writing $\underline{g}(i,j)$ for the vector whose components are $g(i-m,j-n)$, $g(i-m+1,j-n),\ldots,$ $g(i+m,j+n)$, we just need to compute

$$\mu_P(i,j) = \mu_T(\underline{g}(i,j))$$

to get the number we require.

Construction of μ_T in the first place will be no simple task, however. One possibility would be to get an expert to rate a large number of images either verbally or numerically. When shown a template-sized image, the expert would respond with how possible it is that what he is seeing represents the object we are looking for; he would either give a membership number, or a verbal response, such as 'highly possible,' 'impossible,' etc., which would then be given a numerical interpretation. After a large number of responses, the membership function would be computed by interpolation (possibly linear). Such a method would be capturing the expertise of a human expert within the computer system--one of the original emphases in expert system research. Notice that this method would have a considerable advantage over other methods in that different orientations, sizes and shapes for the building, as well as different levels of illumination could be handled effectively. A problem might be that sharp dips or peaks which should be present in the multi-dimensional membership function might not be created by a method based on linear interpolation. The alternative method of constructing μ_T by making plausible arguments from first principles may be feasible in certain circumstances, but its feasibility is likely to depend on the size of the template and the nature of the object being sought.

We have seen then how fuzzy set theory may be used as a calculus for imprecise reasoning in template matching in two distinct ways. Both ways should be applied to real data to test their efficiency.

A.3.5 Shaferian template matching. Shafer's theory is designed to provide a method of combining information from distinct sources in the light of what is known about the reliability of those sources. The most obvious way to apply this theory to the template matching problem, then, is to consider the pixel gray levels in the image as being separate data sources, each of which may support the hypothesis that the template matches. This is similar to the case of uncorrelated noise in the Bayesian analysis; we are assuming that if the hypothesis is true (the template fits), then the only reason that the individual gray levels in the pixels are different from those in the template is that some random error in the optical image representation has occurred and that these errors are independent. The concept of independence in Shafer's theory is still being developed, but it is clear that what we need to assume is that it is appropriate to combine evidence using Dempster's rule.

Let us change the notation slightly for convenience of exposition. Label the pixels in the template from 1 to N, rather than with the two indices i and j as before. If t_i is the gray level in the template at the ith pixel, and g_i that in the image for a particular positioning of the template, then our sources of evidence are in pairs (t_i,g_i). If H is the hypothesis that the template fits, then it seems sensible to ascribe a set of support functions by relations of the type

$$\begin{aligned} m_i(H) &= f_1(t_i,g_i) \\ m_i(\bar{H}) &= f_2(t_i,g_i) \\ m_i(H \text{ or } \bar{H}) &= f_3(t_i,g_i) \end{aligned}$$

for some functions $f_j(\cdot,\cdot)$ satisfying $\sum_{i=1}^{3} f_j(t,g) = 1$. The precise form of these functions would depend on what is known about the optical blurring produced when an image is distorted. It might be, for example, that if t and g are both at an extreme of the range of gray levels, then strong support is provided for H, while if t and g are far enough apart, support is given to $\bar{H}$, and if either of them is central while the other is extreme, we can give support to neither (thus giving our support to (H or $\bar{H}$)). Suitable functions displaying these properties are the following:

$$\begin{aligned} f_1(t,g) &= [1-4t(1-t)][1-4g(1-g)][1-(t-g)^2] \\ f_2(t,g) &= [(t-g)^2] \\ f_3(t,g) &= [4t(1-t)+4g(1-g)+16gt(1-g)(1-t)][1-(t-g)^2]. \end{aligned}$$

The combination of these N separate support functions is effected by the repeated application of Dempster's rule. We need some more notation to express this rule here. Let c_i be a variable name for the hypothesis supported by $m_i(\cdot)$; that is $c_i \,\varepsilon (H,\bar{H},(H \text{ or } \bar{H}))$. Then let S_1 be the set of $(c_1,\ldots,c_N)$ whose intersection is H, S_2 the set whose intersection is $\bar{H}$, S_3 the set whose intersection is (H or $\bar{H}$), and S_4 the set whose intersection is the null set.

With these definitions, we can apply Dempster's rule repeatedly, to get the following support functions for the hypotheses:

$$m(H) = \frac{\sum\limits_{S_1} \prod\limits_{i=1}^{N} m_i(c_i)}{1 - \sum\limits_{S_4} \prod\limits_{i=1}^{N} m_i(c_i)}$$

$$m(\bar{H}) = \frac{\sum\limits_{S_2} \prod\limits_{i=1}^{N} m_i(c_i)}{1 - \sum\limits_{S_4} \prod\limits_{i=1}^{N} m_i(c_i)}$$

$$m(H \text{ or } \bar{H}) = \frac{\sum\limits_{S_3} \prod\limits_{i=1}^{N} m_i(c_i)}{1 - \sum\limits_{S_4} \prod\limits_{i=1}^{N} m_i(c_i)}$$

To understand the implications of these expressions, we have computed them for 15 hypothetical example cases when $N = 5$, that is, a five-pixel template. The results are expressed in the table below.

Table A-1: Final Support Functions for a Five-Pixel Template

Case	t_1	g_1	t_2	g_2	t_3	g_3	t_4	g_4	t_5	g_5	m(H)	m($\bar{H}$)	m(H or $\bar{H}$)	$\sum_{i=1}^{5} (t_i-g_i)^2$
1	0.71	0.68	0.48	0.99	0.64	0.13	0.37	0.59	0.13	0.93	0.09	0.79	0.12	1.22
2	0.22	0.31	0.23	0.90	0.59	0.46	0.00	0.83	0.03	0.37	0.11	0.80	0.09	1.26
3	0.50	0.55	0.11	0.36	0.82	0.87	0.41	0.71	0.00	0.60	0.18	0.38	0.44	0.43
4	0.48	0.77	0.10	0.54	0.88	0.82	0.18	0.53	0.18	0.20	0.27	0.26	0.47	0.41
5	0.32	0.37	0.92	0.19	0.58	0.86	0.70	0.13	0.01	0.83	0.08	0.87	0.05	1.62
6	0.04	0.32	0.66	0.08	0.54	0.09	0.64	0.42	0.75	0.73	0.11	0.50	0.39	0.68
7	0.97	0.68	0.21	0.56	0.72	0.66	0.52	0.89	0.96	0.18	0.14	0.69	0.17	0.82
8	0.22	0.04	0.76	0.05	0.87	0.33	0.17	0.07	0.05	0.05	0.66	0.25	0.09	1.68
9	0.26	0.07	0.31	0.52	0.08	0.85	0.21	0.06	0.70	0.23	0.25	0.59	0.15	0.92
10	0.32	0.99	0.19	0.02	0.20	0.59	0.60	0.03	0.41	0.22	0.19	0.59	0.22	0.61
11	0.72	0.73	0.48	0.51	0.65	0.62	0.37	0.37	0.13	0.14	0.31	0.00	0.69	0.0014
12	0.22	0.21	0.23	0.24	0.59	0.58	0.00	0.00	0.03	0.03	0.99	0.01	0.00	0.0001
13	0.51	0.51	0.11	0.11	0.82	0.85	0.41	0.42	0.00	0.00	0.99	0.00	0.01	0.0010
14	0.48	0.49	0.10	0.10	0.88	0.91	0.18	0.18	0.18	0.17	0.76	0.00	0.24	0.0010
15	0.32	0.31	0.92	0.89	0.58	0.60	0.70	0.67	0.00	0.00	0.98	0.00	0.02	0.0019

In the first 10 cases, the pixel gray levels in both the template and image have been chosen at random. As may be observed, in none of these cases are the gray levels close to each other as is evidenced by the moderate values of the sum of squared differences, which we have computed in the last column of the table. Unsurprisingly therefore, little support is given to H in these cases. The only case where m(H) is moderately high, case 8, corresponds to a case where one of the pixels matches very closely, and at an extreme value (the fifth) while the others yield quite inconclusive evidence (the values of m(H or $\bar{H}$) for the four other pixels are 0.71, 0.37, 0.65, 0.67).

Cases 11 to 15 are arranged to be highly correlated, as can be seen from the very small values of the sum of squared differences. In three out of these five cases, as we might expect, very high support is given to H, and in every case virtually no support is given to $\bar{H}$. Case 11 is interesting in that, despite the high correlation, the uncommitted support is still 0.69. This derives from the intermediate values of the gray levels; we constructed our support function so that support for H is only high if t and g are close and at an extreme end of the range.

