COMPARING THE VALIDITY OF ALTERNATIVE BELIEF LANGUAGES: AN EXPERIMENTAL APPROACH

Shimon Schocken
Department of Information Systems
Leonard N. Stern School of Business
New York University
44 West 4th Street, Room 9-80
New York, NY 10012-1126
(212) 998-0841
E-mail: sschocke@stern.nyu.edu

Revised August 1991

Replaces #IS-88-94

Working Paper Series
Stern #IS-91-31

Abstract

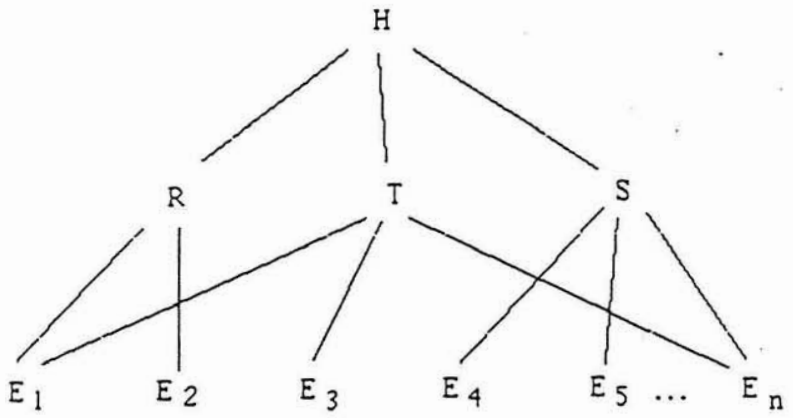
The problem of modeling uncertainty and inexact reasoning in rule-based expert systems is challenging on normative as well on cognitive grounds. First, the modular structure of the rule-based architecture does not lend itself to standard Bayesian inference techniques. Second, there is no consensus on how to model human (expert) judgement under uncertainty. These factors have led to a proliferation of quasi-probabilistic belief calculi which are widely-used in practice. This paper investigates the descriptive and external validity of three well-known "belief languages:" the Bayesian, ad-hoc Bayesian, and the certainty factors languages. These models are implemented in many commercial expert system shells, and their validity is clearly an important issue for users and designers of expert systems. The methodology consists of a controlled, within-subject experiment designed to measure the relative performance of alternative belief languages. The experiment pits the judgement of human experts with the recommendations generated by their simulated expert systems, each using a different belief language. Special emphasis is given to the general issues of validating belief languages and expert systems at large.

1. Rule-Based Belief Languages

Consider the following familiar problem: a faculty recruiting committee attempts to estimate the academic potential of a candidate for a junior faculty position, based on the resume and recommendation letters submitted by the candidate. We will assume henceforth that the candidate's profile can be credibly encoded through a set of attributes, e.g. "has an MBA degree," "is a foreign citizen," etc. Formally speaking, the recruiter's task can be described as one of classifying a set of instantiated attributes (representing a particular candidate) into the two categories "academic success" or "academic failure." This inexact classification can be made continuous by assigning degrees of likelihood to the two alternative hypotheses. Under these assumptions, the recruiter's task can be formalized using such models as utility theory, bootstrapping, or the Analytic Hierarchy Process. In this paper, though, we wish to cast the faculty selection problem (which is just an example) in what is termed in artificial intelligence a "rule-based" framework. The resulting model is rather subjective; it is based on our own set of values and experience regarding the selection and promotion of prospective faculty members.

We begin by describing our recruiting rationale (rule-base) and the candidate's profile (fact-base) in terms of hypotheses, pieces of evidence, and rules. Our ultimate goal is to evaluate the likelihood of an "academic success" hypothesis (H). We break

down this overall hypothesis into the three sub-hypotheses "research ability (R)," "teaching ability (T)," and "service potential (S)." These sub-hypotheses, in turn, can be linked to the set of attributes drawn from the candidate's information. For example, if we think that "MBA degree" (E1) is related to "teaching ability" (T), we connect these two propositions by the rule <if E1 then T with degree of belief Bel(T,E1)>. The resulting rule-base can be pictorially presented as an inference network, like the one depicted in Picture 1.