Once the support functions for the template matching at a particular position have been calculated, we must decide what to do next. One procedure would be to choose the location which maximizes Shafer's plausibility function, which in this case is equal to $m(H) + m(H \text{ or } \bar{H})$. Alternatively we could use the fact that the probability of H is bounded by m(H) and $1-m(\bar{H})$ in this case, and carry out a loss function computation as in the Bayesian analysis of Section A.3.3.2. Since the probability of H would lie in a range, the expected loss would also lie in a range. A further heuristic would be needed (such as minimax loss) to derive a definite conclusion.

We do not pretend that the functions we have used in this analysis are a proper reflection of the best available understanding of the physics of the template matching problem; nor do we believe that the neglect of the relationship between the information connecting pixel data is likely to lead to the best possible analysis; we do believe, however, that a belief function analysis can give insights which simple filtering may not be able to echo.

A.3.6 Summary. As we mentioned in the introduction to this chapter, template matching at the pixel level is subject to problems owing to the imprecision in possible templates, and our uncertainty over how optical conditions might affect the photographic image of the object. We have outlined above how the procedures of Bayesian decision theory, fuzzy set theory, and belief function theory might be applied to this problem to improve the performance of an automatic procedure for searching for a particular object in photographs.

A.4 Relaxation and Scene Labeling

A.4.1 The problem. A common need in interpreting aerial images is to combine tentative identifications for small regions of the image with more general information about the possible relationships of one region to other neighboring regions. An example of this problem, at the pixel level, is how to relate a categorization for each pixel, (i.e., as field, road, water, etc.), to the classifications of neighboring pixels, to ensure reasonable consistency. The seminal paper by Rosenfeld et al. (1976) suggested a method for doing this, which has come to be termed "probabilistic relaxation." A considerable literature has built up on this technique (where it is often described as "standard"), and there is also much experience now of using it in practice (see, for example, Peleg, 1980; Ballard and Brown, 1982; Crombie et al., 1982; Haralick, 1983; and Kittler, 1983). As Haralick (1983) has pointed out, however, "probabilistic relaxation has been a mechanism whose theory has not been well understood." It was developed to attempt modification of crude probabilistic estimates of the labeling (or categorization) of each basic unit, in the light of information at neighboring units. As Haralick (1983) suggests, however, there are alternative ways of achieving this goal, particularly if one sets the problem in a larger context than low-level "pixel-pushing" (to use a phrase of Haralick's (private communication)).

In this chapter, we shall present a Bayesian formulation of the problem much as Haralick (1983) does; but we shall show how a slightly different formulation can work on the scene labeling problem first suggested in Rosenfeld's 1976 paper. We shall generalize this as an example of conflict resolution when different kinds of basic labeling algorithms are available. Then we discuss Shafer's account of Rosenfeld's problem, and show how his theory may be combined with the Bayesian one. Finally, we discuss Rosenfeld's own application of fuzzy set theory to this problem, and how it might be modified.

A.4.2 Bayesian analysis. Suppose we wish to label n objects with a set of labels $L \equiv \{\lambda_j : j=1,\ldots,m\}$. This could either be the pixel labeling problem, or, at a higher level of image understanding, scene labeling once a segmentation algorithm has been applied to identify elemental regions of the image. For each of the n objects separately, data D_i is available on which to base the choice of label for that object. Moreover, we have prior information about which sets of labelings are more likely than others which we assume can be expressed as a prior probability distribution

$$p(\underline{l}) = \Pr[\text{label of the ith object is } l_i,\ i=1,\ldots,n].$$

This will be zero for labeling combinations, $\underline{l}$, that are impossible; unlike the assumption made by Haralick (1983, p.422), we observe that some labelings $\underline{l}$ with non-zero probability may be more likely than others, and this will be determined by our *prior* knowledge of the kinds of sets of objects which we may *expect* to find in an image of the kind we are looking at. We will discuss how to specify our prior distribution in the example of the next section. The quantity of interest to us is what chance should be associated with each labeling $\underline{l}$, in the light of the data set $\{D_i\colon i=1,\ldots,n\}$. We use Bayes' formula to express this quantity as

$$p(\underline{l}\,|\,\{D_i\}) = \frac{\Pr[\{D_i\}\,|\,\underline{l}]}{\Pr[\{D_i\}]} \cdot p(\underline{l}).$$

Now we follow Haralick, and suggest that since for any object the data D_i will depend only on the true labeling of that object, we can express

$$\Pr[\{D_i\}\,|\,\underline{l}] = \prod_{i=1}^{n} \Pr[D_i\,|\,l_i].$$

For example, in the scene labeling problem, the data D_i might be a texture vector which should discriminate between water, forests, buildings, etc. The chance of getting a particular texture vector from an object which is *really* a field should not depend (it can be plausibly argued) on whether the neighboring regions are buildings, forests or lakes, or on the texture vectors obtained from neighboring regions.

Using these equations, we get

$$p(\underline{l}\,|\,(D_i)) = \frac{\prod_{i=1}^{n} \Pr[D_i|l_i]}{\sum_{\underline{l}'} \prod_{i=1}^{n} \Pr[D_i|l_i']\,p(\underline{l}')}\; p(\underline{l}). \tag{A.2}$$

Now we see that our result depends only on $p(\underline{l})$, and $\Pr[D_i|l_i]$. We have discussed the first of these above. The second could be assessed directly, as Haralick (1983) implicitly assumes, and we suggest that this may be the most satisfactory approach. One of our purposes here, however, is to show how a Bayesian approach differs from the non-linear relaxation method of Rosenfeld et al. (1976). The inputs in that process are not the conditional probabilities on the data given the label, but the inverse conditional probabilities, $\Pr[l_i|D_i]$. If we are to be coherent, it is not possible to specify these probabilities independently of $p(\underline{l})$, our prior opinion on labels, since

$$\sum_{D_i} \Pr[l_i|D_i]\Pr[D_i] = \Pr[l_i].$$

$\Pr[D_i]$ will not need to be assessed in our subsequent analysis; all we need is to assure ourselves that a set of probabilities $\Pr[D_i]$ (or a distribution, if the data are continuous) exists which allows a particular assessment of $\Pr[l_i]$ to be consistent with the algorithm for finding $\Pr[l_i|D_i]$. This will be the case so long as the m-vector $\Pr[l_i=\lambda_k]$, $k=1,\ldots,m$, is in the convex hull of the vector $\Pr[l_i=\lambda_k\ D]$, $k=1,\ldots,m$, for all D which are possible. This is unlikely to be much of a restriction, and can be checked in a working algorithm. We shall continue our analysis assuming that $\Pr[l_i|D_i]$ and $\Pr[l_i]$ can be separately specified.

Now given that we can take the statistical interaction between the label and the data to be localized, we have

$$\Pr[D_i | l_i] = \frac{\Pr[l_i | D_i]}{\Pr[l_i]} \Pr[D_i]$$

and inserting this in the formula above, we get

$$p(\underline{l} | \{D_i\}) = K \left[\left\{ \prod_{i=1}^{n} \Pr[l_i \; D_i] \right\} \left\{ \frac{p(\underline{l})}{\prod_{i=1}^{n} \Pr[l_i]} \right\} \right], \qquad \text{(A.3)}$$

where K is a normalization factor which ensures $\sum_{\underline{l}} p(\underline{l} | \{D_i\}) = 1$, i.e., that we are really dealing with a probability distribution over possible labelings $\underline{l}$. Notice that in this formulation we do not have to assess probabilities of getting the data $\{D_i\}$ either conditional on the labeling, or marginal over labelings. This assessment task, which could be very difficult in the case of continuous multidimensional variables, such as texture vectors, has been replaced by the apparently more tractible problem of assessing conditionals on labels given the data, for each object independently. (We note that the advantage in doing it this way may be more apparent than real, however.)

A second apparent advantage of this formulation is that is separates (a) assessment of the probability of each l_i considering only the corresponding D_i, from (b) assessment of the impact of interdependencies among the set of l_i on the probability of $\underline{l}$. Note that the ratio on the right hand side, between $p(\underline{l})$ and the product of the $\Pr[l_i]$ is a measure of the degree to which non-independence among the l_i supports or detracts from the likelihood of a particular set of labels, $\underline{l}$. To the extent that the ratio exceeds (falls below) 1.0, the l_i (do not) "belong together" and $p(\underline{l} | \{D_i\})$ is increased (decreased).

We suggest that this scheme is a more satisfactory way of handling the input information which Rosenfeld uses in his nonlinear probabilistic relaxation method than the procedures of that method itself. This is not to say that probabilistic relaxation should not be used, since as a numerical method it can clearly produce sensible practical results. Rather, we should interpret the computations of probabilistic relaxation either as Haralick (1983) does, as a process of sequentially including more and more information; or, as Hummel and Zucker (1983) do, as not being probabilistic at all. With the latter interpretation, we can think of relaxation as being a sensible heuristic technique for deriving consistent labelings, or even as a non-probabilistic method for generating probabilities, to be contrasted with the more intelligible probabilistic approach, given by the formula above.

A.4.3 Rosenfeld's example. To illustrate the difference between our suggested method, and non-linear relaxation, we shall apply it to the example that is used in Rosenfeld et al. (1976). A triangle is identified in an image, and the scene interpreter has to make a three-dimensional interpretation of this triangle on the basis of information about each of the three lines. Each line can be labeled with one of four labels, which we shall call λ_1, λ_2, λ_3, and λ_4, and of the 4^3=64 possible labelings, only eight are possible, as listed in the table below. The reader is referred to Rosenfeld et al. (1976) for a precise meaning of these labels and the eight interpretations of the triangle.

Table A-2: The Eight Possible Labelings

Labeling of side:	$\underline{l}^{(1)}$	$\underline{l}^{(2)}$	$\underline{l}^{(3)}$	$\underline{l}^{(4)}$	$\underline{l}^{(5)}$	$\underline{l}^{(6)}$	$\underline{l}^{(7)}$	$\underline{l}^{(8)}$
1	λ_1	λ_2	λ_1	λ_1	λ_3	λ_2	λ_2	λ_4
2	λ_1	λ_2	λ_3	λ_1	λ_1	λ_4	λ_2	λ_2
3	λ_1	λ_2	λ_1	λ_3	λ_1	λ_2	λ_4	λ_2

Prior information is that each of these labelings is equally likely; this being so, $p(\underline{l}^{(k)}) = 1/8$, for each k. Moreover, we must use this information to give the prior marginals for each label on each side. For side 1, this gives $p(l_1=\lambda_1)=3/8$; $p(l_1=\lambda_2)=3/8$; $p(l_1=\lambda_3)=1/8$; $p(l_1=\lambda_4)=1/8$. (For example, $p(l_1= {}_2) =$

$p(\underline{l}^{(2)}) + p(\underline{l}^{(6)}) + p(\underline{l}^{(7)}.)$ But because of the symmetry in the prior information, we find the marginals to have the same values for sides 2 and 3 as they have for side 1. We can now compute the second factor in braces in the expression for the posterior distribution, $p(\underline{l}|\{D_i\})$, given at the end of the last section, i.e., the interpendence ratio discussed in the last section. This is the joint distribution for the labeling input, divided by the product of the marginals:

	Interdependence Ratio
$\underline{l}^{(1)}$	2.37
$\underline{l}^{(2)}$	2.37
$\underline{l}^{(3)}$	7.11
$\underline{l}^{(4)}$	7.11
$\underline{l}^{(5)}$	7.11
$\underline{l}^{(6)}$	7.11
$\underline{l}^{(7)}$	7.11
$\underline{l}^{(8)}$	7.11

The lower ratios for $\underline{l}^{(1)}$ and $\underline{l}^{(2)}$ reflect the fact that the labels they involve (λ_1 and λ_2) are more frequent in the possible labelings than λ_3 or λ_4; thus, for example, the cooccurrence of λ_1's in $\underline{l}^{(1)}$ may more due to chance (rather than interdependence) than the occurrence of λ_3, λ_1, and λ_1 in $\underline{l}^{(5)}$.

In order to make a comparison between our method and that of Rosenfeld, we have computed the posterior probabilities by our formula using these ratios, for each of the eight examples of input probabilities suggested by Rosenfeld, as given in Table A-3.

Table A-3: Input Identification Probabilities

Case	$p(l_1 \mid D_1)$				$p(l_2 \mid D_2)$				$p(l_3 \mid D_3)$			
	$l_1 =: \lambda_1$	λ_2	λ_3	λ_4	$l_2 : \lambda_1$	λ_2	λ_3	λ_4	$l_3 : \lambda_1$	λ_2	λ_3	λ_4
A	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
B	0.5	0	0.5	0	0.5	0	0.5	0	0.5	0	0.5	0
C	0.5	0	0.5	0	0.4	0	0.6	0	0.5	0	0.5	0
D	0.5	0	0.5	0	0.3	0	0.7	0	0.5	0	0.5	0
E	0.3	0	0.7	0	0.3	0	0.7	0	0.5	0	0.5	0
F	0.2	0	0.8	0	0.3	0	0.7	0	0.5	0	0.5	0
G	0.3	0.2	0.3	0.2	0.3	0.2	0.3	0.2	0.3	0.2	0.3	0.2
H	0.3	0.2	0.3	0.2	0.25	0.25	0.25	0.25	0.2	0.2	0.4	0.2

Table A-4 below contains the results of the computations, giving the posterior probability of each of the possible interpretations being correct, based on our Bayesian formula (B), and on Rosenfeld's non-linear relaxation method (R).

Table A-4: Posterior Probabilities

Case:	A		B		C		D		E		F		G		H	
Labeling	B	R	B	R	B	R	B	R	B	R	B	R	B	R	B	R
$\underline{l}^{(1)}$	1/20	1/8	1/10	1	2/23	1	1/14	0	1/18	1	1/23	0	27/350	1	3/59	0
$\underline{l}^{(2)}$	1/20	1/8	0	0	0	0	0	0	0	0	0	0	8/350	0	2/59	0
$\underline{l}^{(3)}$	3/20	1/8	3/10	0	9/23	0	7/14	1	7/18	0	7/23	0	81/350	0	9/59	0
$\underline{l}^{(4)}$	3/20	1/8	3/10	0	6/23	0	3/14	0	3/18	0	3/23	0	81/350	0	18/59	1
$\underline{l}^{(5)}$	3/20	1/8	3/10	0	6/23	0	3/14	0	7/18	0	12/23	1	81/350	0	9/59	0
$\underline{l}^{(6)}$	3/20	1/8	0	0	0	0	0	0	0	0	0	0	24/350	0	6/59	0
$\underline{l}^{(7)}$	3/20	1/8	0	0	0	0	0	0	0	0	0	0	24/350	0	6/59	0
$\underline{l}^{(8)}$	3/20	1/8	0	0	0	0	0	0	0	0	0	0	24/350	0	6/59	0

We have represented the probabilities in Table A-4 as fractions rather than decimals in order for the reader to see probability ratios more easily.

Notice that in cases D, F and H, the relaxation result is to pick out the most likely labeling; what is more interesting are cases B, C, E and G where a labeling which is not the most likely is chosen (in case E it is only 1/7 as likely). The results of the Bayesian algorithm in case A may seem surprising; since the data gives each label to be equally likely for each side, and each interpretation to be equally likely, would it not seem more reasonable to use the relaxation result, that each labeling should be equally likely, posterior to getting the data? This inference is false, however, because the labels are <u>not</u> distributed uniformly in the possible labelings; if the data suggest that a side is just as likely to have label λ_3 as λ_1 for example, this favors labelings $\underline{l}^{(3)}$, $\underline{l}^{(4)}$ and $\underline{l}^{(5)}$, over $\underline{l}^{(1)}$, since it must give more weight to the few appearances of label λ_3.

A.4.4 <u>An alternative Bayesian analysis</u>. An important observation can be made regarding the Bayesian analysis in the last section, namely that the meaning of the input conditional probabilities, $p(l_i|D_i)$, may in some cases be unclear. To illustrate this point, and also to illuminate the triangle example, we shall now construct a simple example of a labeling problem and discuss the issue in the context of that problem.