Picture 1

The meaning of the belief function Bel(.|.) depends on our choice of a belief language. According to Shafer and Tversky (1985), a belief language consists of syntax, calculus, and semantics. The syntax corresponds to the set of degrees of belief associated with various rules (arcs) and propositions (nodes). Typically, a set of atomic degrees of belief is

elicited from a domain expert (e.g. an experienced recruiter), while posterior beliefs in hypotheses are computed by a set of operators collectively known as a belief calculus. When a belief language is used in the context of an expert system, the computed posterior beliefs influence, if not determine, the system's judgement. Therefore, the semantics of the belief language can be viewed as a measure of the system's validity. Descriptive validity concerns the proximity of the system's recommendations to actual human judgement. External validity concerns the consistency of the system's judgments with the actual state of the world. More about that, later.

The validity of alternative belief languages is an interesting question on theoretical as well as on practical grounds. Belief languages are typically built into expert system shells, the canned programs used to develop applied expert systems. Thus, the credibility of these languages sheds light on the integrity of expert systems at large. This paper investigates the descriptive and external validity of three well-known belief languages: the Bayesian, ad-hoc Bayesian, and the certainty factors languages. These languages are widely-used in practice, and versions of them are employed in most commercial expert systems. We now turn to present a brief review of the three languages, without getting into unnecessary technical clutter. Detailed descriptions of these models can be found elsewhere, and the reader will be directed to these references as we go along.

The normative Bayesian language gives an inference network like the one depicted in Figure 1 a probabilistic interpretation. The network's nodes are viewed as a set of random variables which are causally interrelated. In a medical diagnosis example, the hypothesis H (a disease) might be viewed as the cause of the syndromes T, R, and S. These syndromes, in turn, may manifest themselves through the pieces of evidence (symptoms) {Ei}. The strength of these causal associations may be measured through conditional probabilities or likelihoos ratios. For example, the degree of belief Bel(H,R) may be captured through P(R|H) or P(R|H)/P(R|not H), P being a probability. Going back to the faculty selection problem, suppose a particular candidate can be encoded through the instantiated attributes set E=<E1,...,En>. Given this terminology, evaluating the academic potential of the candidate E amounts to computing the posterior belief in the hypothesis H in light of E, P(H|E). Unfortunately, this computation is exponential in the number of nodes in the network, and, in fact, is NP-hard (Cooper, 1987).

If, however, the underlying joint-distribution function obeys a set of conditional independence assumptions, there exist efficient Bayesian algorithms that compute P(H|E) in time linear to the size of the network (e.g. Pearl, 1986). These algorithms hinge on the topology of the network, which, in turn, dictates the set of conditional independence assumptions which P is

assumed to possess. If these assumptions don't hold, one can sometimes restructure the network in order to enforce them (Charniak, 1983).

To sum up, given that the rule-base's structure is consistent with a set of simplifying assumptions on P, there exist Bayesian belief-update algorithms which do not violate the axioms of probability theory. In contrast, the majority of the languages employed by so-called "Bayesian" expert systems like PROSPECTOR or AL/X are quasi-probabilistic. Like the normative case, the syntax of these languages consists of a set of conditional probabilities or likelihood-ratios, elicited from a domain expert. At the same time, the calculus of the Ad-Hoc Bayesian (AHB) language is basically a heuristic version of Bayes rule, designed to "adjust" the computation of probabilities to the deductive nature of rule-based inference. This is done by introducing "parallel" and "sequential" combination functions which prune the inference net recursively until a set of posterior beliefs is computed (Duda et al, 1977).

To illustrate, consider the application of the AHB calculus to the inference net depicted in Picture 1. The process begins by applying the parallel combination function (which is basically Bayes rule under the assumption of conditional independence) three times to compute the posterior beliefs in the sub-hypotheses "research," "teaching," and "service." These

posterior beliefs, in turn, serve to "attenuate" the original degrees of belief rendered by the three sub-hypotheses to the "academic success" hypothesis (H). This attenuation is carried out by the sequential combination function. Finally, the attenuated degrees of belief are combined by the parallel combination function, yielding the overall posterior belief in the root hypothesis, H.