Suppose that a room contains a large number of urns, of two types, A and B. Type A urns contain 50% black balls and 50% white balls, while type B urns contain 80% black balls and 20% white balls. A probabilistic labeling procedure (analogous to the line labeling algorithm for the previous example) consists of taking a random sample of size n from any urn, with replacement. This will give the following probabilities for getting r black and n-r white balls from the urn.

$$\Pr[r|A] = \binom{n}{r}(0.5)^n$$

$$\Pr[r|B] = \binom{n}{r}(0.8)^r(0.2)^{n-r}$$

So the algorithm yields, in the general notation $\Pr[D_i|l_i]$, and not $\Pr[l_i|D_i]$. As we mentioned previously, it would be much more straightforward to do a Bayesian analysis supposing that $\Pr[D_i|l_i]$ were the numbers produced by the line labeling algorithm in the triangle case; indeed Haralick's analysis of the general case does make this assumption. Let us suppose, however, that we must deal with $\Pr[l_i|D_i]$.

Suppose, in our simple example, we are now presented with a pair of urns, and we are asked for a labeling of the pair. We have, from Bayes' Theorem, and using an obvious notation,

$$\Pr[A_1,A_2|r_1,r_2] = \frac{\Pr[r_1,r_2|A_1,A_2]}{\Pr[r_1,r_2]}\Pr[A_1,A_2]$$

with similar expressions for the other labeling pairs (A_1,B_2), (B_1,A_2) and (B_1,B_2). The analysis of Section A.4.2 now gives

$$\Pr[A_1,A_2|r_1,r_2] = K\left[\Pr[A_1|r_1]\Pr[A_2|r_2]\left\{\frac{\Pr[A_1,A_2]}{\Pr[A_1]\Pr[A_2]}\right\}\right]$$

But now we must ask how $\Pr[A_1|r_1]$ is computed. Clearly in the triangle example it should be determined by the very formula that led to its inclusion in the expression above, namely

$$\Pr[A_1|r_1] = \frac{\Pr[r_1|A_1]\Pr[A_1]}{\Pr[r_1]} \tag{A.4}$$

Substitution of (A.4) in the previous equation leads to the equivalent, in this context, of Haralick's equation, (A.2). If, of course, $\Pr[A_1]$ is subjectively assessed, then there is no reason why we should not think of $\Pr[A_1|r_1]$ as also being subjectively assessed. But even if this is the case, it is clear that its assessment must be made in awareness of the relationship (A.4) above which must hold. In summary then, the identification of the input numbers in the examples of Section A.4.3 as conditional probabilities of labels given data is appropriate only in the <u>absence</u> of an understanding of the data generation process comparable to the understanding we have in the urn sampling example; i.e., if we clearly understand how often a given true label will produce a given set of data D_i, we should use equation (A.2) rather than equation (A.3).

Let us suppose, then, that we have such an understanding. We can offer an alternative Bayesian interpretation of the triangle example of the last section, which utilizes Rosenfeld's data, if the numbers in Table A-3 are taken, not as probabilities of the labels given the data, but as the relative sizes of the probabilities of data given the labels. For example, we might have, in case A;

$$\Pr[D_1|l_1=\lambda_1]:\Pr[D_i|l_1=\lambda_2]:\Pr[D_1|l_1=\lambda_3]:\Pr[D_1|l_1=\lambda_4]$$

$$= 0.25:0.25:0.25:0.25.$$

With this revised interpretation, we can recompute the posterior probabilities using equation (A.2). The table below gives the results of this calculation, again with Rosenfeld's solutions for comparison.

Table A-4': Posterior Probabilities--Revised Interpretation

Case:	A		B		C		D		E		F		G		H	
Labeling	B	R	B	R	B	R	B	R	B	R	B	R	B	R	B	R
$\underline{l}^{(1)}$	1/8	1/8	1/4	1	2/9	1	3/16	0	3/20	1	3/25	0	27/140	1	3/25	0
$\underline{l}^{(2)}$	1/8	1/8	0	0	0	0	0	0	0	0	0	0	8/140	0	2/25	0
$\underline{l}^{(3)}$	1/8	1/8	1/4	0	3/9	0	7/16	1	7/20	0	7/25	0	27/140	0	3/25	0
$\underline{l}^{(4)}$	1/8	1/8	1/4	0	2/9	0	3/16	0	3/20	0	3/25	0	27/140	0	6/25	1
$\underline{l}^{(5)}$	1/8	1/8	1/4	0	2/9	0	3/16	0	7/20	0	12/25	1	27/140	0	3/25	0
$\underline{l}^{(6)}$	1/8	1/8	0	0	0	0	0	0	0	0	0	0	8/140	0	2/25	0
$\underline{l}^{(7)}$	1/8	1/8	0	0	0	0	0	0	0	0	0	0	8/140	0	4/25	0
$\underline{l}^{(8)}$	1/8	1/8	0	0	0	0	0	0	0	0	0	0	8/140	0	2/25	0

Once again there are marked differences from the Rosenfeld analysis.

Further evaluation of the Bayesian inference schemes we have developed above will depend on their application to real scene labeling problems, as an alternative to relaxation labeling, to determine if empirically useful results can be obtained.

A.4.5 Bayesian analysis of conflict from more than one labeling algorithm. In some cases more than one probabilistic classifier is available to give input probabilities for the labeling of each object in the light of data, $Pr[l_i|D_i]$ or $Pr[D_i|l_i]$. We can think of these as being different because they are based on different data, D_i and D_i', say. This is not unreasonable, if the methods are based on different ways of handling the fundamental inputs of image analysis, namely the gray levels at the pixels. We shall consider an alternative interpretation, namely that the methods have different reliabilities, in a later section.

We are now interested in computing the posterior probability on $\underline{l}$ given the two data sources, $\{D_i\}$ and $\{D_i'\}$. This is given by

$$p(\underline{l}|\{D_i\},\{D_i'\}) = \frac{Pr[\{D_i'\}|\{D_i\},\underline{l}]Pr[\underline{l}|\{D_i\}]}{Pr[\{D_i'\}|\{D_i\}]}$$

$$= \frac{Pr[\{D_i'\}|\{D_i\},\underline{l}]Pr[\{D_i\}|\underline{l}]p(\underline{l})}{Pr[\{D_i'\}|\{D_i\}]Pr[\{D_i\}]}.$$

Now once a labeling $\underline{l}$ has become known, the chance of getting particular data $\{D_i'\}$ will not depend on $\{D_i\}$. Hence, we may write

$$Pr[\{D_i'\}|\{D_i\},\underline{l}] = Pr[\{D_i'\}|\underline{l}].$$

We could leave matters there, and simply input values of $Pr[\{D_i\}|\underline{l}]$ and $Pr[\{D_i'|\underline{l}]$. But to follow our comparison with Rosenfeld's analysis, we could adopt the first Bayesian interpretation (of Section A.4.2) to get

$$p(\underline{l}|\{D_i\},\{D_i'\}) = K'\left[\left\{\prod_{i=1}^{n}\frac{Pr[l_i|D_i']}{Pr[l_i]}\right\}\cdot\left\{\prod_{i=1}^{n}\frac{Pr[l_i|D_i]}{Pr[l_i]}\right\}p(\underline{l})\right]$$

where K' is another normalizing constant. This expression is symmetric in the two data sources, as we would expect.

To see how this would affect the computations, suppose the first data source yields the identification probabilities given by entry A in Table A-3, but that the second data source yields the identification probabilities of case B in that table. In this case, the posterior probabilities for the 8 possible labelings, $\underline{l}^{(1)},\ldots,\underline{l}^{(8)}$ are, respectively 1/28(1,0,9,9,9,0,0,0). As we would expect, this gives an interpretation which is different from A and B. Like B, it gives zero probability to four of the labelings, since one of the methods has shown them to be impossible; it also suggests $\underline{l}^{(1)}$ is less likely than <u>either</u> independent data source would suggest; here the second method, B,is confirming the small change indicated by A, thus reducing it.

A.4.6 <u>Shafer's approach to the triangle identification problem</u>. In a discussion of how to apply his belief theory to the problem of combining dependent evidence Shafer (1984b) touches upon Rosenfeld's scene labeling problem. Shafer's criticism of Rosenfeld's method, as an argument for the proper selection of frames when combining evidence, is of less interest to us than his recommendation of how the problem should be analyzed.

He suggests that the data which give probabilistic labelings for each side of the triangle should lead to the construction of three independent belief functions over the frame consisting of the 64 labeling combinations. The first three of these are derived from the pixel data for each side; the fourth comes from the prior information regarding which interpretations are possible. The pixel information corresponds to case B of Table A-4 above. Table A-5 below gives Shafer's allocation of support; the notation is self-explanatory, and we only quote the subsets of the set of hypotheses which are given non-zero support.