This ad-hoc calculus is quite similar to the one employed by the Certainty Factors (CF) language. This language was first implemented in the MYCIN expert system (Shortliffe, 1976) and was subsequently incorporated in the EMYCIN and M1 (van Melle, 1984) expert system shells. In the CF terminology, the degree of belief associated with the diagnostic rule <if E then H> is the certainty factor CF(H|E). CF(H|E), which is elicited from a domain expert, measures the increased belief (or disbelief) in H in light of the piece of evidence E. The CF function, though, is not a probability. It varies from -1 to 1, corresponding to "E confirms not H with certainty" and "E confirms H with certainty," respectively. If E is irrelevant to H, the certainty factor CF(H|E) is set to 0. In sum, CF(H|E) measures the strength of the logical entailment E->H, in the spirit of Carnap's (1954) confirmation function and inductive logic. Atomic CF's are elicited from domain experts. If a single hypothesis is backed by several rules, its posterior CF is computed by the CF parallel combination function. If an hypothesis H is backed by

a reasoning chain, say, E->S->H, its posterior belief is computed by the CF sequential combination function. These functions are described in detail in (Buchanan and Shortliffe, 1984).

The mathematical properties (and limitations) of the AHB and the CF languages are now well understood, and the reader is referred to Adams (1984), Grosof (1986), Heckerman (1986), Horvitz et al (1986), and Schocken and Kleindorfer (1987) for detailed analyses. What emerges from this research is that the AHB and the CF languages are essentially special cases of the Bayesian language, involving implicit assumptions of conditional independence. Moreover, these language are mathematically isomorphic to each other. At the same time, this normative proximity does not necessarily guarantee compatibility on other, ex-mathematical grounds. In fact, the Bayesian, AHB, and CF languages involve different cognitive views of inference under uncertainty, different elicitation procedures, and, perhaps, different or incompatible posterior beliefs. To emphasize this point, suppose we replace the AHB language employed by a rule-based medical diagnosis system with a CF language. If all other things are held equal, including the rule-base and the patient, it is still possible that the system will switch its prognosis from one disease to another. Clearly, this potential blunder requires serious investigation: it implies that at least one of the languages under consideration must be invalid.

The descriptive and external validity of a belief language can be tested only in a controlled experiment involving a realistic inference problem and human experts. The posterior beliefs assigned by the language to various hypotheses of interest can be then compared to either (a) a set of likelihoods assigned by a human expert, or, (b) a set of probabilities generated by a monte-carlo simulation. Such "within-language" experiments were carried out by Yu et al (1984) and Yadrick et al (1988), respectively. Alternatively, one can apply several belief languages to the same inference problem, comparing their resulting recommendations to each other. Such "across-languages" experiments were undertaken by Mitchell (1986), Wise (1988), and Kopsco et al (1988).

This paper belongs to the latter category of comparative studies. It involves the application of the Bayesian, AHB, and CF languages to the same inference task, viz, the faculty selection problem. The structure of the remainder of the paper is as follows: Section 2 consists of an example of a simple inference problem designed to illustrate our approach to measuring the descriptive and external validity of competing belief languages. This discussion sets the stage for the experimental design, described in Section 3. Our research hypotheses and results are given in Section 4. A discussion and conclusion sections highlight the key findings.

2. Pitting Human and Machine Judgement

Let E1 and E2 be two independent pieces of evidence supporting a prospective hypothesis, H. Let P be the joint distribution function defined over the space {E1,E2,H}, and consider the following likelihood-ratio notation:

$$L(H) = P(H)/P(\underline{H}) \qquad (\underline{H} \text{ hereafter stands for "not H"})$$

$$R(Ei|H) = P(Ei|H)/P(Ei|\underline{H})$$

$$L(H|E1,E2) = P(H|E1,E2)/P(\underline{H}|E1,E2)$$

In the likelihood-ratio paradigm, L(H) is the prior belief in H, and L(H|E1,E2) is the posterior belief in H in light of the evidence {E1,E2}. R(Ei|H) is the degree of belief in the "symptom" Ei occurring when H is known to be true. Such degrees of belief can be obtained from past records, textbook information, and expert opinions. The question of belief-update, simply put, is this: given L(H), a certain body of evidence {E1,E2}, and a set of degrees of belief {R(E1|H),R(E2|H)}, how does one go about computing the posterior belief L(H|E1,E2)?