Table A-5: Shafer's Four Support Functions

m_1	m_2	m_3	m_4
$m_1(\lambda_1,\{\lambda_i\},\{\lambda_i\})=1/2$ $m_1(\lambda_3,\{\lambda_i\},\{\lambda_i\})=1/2$	$m_2(\{\lambda_i\},\lambda_1,\{\lambda_i\})=1/2$ $m_2(\{\lambda_i\},\lambda_3,\{\lambda_i\})=1/2$	$m_3(\{\lambda_i\},\{\lambda_i\},\lambda_1)=1/2$ $m_3(\{\lambda_i\},\{\lambda_i\},\lambda_3)=1/2$	$m_4(\lambda_1,\lambda_1,\lambda_1)=1/8$ $m_4(\lambda_2,\lambda_2,\lambda_2)=1/8$ $m_4(\lambda_1,\lambda_3,\lambda_1)=1/8$ $m_4(\lambda_1,\lambda_1,\lambda_3)=1/8$ $m_4(\lambda_3,\lambda_1,\lambda_1)=1/8$ $m_4(\lambda_2,\lambda_4,\lambda_2)=1/8$ $m_4(\lambda_2,\lambda_2,\lambda_4)=1/8$ $m_4(\lambda_4,\lambda_2,\lambda_2)=1/8$

The notation $\{\lambda_i\}$ is short for $\{\lambda_1,\lambda_2,\lambda_3,\lambda_4\}$, the union of the hypotheses that each of the four labels is correct.

We now combine these four support functions, using Dempster's rule, to get

$$m_{1234}(\lambda_1,\lambda_1,\lambda_1)=1/4;\ m_{1234}(\lambda_1,\lambda_3,\lambda_1)=1/4;$$

$$m_{1234}(\lambda_1,\lambda_1,\lambda_3)=1/4;\ m_{1234}(\lambda_3,\lambda_1,\lambda_1)=1/4$$

with zero support to all other combinations of hypotheses.

Note that the suggestion of this analysis is that we should give equal support to the labelings $\underline{l}^{(1)}$, $\underline{l}^{(3)}$, $\underline{l}^{(4)}$, and $\underline{l}^{(5)}$; this is in sharp contrast to the results of the first Bayesian analysis of Section A.4.3, where the posterior probabilities were 1/10, 3/10, 3/10, 3/10. The distinction is caused by the handling of prior belief about label λ_3. In the first Bayesian analysis, recognition that we would expect λ_3 to be only 1/3 as likely as λ_1 on any side, instead of just as likely, as the data suggest, leads us to conclude that labelings containing λ_3 are more likely (in fact, three times as likely) as $\underline{l}^{(1)}$ which does not contain λ_3.

The Bayesian analysis would be recovered if different support functions for m_1, m_2, and m_3 were used. If we were to think of the support for the labels given the data as relative to the underlying support for the labels, based on m_4, then we might take

$$m_1(\lambda_1,\{\lambda_i\},\{\lambda_i\})=1/4;\ m_1(\lambda_3,\{\lambda_i\},\{\lambda_i\})=3/4$$

with similar assignments for m_2 and m_3. Using Dempster's rule on these, we recover the Bayesian results. An important point to make here is that the meaning of Shafer's support functions is very significant.

Alternatively, and perhaps more acceptably, we can compare Shafer's analysis with the second Bayesian interpretation above. In that case, Shafer's support function of Table A-5 leads to results which are consistent with column B of Table A-4'.

We conclude that Shafer's approach has nothing to offer over a Bayesian theory when applied in this way to this problem. But there are ways in which it can provide greater insight, as we shall describe in the next section.

A.4.7 Conflict between two or more evidence mechanisms. Let us now suppose, as we did in Section A.4.4, that in making local assessments of the appropriateness of a label for each object separately, we have two competing inference procedures. Instead of imagining, however, that each of these procedures produces probabilities that the label of each object should be a particular label, let us suppose that we specify support functions $m_1(\cdot)$, $m_2(\cdot)$ on the set of all subsets of labelings.

Thus it might be that the data either point unambiguously to label λ_1, with probability α, say; or, with probability β, the data point to $\{\lambda_2,\lambda_3\}$, but fail to distinguish between them; or, with probability $1-\alpha-\beta$ do not tell us anything.

This would lead to the following support function:

$$m(\{\lambda_1\})=\alpha;\ m(\{\lambda_2,\lambda_3\})=\beta;\ m(\{\lambda_1,\lambda_2,\ldots,\ _m\})=1-\alpha-\beta$$

and m(C)=0 for C being any other subset of the set of labels. As we pointed out above, the probabilities could be thought of as relative to the underlying probabilities.

If two different methods were available for labeling on the basis of low-level data about each object, and these labelings were in <u>conflict</u>, we can now see how to use Shafer's theory to combine this evidence, and prior evidence, to illuminate the labeling problem. Specifically, suppose each object can be addressed by two different inference procedures, but that these are applied to each object separately. Application to the ith object will lead to support functions

$$m_{ij}(\ldots;\{\lambda_1,\ldots,\lambda_m\};\ldots;x;\ldots;\{\lambda_1,\ldots,\lambda_m\};\ldots) \qquad \text{for } j=1,2$$

where x is any subset of the set of labels and it is in the ith position in the list of arguments. This notation implies that, while the frame for the support function actually has $(2^m-1)^n$ elements (there are 2^m-1 possible sets of labels for each of the n objects), the inference procedure operating on the ith object does not have anything to say about the other n-1 objects, and so the support function for the ith object allocates positive measure only to the universal set of labels $\{\lambda_1,\ldots,\lambda_m\}$ for all objects except the ith. Dempster's rule is now applied to the 2n support functions thus prescribed, to produce a combined support function $m_D(\cdot)$; this is then, in its turn, combined with the prior support function $m_P(\cdot)$, again by Dempster's rule, to give a final support function for subsets of the set of all labeling n-tuples.

To illustrate this rather complex description, let us return to Rosenfeld's triangle example. Suppose that the six support functions in Table A-6 are obtained by application of two distinct line labeling algorithms to the three sides of the triangle.

Here we have abbreviated the notation. The labels in a support function m_{ij} just refer to the ith object; m_{ij} gives exclusive support to the complete set $\{\lambda_1,\lambda_2,\lambda_3,\lambda_4\}$ for objects other than the ith. Let us demonstrate how Dempster's rule is now used. First let us construct $m_{1,12}(\cdot)$ by combining the first two belief functions in table A-6, again using the abbreviated notation.

$$m_{1,12}(\lambda_1) = \frac{\begin{array}{c} m_{11}(\lambda_1)m_{12}(\lambda_1)+m_{11}(\lambda_1)m_{12}(\lambda_1\lambda_2\lambda_3)+m_{11}(\lambda_1)m_{12}(\lambda_1\lambda_3)+m_{11}(\lambda_1)m_{12}(\lambda_1\lambda_2\lambda_3\lambda_4) \\ +m_{11}(\lambda_1\lambda_3)m_{12}(\lambda_1) + m_{11}(\lambda_1\lambda_2\lambda_3\lambda_4)m_{12}(\lambda_1) \end{array}}{1-m_{11}(\lambda_2\lambda_4)m_{12}(\lambda_1)-m_{11}(\lambda_2\lambda_4)m_{12}(\lambda_1\lambda_3)}$$

The numerator of this expression is the sum of products of support functions for subsets whose intersection is <u>exactly</u> λ_1; the demonimator differs from one by a similar sum over subsets with a null intersection.

Using similar methods, we derive the following support functions.

Table A-7: A First Application of Dempster's Rule

$m_{1,12}$		$m_{2,12}$		$m_{3,12}$	
$m_{1,12}(\lambda_1)$	= 0.744	$m_{2,12}(\lambda_1)$	= 0.097	$m_{3,12}(\lambda_1)$	= 0.904
$m_{1,12}(\lambda_1\lambda_3)$	= 0.116	$m_{2,12}(\lambda_3)$	= 0.861	$m_{3,12}(\lambda_3)$	= 0.048
$m_{1,12}(\lambda_2\lambda_4)$	= 0.023	$m_{2,12}(\lambda_1\lambda_3)$	= 0.014	$m_{3,12}(\lambda_1\lambda_3)$	= 0.048
$m_{1,12}(\lambda_1\lambda_2\lambda_3)$	= 0.047	$m_{2,12}(\lambda_2)$	= 0.014		
$m_{1,12}(\lambda_2)$	= 0.047	$m_{2,12}(\lambda_1\lambda_2\lambda_3)$	= 0.014		
$m_{1,12}(\lambda_1\lambda_2\lambda_3\lambda_4)$	= 0.023				

Table A-6: Support Functions

m_{11}	m_{12}	m_{21}	m_{22}	m_{31}	m_{32}
$m_{11}(\lambda_1) = 0.4$ $m_{11}(\lambda_1\lambda_3) = 0.2$ $m_{11}(\lambda_2\lambda_4) = 0.2$ $m_{11}(\lambda_1\lambda_2\lambda_3\lambda_4) = 0.2$	$m_{12}(\lambda_1) = 0.6$ $m_{12}(\lambda_1\lambda_2\lambda_3) = 0.2$ $m_{12}(\lambda_1\lambda_3) = 0.1$ $m_{12}(\lambda_1\lambda_2\lambda_3\lambda_4) = 0.1$	$m_{21}(\lambda_1) = 0.1$ $m_{21}(\lambda_3) = 0.6$ $m_{21}(\lambda_1\lambda_3) = 0.1$ $m_{21}(\lambda_2) = 0.1$ $m_{21}(\lambda_1\lambda_2\lambda_3) = 0.1$	$m_{22}(\lambda_1) = 0.2$ $m_{22}(\lambda_3) = 0.7$ $m_{22}(\lambda_1\lambda_2\lambda_3) = 0.1$	$m_{31}(\lambda_1) = 0.8$ $m_{31}(\lambda_1\lambda_2\lambda_3) = 0.2$	$m_{32}(\lambda_1) = 0.6$ $m_{32}(\lambda_3) = 0.2$ $m_{32}(\lambda_1\lambda_3) = 0.2$

The next step of combining these support functions into a single support function over the labeling triplets for the triangle will give support to 90 different elements. Rather than compute all these, let us introduce the prior support function at this stage.

Let us first take $m_P(\cdot)$ to be the simple support function suggested by Shafer in his work on this example giving equal support to the eight possible labelings. This allocates no support to anything other than single labeling triplets (rather than sets of labels for one or more of the sides) and, as a result of joining this with the support functions in Table A-7, the combined support function will be of the same type. The calculations using Dempster's rule on the four support functions, give:

$m_{PD}(\lambda_1,\lambda_1,\lambda_1)=0.119$; $m_{PD}(\lambda_1,\lambda_3,\lambda_1)=0.845$; $m_{PD}(\lambda_1,\lambda_1,\lambda_3)=0.012$; $m_{PD}(\lambda_3,\lambda_1,\lambda_1)=0.024$.

Because of the special structure of this support function, these are, in fact, probabilties for each of the four labelings, and may now be used with a loss function, as suggested by Haralick, 1983, to make a labeling decision.

It will be more interesting, however, to investigate the implications of Shafer's theory when the input support functions give positive support to some combination of simple hypotheses. In particular, suppose $m_P(\cdot)$ gives support of 1 to the set of labelings $\{(\lambda_1,\lambda_1,\lambda_1),(\lambda_2,\lambda_2,\lambda_2),(\lambda_1,\lambda_3,\lambda_1),(\lambda_1,\lambda_1,\lambda_3),(\lambda_3,\lambda_1,\lambda_1),(\lambda_2,\lambda_4,\lambda_2),(\lambda_2,\lambda_2,\lambda_4),(\lambda_4,\lambda_2,\lambda_2)\}$. Thus, instead of supposing, with the Bayesians, that each of the labelings $\underline{l}^{(1)},\ldots,\underline{l}^{(8)}$ is equally likely, we just give all our support to the set of all 8 labelings. This highlights the distinction between the Shaferian and Bayesian representations of lack of knowledge. It is now a tedious, but straightforward matter to compute the final support function, and the associated belief and plausibility functions of the sets of hypotheses (labels).

Table A-8: Computed Belief Functions

Label Set	Support	Belief	Plausibility
$\underline{l}^1$	0.0766	0.0766	0.1311
$\underline{l}^3$	0.8633	0.8633	0.8924
$\underline{l}^4$	0.0066	0.0066	0.0132
$\underline{l}^5$	0	0	0.0261
$(\underline{l}^1,\underline{l}^3)$	0.0221	0.9620	0.9934
$(\underline{l}^1,\underline{l}^4)$	0.0041	0.0873	0.1367
$(\underline{l}^1,\underline{l}^5)$	0.0192	0.0958	0.1301
$(\underline{l}^3,\underline{l}^4)$	0	0.8699	0.9042
$(\underline{l}^4,\underline{l}^5)$	0	0.0066	0.0380
$(\underline{l}^3,\underline{l}^5)$	0	0.8633	0.9127
$(\underline{l}^1,\underline{l}^3,\underline{l}^4)$	0.0012	0.9739	1.0000
$(\underline{l}^1,\underline{l}^4,\underline{l}^5)$	0.0011	0.1076	0.1367
$(\underline{l}^1,\underline{l}^3,\underline{l}^5)$	0.0056	0.9868	0.9934
$(\underline{l}^3,\underline{l}^4,\underline{l}^5)$	0	0.8699	0.9234
$(\underline{l}^1,\underline{l}^3,\underline{l}^4,\underline{l}^5)$	0.0002	1.0000	1.0000

We have not included in the label sets any set of labels which includes a label triplet not in the allowable four ($\underline{l}^1$, $\underline{l}^3$, $\underline{l}^4$ or $\underline{l}^5$). It is clear that $\underline{l}^3$ has the strongest support of any simple labeling; moreover, one sensible procedure for making a conclusion from an analysis of this kind is to adopt the simple labeling with the maximum plausibility. Once again, this is $\underline{l}^3$ in this case.

This analysis does not give us a probability for a hypothesis, but it does lead to (approximate) bounds on that probability, given by Bel(·) and Pl(·). Using these bounds in a loss function calculation might still given an unequivocal labeling decision, or, more likely, will lead to indeterminacy. This may well be the proper output of the labeling procedure, since it corresponds to the inherent indeterminacy in the input information.

We have seen how Shafer's theory may be applied to handle the object labeling problem. It can be a more sensible way of representing what the data tells us, and we recommend the construction of a labeling program, and low-level labeling algorithms, which are consistent with this philosophy.

A.4.8 Fuzzy labeling. In this section we examine the potential of fuzzy set theory for the scene labeling problem. We will first describe in outline the use suggested by Rosenfeld et al. (1976), and give a critique of that use. Then we shall suggest an alternative way that fuzzy measures can illuminate the scene labeling problem.

Rosenfeld et al. start by presuming the existence of an object labeling algorithm which is able to produce for each object i, and each label, λ_k, a number $\mu_i(\lambda_k)$ between 0 and 1. This defines the degree to which it is possible to label object i with label λ_k. They also define a number $\Psi_{ij}(\lambda_k,\lambda_1)$ as the degree to which label λ_k for object i is compatible with label λ_1 for object j; this number is presumed to derive from some discussion of physically possible relationships between objects. As before, in our discussions of the object labeling problem, we see that the task is to combine two types of information, namely, intrinsic information derived from each object about appropriate labels for that object, and more global information about the compatibilities of different combinations of labels for the different objects. In this case, this information is given by $\mu_i(\cdot)$ and $\Psi_{ij}(\cdot,\cdot)$, respectively.

Then a procedure has to be defined to operate on these input numbers to produce a combined opinion about appropriate labelings for the set of objects. Rosenfeld et al. do this in two ways. They are not explicit, but appear to compute, for any labeling $l_1,l_2,\ldots,l_n$, the expression

$$\min_{i,j}(\mu_i(l_i),\Psi_{ij}(l_i,l_j))$$

This represents the degree to which the labeling is compatible both with the data at each object and with the relationships between objects. One could then choose the labeling, $\underline{l}$, for which this expression is largest.