Under the assumption that E1 and E2 are ratio-independent with respect to H (Grosof, 1986), the normative posterior belief, denoted L_T, may be derived from Bayes rule, as follows:

$$L_T(H|E1,E2) = L(H) * R(E1|H) * R(E2|H) \qquad (1)$$

If the R(Ei|H) are elicited from human experts, which is normally the case, we must replace them with their estimates, R'(Ei|H). Moreover, it is well known by now that human judgement under uncertainty does not conform to (1). When left to their own devices, people's judgement under uncertainty is prone to a number of systematic biases, e.g. representativeness (Tversky and Kahnemna, 1974). For example, if the symptom E1 is very representative of H, most humans will unduly overweight its diagnostic impact on the likelihood of H. This judgmental bias might be represented in the following descriptive model:

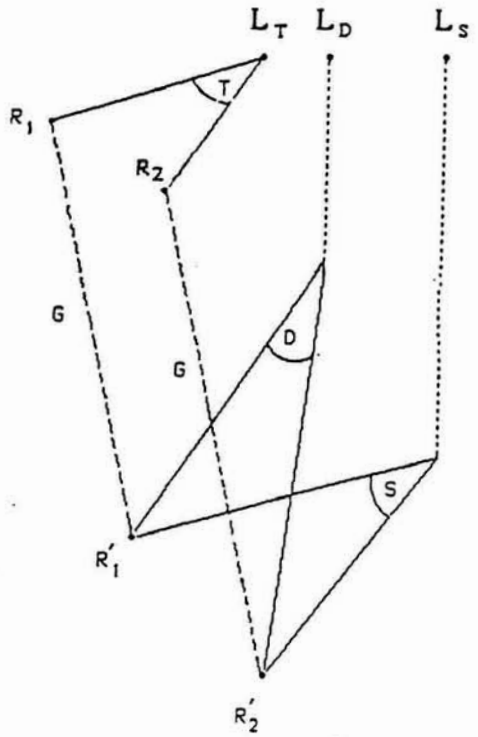
$$L_D(h|E1,E2) = L(H)^{\alpha} * R'(E1|H)^{\beta} * R'(E2|H)^{\gamma} \quad (2)$$

The representativeness heuristic is modeled in (2) through the parameters $\{ \alpha , \beta , \gamma \}$. Any assignment of values other than $\alpha = \beta = \gamma = 1$ consists of a violation of Bayes rule. Interestingly, it can be shown that parameterized versions of (2) are isomorphic to many non-Bayesian belief languages, e.g. the certainty factors and the contrast-inertia languages (Einhorn and Hogarth, 1987).

One obvious way to debias human judgement is to design a "Bayesian machine" like (1) whose inputs consist of elicited degrees of belief. This prescriptive approach can be modeled as follows:

$$L_S(h|E1,E2) = L(H) * R'(E1|H) * R'(E2|H) \qquad (3)$$

in order to investigate the comparative validity of (1), (2), and (3), we must assume that the three models are based on the same human expert working on a fixed inference problem. This is emphasized in Picture 2. The three triangles T, D, and S correspond to the belief calculi (1), (2), and (3), respectively. Each of these models takes as input the body of evidence E={E1,E2} and goes on to compute the posterior belief in the hypothesis H in light of E. The notation R_i is an abbreviation of R(Ei|H). The T triangle represents the normative belief-update model (1) which generates the true posterior belief, L_T. G is an elicitation operator replacing true degrees of belief, R_i, with their human-provided estimates, R_i'. The D triangle represents the expert's own, abstract, decision process (model (2)): when presented with the body of evidence, E, the expert sets his posterior belief in H to L_D, which may or may not coincide with the true posterior belief, L_T. The S triangle represents model (3), i.e. a Bayesian "inference engine" operating on human inputs.



Picture 2

The preceding discussion and Picture 2 give rise to the following definitions of three relevant performance measures:

$|L_D - L_T|$: expert's external validity

$|L_S - L_T|$: system's external validity

$|L_S - L_D|$: system's descriptive validity

The distinction between expert's and system's external validity was illustrated dramatically in the MYCIN experiments. In order to test the plausibility of MYCIN's therapeutic advice, the program was fed with diagnostic information regarding ten

patients with infectious meningitis (Yu et al, 1984). MYCIN's recommended therapy (analog of L_S) was then evaluated by a panel of leading medical experts whose opinions were taken to represent the truth (analog of L_T). The panel judged MYCIN to be correct 75% of the time. Viewed in isolation, this measure of system's external validity appeared to be rather discouraging. However, these same cases were then presented to a group of eight Stanford physicians. Surprisingly, the recommended therapy of the human experts (analog of L_D) received an external validity rating which was uniformally lower than 75% in view of the same panel of experts who evaluated MYCIN.

In general, the external validity of an expert system might exceed that of its underlying expert. Indeed, the management science literature is rife with examples in which computer-based models, e.g. linear models, have been known to systematically outperform human experts (Dawes and Corrigan, 1974). Incidentally, these mechanical variance-minimizing models have no descriptive appeal whatsoever. This also gives rise to the argument that, unlike other efforts in artificial intelligence, descriptive validity is not necessarily a good measure of expert system performance: "Evidence about human reasoning including introspection may give us excellent ideas for devising new and better systems, but the criteria for judging their usefulness should be the quality of their performance, rather than how well they simulate human thought processes (Henrion, 1986).

Finally, note that the expert's external validity is an intrinsic property of the expert. At the same time, we expect that the external and descriptive validity of an expert system will vary with our choice of a belief language. In Picture 2, there is only one such prescriptive language, represented by the S triangle. More such languages may be considered, with the provision that all languages draw on the same expert who is working on a fixed inference problem. These different languages are likely to yield different posterior beliefs. By comparing these beliefs to L_T and L_D (which are fixed), we can make statements about the relative validity of the underlying languages. Ideally, these statements should withstand the test of statistical significance. This is the crux of our experiment.

3. Experimental Design

The experimental task involved the faculty selection problem discussed in the beginning of the paper. An opening of a tenure-track faculty position in a major university typically attracts dozens of candidates. Each candidate submits a resume and recommendation letters, which are then scrutinized by a recruiting committee. The committee has to decide which candidates should be invited to on-site interviews. This task is normally carried out through some sort of a "phased" strategy consisting of screening and ranking (Bettman, 1979). First, inferior candidates are eliminated sequentially from

consideration. The remaining candidates are then compared to each other in a more holistic sense. For example, the first phase might employ an elimination by aspects strategy, in which a criterion is chosen, e.g. "research interests," and all the candidates who do not measure up are rejected. This process is repeated with additional criteria, until a smaller but more focused pool of candidates remains for further consideration. The process terminates when candidates can no longer be evaluated on the basis of a single criterion. At that stage, the decision maker resorts to a compensatory, holistic strategy which considers several attributes simultaneously. This latter stage is the general context in which our experiment took place.

The subjects in the experiment were 12 senior Ph.D. students and 3 professors at a decision sciences department. Each subject was randomly assigned to one of two groups, Group I and Group II. The experiment consisted of three stages, as follows:

Human Rankings: The subject was given a set of ten resumes of hypothetical candidates who presumably applied for a job at the subject's decision sciences department. The subject was told that the experiment evolves around determining the potential academic success of these candidates. A measure of "academic success" was explicitly defined as the answer to the following question: what is the likelihood that a particular candidate will be offered tenure in the decision sciences department within 6

years from his or her first appointment, given all the information that can be extracted from the candidate's resume?

The subject was then asked to rank-order the ten candidates in decreasing order of academic success. The resulting ranking, termed "human ranking," is denoted L_{H1}. The subject performed this ranking without the interruption or assistance of any formal model, although he or she was allowed to use paper, pencil, and a calculator.

The subject was then told that a rule-based inference system based on his individual preferences will now be constructed, and that the system-generated ranking of the ten candidates will be compared to his original human ranking. A financial compensation was offered as follows: the subject received a flat $5 participation fee, plus a performance bonus which was proportional to the correlation found between the subject's ranking and the system's ranking. This bonus ranged from $1 to $50 for the worst and best correlation detected in the experiment, respectively.

Knowledge Elicitation: Following the human ranking, each subject underwent an elaborate knowledge elicitation procedure administered by the experimenter, who played the role of a knowledge engineer. First, the general principles of rule-based inference were presented to the subject, who was allowed to ask

questions and receive further clarifications. The rule-base depicted in Picture 1 was offered as a point of departure toward the subject's specific rule-base. First, the experimenter suggested that the information extracted from the candidates' resumes could be encoded through a set of attributes. The subject was then encouraged to refine this set by deleting irrelevant attributes or adding new ones which he or she perceived important. The refined set amounted to the bottom tier of the inference net depicted in Picture 1. The subject was then asked to connect all the nodes in the network which he thought were causally related to each other. The topology and contents of the resulting network (rule-base) varied widely across subjects.