As an alternative, they suggest that a sequence of membership functions should be derived using the relationship

$$\mu_i^{(k+1)}(l_i) = \min_j[\max_{l_j}[\min(\mu_j^{(k)}(l_j),\Psi_{ij}(l_i,l_j))]].$$

This is a kind of relaxation, justified intuitively. The expression in the inner square brackets is the degree to which labels l_i, l_j for objects i and j are possible. The expression in the outer brackets is the degree to which l_i and l_j* are possible, where l_j* is the most plausible label for object j consistent with label l_i for object i. Finally, the overall possibility of the label l_i for object i is the least of these degrees of possibility over all other objects j.

Rosenfeld et al. report that the behavior of this latter algorithm is unsatisfactory when applied to real labeling problems, since degrees of possibility may decrease, but never increase, by using it.

As an alternative to Rosenfeld et al.'s approach, consider the following, which is, in essence, a generalization of their first method. Suppose that instead of representing our knowledge about the consistency of labelings by relationships between pairs of objects, we look at the whole set of objects at once. Thus, instead of $\Psi(\cdot,\cdot)$, we specify $\phi(l_1,l_2,\ldots,l_n)$ to be the extent to which the labels $l_1,\ldots,l_n$ for the objects $1,\ldots,n$, are possible. We then compute the overall possibility of a labeling to be

$$\min(\min_i(\mu_i(l_i)),\phi(l_1,\ldots,l_n)) \qquad (A.5)$$

and we could then adopt the labeling for which this measure is biggest. In the case that

$$\phi(l_1,\ldots,l_n) = \min_{i,j} \Psi_{ij}(l_i,l_j)$$

this reverts to Rosenfeld et al.'s first method. Our method allows greater generality than theirs, however, since we can ask for more general information than the compatibility of pairs: it may be, for example, that label 1 for object 1 is compatible with label 3 for object 6 only if object 7 has label 2; this information cannot be represented in the function $\Psi(\cdot,\cdot)$.

As an example of our approach, consider once again the triangle labeling problem. Suppose that for some image of a triangle, we have the following possibilities:

Table A-9: Input Possibilities (1)

	λ_1	λ_2	λ_3	λ_4
$\mu_1(\cdot)$	1	0.1	0.9	0.2
$\mu_2(\cdot)$	0.7	0.3	0.95	0.6
$\mu_3(\cdot)$	1	0.1	1	0.3

This says that for side 1 labels λ_1 and λ_3 are very possible while labels λ_2 and λ_4 are well-nigh impossible, and so on. Further suppose that the following values of ϕ are given for the labelings $\underline{l}^1$ to $\underline{l}^8$, respectively, using the notation of Table A-2.

1, 0.1, 1, 0.85, 1, 0, 0.1, 0

with zero possibility for all other labelings. Then the values of (A.5) for the eight labelings are, respectively,

0.7, 0.1, 0.95, 0.7, 0.7, 0, 0.1, 0.

Thus the most possible labeling is $\underline{l}^3$. Notice that even if all of the eight labelings were thought to be totally possible ($\phi(\underline{l}^k)=1$, $k=1,\ldots,8$), we would get

0.7, 0.1, 0.95, 0.7, 0.7, 0.1, 0.1, 0.1

from applying (A.5), a barely noticeable difference.

The dependence of the output of this algorithm on the smallest numbers around is intuitively unsatisfactory. Part of the problem may be interpretation of the possibilities as probabilities. In fact, as Zadeh points out, generally speaking possibilities will be bigger than probabilities. A label may be very possible, but improbable. A highly probable label will not be almost impossible. That being so, it may be that more plausible input possibilities may be as below:

Table A-10: Input Possibilities (2)

	λ_1	λ_2	λ_3	λ_4
$\mu_1(\cdot)$	1	0.5	1	0.5
$\mu_2(\cdot)$	1	1	1	1
$\mu_3(\cdot)$	1	0	1	0.8

If we combine this with the total possibility ($\phi=1$) of the eight labelings $\underline{l}^1,\ldots,\underline{l}^8$, using (A.5), we get, respectively,

1, 0, 1, 1, 1, 0, 0.5, 0.

This is not very informative; it excludes three possible labelings ($\underline{l}^2$, $\underline{l}^6$ and $\underline{l}^8$) on the grounds that label λ_2 for side 3 is not possible, and leaves us with the information that four labelings remain totally possible. We suspect that this phenomenon is endemic in uses of fuzzy set theory in this way. We conclude, therefore, as Rosenfeld et al. did, that using fuzzy logic on the scene labeling problem is not likely to be very useful.

REFERENCES

Adams, J.B. Probabilistic reasoning and certainty factors. In Buchanan, B.G., and Shortliffe, E.H. (Eds.), Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project, Reading, MA: Addison-Wesley Publishing Co., 1984, 263-272.

Ballard, D.H., and Brown, C.M. Computer vision. Prentice-Hall, 1982.

Brown, R.V., and Lindley, D.V. Improving judgment by reconciling incoherence. Theory & Decision, 1982, 14, 113-132.

Buchanan, B.G., Barstow, D., Bechtel, R., Bennett, J., Clancey, W., Kulikowski, C., Mitchell, T., and Waterman, D.A. Constructing an expert system. In Hayes-Roth, F., Waterman, D.A., and Lenat, D.B. (Eds.), Building expert systems, Vol. I. Reading, MA: Addison-Wesley Publishing Co., Inc., 1983.

Buchanan, B.G., and Duda, R.O. Principles of rule-based expert systems (Report No. STAN-CS-82-926). Stanford, CA: Stanford University, Aguust 1982.

Buchanan, B.G., and Shortliffe, E.H. Rule-based expert systems. Addison-Wesley, 1984.

Cambier, J.L., Reid, W.J., Barth, S., and Barrett, S.A. Advanced pattern recognition (Technical Report 83-1). PAR Technology Corporation, May 1983. (NTIS AD A132339)

Cheng, A.C., Ledley, R.S., Pollock, D.K., and Rosenfeld, A. (Eds.). Pictorial pattern recognition. Washington, D.C.: Thompson Book Co., 1969.

Chinnis, J.O., Jr., Cohen, M.S., and Bresnick, T.A. Human and computer task allocation in air-defense systems (Technical Report 84-2). Falls Church, VA: Decision Science Consortium, Inc., August 1984.

Cohen, L.J. The probable and the provable. Oxford, England: Clarendon Press, 1977.

Cohen, L.J. Can human irrationality be experimentally demonstrated? The Behavioral & Brain Sciences, 1981, 4(3), 317-330.

Cohen, M.S. Status of the rationality assumption in psychology. The Behavioral & Brain Sciences, 1981, 4(3).

Cohen, M.S., Bromage, R.C., Chinnis, J.O., Jr., Payne, J.W., and Ulvila, J.W. A personalized and prescriptive attack planning decision aid (Technical Report 82-4). Falls Church, VA: Decision Science Consortium, Inc., July 1982.

Cohen, M.S., Mavor, A., and Kidd, J. Research on the elicitation of expert knowledge (Proposal). Falls Church, VA: Decision Science Consortium, Inc., March 1984.

Cohen, P.R., and Feigenbaum, E.A. (Eds.) The handbook of artificial intelligence, Stanford, CA: HeurisTech Press, 1982, Vol. III.

Crombie, M.A., Rand, R.S., and Friend, N. An analysis of the max-min texture measure (Report No. ETL-2080). Fort Belvoir, VA: U.S. Army Corps of Engineers, Engineer Topographic Laboratories, January 1982.

de Dombal, F.T. Surgical diagnosis assisted by computer. Proceedings of the Royal Society, 1973, V-184, 433-440.

de Finetti, B. Foresight: Its logical laws, its subjective sources. English translation in H.E. Kybert, Jr., and H.E. Smokler (Eds.), Studies in subjective probability. New York: Wiley, 1964. (Original: 1937)

DeGroot, M.H. Comment (On Lindley's paradox). Journal of the American Statistical Association, June 1982, 77(378), 336-339.

Doyle, J. A truth maintenance system. Artificial Intelligence, 1979, 12(3), 231-272.

Dubois, D., and Prade, H. Fuzzy logics and the generalized modus porens revisited (Working Paper). Toulouse, France: Laboratoire Langages et Systemes Informatiques, Universite Paul Sabatier, 1984.

Duda, R., Gaschnig, J., and Hart, P. Model design in the PROSPECTOR consultant system for mineral exploration. In D. Michie (Ed.), Expert systems in the microelectronic age, Edinburgh University Press, 1979, 153-167.

Edwards, W. (Ed.). Revisions of opinions by men and man-machine systems. IEEE Transactions on Human Factors in Electronics, 1966, 7(1).