Next, the experimenter proceeded to elicit the degrees of belief associated with each rule in the subject's rule-base. Each group of subjects received a different belief language "treatment:" subjects in Group I and II were asked to express their degrees of belief using the language of conditional probabilities and certainty factors, respectively. For example, let's assume that a subject thought that "consulting experience" is relevant to "teaching ability." If the subject belonged to Group I, he or she was posed with the following pair of causal questions:

Assume that x is a good teacher.

What is your belief, as a subjective probability, that x has consulting experience?

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+---+
0    .1    .2    .3    .4    .5    .6    .7    .8    .9   1
```

Assume that x is not a good teacher.

What is your belief, as a subjective probability, that x has consulting experience?

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+---+
0    .1    .2    .3    .4    .5    .6    .7    .8    .9   1
```

If the subject belonged to Group II, he or she was asked to answer the following pair of <u>diagnostic</u> questions:

Assume that x has consulting experience.

To what extent does this fact increase (or decrease) your belief that x will become a good teacher?

```
 +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
-1 .9 .8 .7 .6 .5 .4 .3 .2 .1  0 .1 .2 .3 .4 .5 .6 .7 .8 .9 +1
```

Assume that x has no consulting experience.

To what extent does this fact increase (or decrease) your belief that x will become a good teacher?

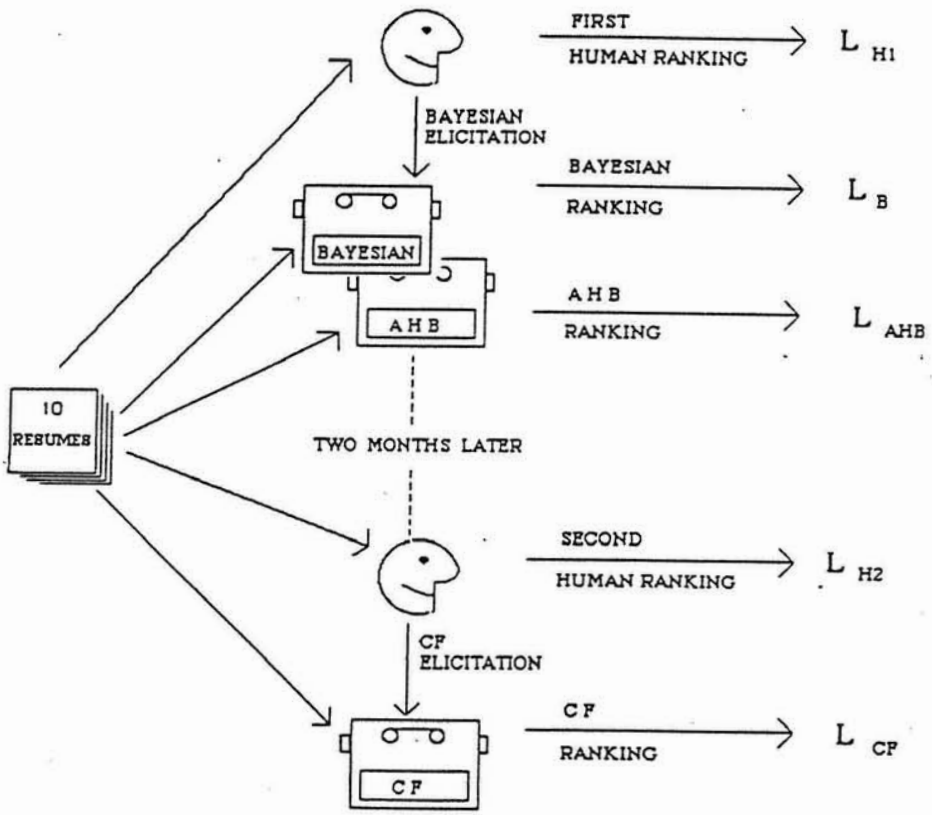
```
 +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
-1 .9 .8 .7 .6 .5 .4 .3 .2 .1  0 .1 .2 .3 .4 .5 .6 .7 .8 .9 +1
```

A series of similar questions then ensued, one pair for each rule in the subject's rule-base. This completed the construction of a rule-base which presumably captured the ranking rationale of the human expert (subject).