Engelman, C., Berg, C.H., and Bischoff, M. KNOBS: An experimental knowledge based tactical air mission planning system and a rule based aircraft identification simulation facility. Proc. 6th Int. Joint Conf. on A.I., Tokyo, 1979, 247-249.

Feigenbaum, E.A., and McCorduck, P. The fifth generation. Reading, MA: Addison-Wesley Publishing Co., 1983.

Freeling, A.N.S. Fuzzy sets and decision analysis. IEEE Transactions on Systems, Man and Cybernetics, 1980, SMC-10, 341-354.

Freeling, A.N.S., and Sahlin, N. Combining evidence. In P. Gardenfors, B. Hansson, and N. Sahlin (Eds.), Evidentiary value: Philosophical, judicial, and psychological aspects of a theory. Lund, Sweden: C.W.K. Gleerups, 1983.

Glymour, C. Theory and evidence. Princeton, NJ: Princeton University Press, 1980.

Gordon, J., and Shortliffe, E.H. The Dempster-Shafer theory of evidence. In Buchanan, B.G., and Shortliffe, E.H. (Eds.), Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project, 1984, 272-295.

Haralick, R.M. Decision making in context. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1983, PAMI-5, 417-428.

Hayes-Roth, F., Waterman, D.A., and Lenat, D.B. Building expert systems. Reading, MA: Addison-Wesley Publishing Co., Inc., 1983.

Hummel, R.A., and Zucker, S.W. On the foundations of relaxation labelling processes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1983, PAMI-5, 267-287.

Kahneman, D., Slovic, P., and Tversky, A. (Eds.) Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press, 1982.

Kandel, A. Fuzzy techniques in pattern recognition. Wiley, 1982.

Kim, J.H. CONVINCE: A conversational inference consolidation engine (Doctoral Dissertation). Los Angeles, CA: University of California, 1983.

Kittler, J. Image processing for remote sensing. Philosophical Transactions of the Royal Society of London, 1983, A309, 323-369.

Levi, I. Consonance, dissonance and evidentiary mechanisms. In Gardenfors, P., Hansson, B., and Sahlin, N. (Eds.) Evidentiary value: Philosophical, judicial and psychological aspects of a theory, Lund, Sweden: C.W.K. Gleerups, 1983.

Lindley, D.V. Scoring rules and the inevitability of probability. International Statistical Review, 1982, 50, 1-26.

Lindley, D.V., Tversky, A., and Brown, R.V. On the reconciliation of probability assessments. Journal of the Royal Statistical Society, 1979, A-142, 146-180.

Lindsay, R., Buchanan, B.G., Feigenbaum, E.A. and Lederberg, J. Applications of artificial intelligence for organic chemistry: DENDRAL. NY: McGraw Hill, 1980.

Lowrance, J.D., and Garvey, T. Evidential reasoning: Am implementation for multisensor integration (Technical Note 307). SRI International, December 1983.

Mamdani, E.H., and Gaines, B.R. Fuzzy reasoning and its applications. London: Academic Press, 1981.

McCarthy, J. Circumscription--A form of non-monotonic reasoning. Artificial Intelligence, 1980, 13(1,2), 27-39.

McDermott, D. Duck: A lisp-based deductive system. McLean, VA: Smart Systems Technology, May 1983.

McDermott, D., and Doyle, J. Non-monotonic Logic I. Artificial Intelligence, 1980, 13, 41-72.

Pearl, J. Distributed Bayesian belief maintenance. Proceedings of the Second National Conference on Artificial Intelligence. Los Altos, CA: William Kaufmann, Inc., 1982.

Peleg, S. A new probabilistic relaxation scheme. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1980, PAMI-2, 362-369.

Quine, W.V. Two dogmas of empiricism. In Quine, W.V., From a logical point of view. New York: Harper & Row, Inc., 1953.

Reiter, R. A logic for default reasoning. Artificial Intelligence, 1980, 13, 81-132.

Rosenfeld, A. Picture processing: A review. Computer vision, graphics, and image processing, 1983, 22(3), 339-387.

Rosenfeld, A. Image analysis: Problems, progress and prospects. Pattern Recognition, 1984, 17(1), 3-12.

Rosenfeld, A., Hummel, R.A., and Zucker, S.W. Scene labelling by relaxation operations. IEEE Transactions on Systems, Man and Cybernetics, 1976, SMC-6, 420-433.

Schum, D.A. A review of a case against Blaise Pascal and his heirs. Michigan Law Review, Jan-Mar, 1979, 77(3), 446-484.

Schum, D.A. Current developments in research on cascaded inference processes. Chapter 10 of T.S. Wallsten (Ed.), Cognitive process in choice and decision behavior. Hillsdale, NJ: Lawrence Erlbaum Associates, 1980.

Schum, D.A. Sorting out the effects of witness sensitivity and response criterion placement upon the inferential value of testimonial evidence. Organizational Behavior and Human Performance, 1981, 2.

Schum, D.A. and Martin, A.W. Probabilistic opinion revision on the basis of evidence at trial: A Baconian or a Pascalian process? (Report 80-02). Houston, TX: Rice University, 1980.

Shafer, G. A mathematical theory of evidence. Princeton, NJ: Princeton University Press, 1976.

Shafer, G. Jeffrey's rule of conditioning. Phil. of Sci., 1981, 48, 337-362.

Shafer, G. Lindley's paradox. Journal of the American Statistical Association, June 1978, 77(378), 325-351.

Shafer, G. Probability judgment in artificial intelligence and expert systems. Lawrence, KS: University of Kansas, School of Business, December 1984. (a)

Shafer, G. The problem of dependent evidence (Working Paper No. 164). Kansas: University of Kansas, School of Business, 1984. (b)

Shafer, G. Belief functions and possibility measures. Lawrence, KS: University of Kansas, School of Business, in press.

Shafer, G., and Tversky, A. Weighing evidence: The design and comparison of probability thought experiments. Stanford, CA: Stanford University, June 1983.

Shimony, A. Scientific inference. In Colodny, R.G. (Ed.), The nature & function of scientific theories. Pittsburgh, PA: University of Pittsburgh Press, 1970.

Shortliffe, E.H. Computer based medical consultation: MYCIN. Elsevier, 1976.

Slovic, P., and Tversky, A., Who accepts Savage's axiom? Behavioral Science, 1974, 19, 368-373.

Stallman, R.M., and Sussman, G.J. Problem solving about electrical circuits. In Proceedings of the Fifth International Joint Conference on Artificial Intelligence, August 22-25, 1977, Cambridge, MA, pp. 299-304.

Watson, S.R., Weiss, J.J., and Donell, M.L. Fuzzy decision analysis. IEEE Transactions of Systems, Man and Cybernetics, 1979, SMC-9, 1-9.

Williams, P.M. On a new theory of epistemic probability. The British Journal for the Philosophy of Science, 1978, 29, 375-387.

Yu, V.L., Fagan, L.M., Bennett, S.W., Clancey, W.J., Scott, A.C., Hannigan, J.F., Buchanan, B.G., and Cohen, S.M. An evaluation of MYCIN's advice. In Buchanan, B.G., and Shortliffe, E.H. (Eds.), *Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project*, Reading, MA: Addison-Wesley Publishing Co., 1984, 589-599.

Zadeh, L.A. Fuzzy sets. *Inf. and Contr*., 1965, *8*, 338-353.

Zadeh, L.A. The concept of a linguistic variable and its application to approximate reasoning. *Information Science*, 1975, *8*, 199-249; 301-357; *9*, 43-80.

Zadeh, L.A. *Fuzzy probabilities and their role in decision analysis* (Technical Report). Berkeley, CA: University of California, Computer Science Division, 1981.

Zadeh, L.A. Possibility theory and soft data analysis. In Cobb, L., and Thrall, R.M. (Eds.), *Mathematical frontiers of the social and policy sciences*. AAAS: Washington, 1981, 69-129.

Zadeh, L.A. The role of fuzzy logic in the management of uncertainty in expert systems. *Fuzzy Sets and Systems*, 1983, *11*, 199-227.

Zadeh, L.A. Making computers think like people. *IEEE Spectrum*, August 1984. (a)

Zadeh, L.A. Review of Shafer's A Mathematical Theory of Evidence. *AI Magazine*, 1984(b), *5*(3), 81-83.