Two months later, the subjects were recalled, and the very same sequel ensued: first, the subject was given the resumes of the very same ten candidates that he or she has evaluated two months earlier. Each subject then generated a second ranking of the candidates, denoted L_{H2}. Next, each subject was presented with exactly the same rule-base that he or she has constructed originally, with one exception: the degrees of belief which parameterized the rules were omitted. Finally, the elicitation treatment was switched: subjects from group I were asked to express their degrees of belief in each rule using the certainty factors language, while group II subjects expressed their belief in terms of conditional probabilities. This completed the subject's participation in the experiment.

Machine rankings: Using the inputs provided by the subjects, three expert systems (per subject) were constructed. These systems were implemented through a Prolog-based inference engine designed specifically to take a belief language as an external parameter (Schocken and Finin, 1987). The three systems, which operated on the same rule-base, varied only in their dependance on a CF, Bayesian, and ad-hoc Bayesian belief calculi. The three systems were then fed with the ten encoded resumes, and went on to generate three candidate rankings in terms of posterior probabilities, ad-hoc posterior probabilities, and certainty factors. These rankings are denoted L_B, L_{AHB}, and L_{CF},

respectively. Note that the three systems were identical in their reliance on the same rule-base, fact-base, and expert. These factors were tightly controlled, varying only the belief calculus "treatment."



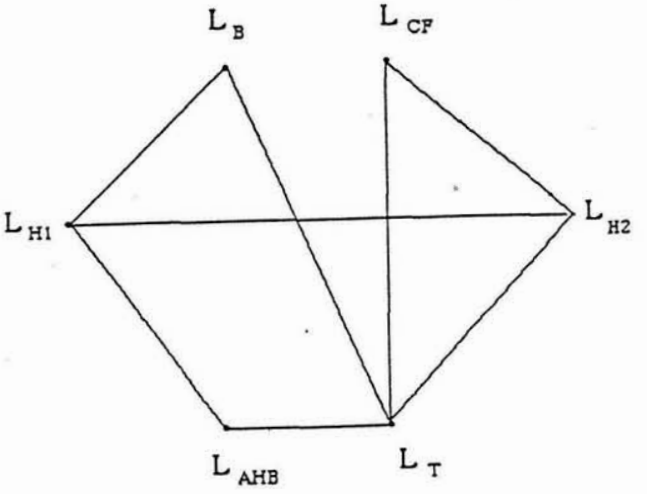
Picture 3

Picture 3 is a summary of the experiment, describing the various stages undertaken by Group I subjects (for Group II subjects, replace the order of the Bayesian and the CF treatments). Altogether, each subject generated two direct human rankings (L_{H1}

and L_{H2}) and three indirect machine-based rankings (L_B, L_{AHB}, L_{CF}). The correlations among these rankings were used as measures of subject's reliability and system's descriptive and external validity. These measures are discussed in the next section.

4. Hypotheses and Results

Picture 4 depicts the various rankings and correlations investigated in the experiment. Each node (excluding L_T, which will be discussed shortly) represents a ranking of the ten resumes, generated either by a human subject or by an expert system simulating the same subject. L_{H1} and L_{H2} are the two human rankings, spaced two months apart. The remaining three nodes correspond to the rankings generated by the three expert systems. All three systems drew on the same rule-base of the subject that generated L_{H1} and L_{H2}, and all operated on the same fact-base (i.e. the ten resumes).



Picture 4

(This figure depicts the comparisons made within Group I (Bayes first, then CF). For Group II comparisons (CF first, then Bayes), replace L_{H1} with L_{H2}).

L_T represents the "true," ex-post ranking of the ten resumes. That is, assuming that we have a crystal ball, L_T gives the actual, rather than the predicted academic performance of the ten candidates. In the context of medical diagnosis, L_T would represent the actual physical state of the patient, which, unfortunately, can often be determined only through a radical procedure such as autopsy or surgery. The need for a "gold standard" and the realization that such standard is not necessarily subject to observation was addressed by Buchanan and Shortliffe (1984): "In general there are two views of how to define a gold standard for an expert system's domain: (1) what eventually turns out to be the "correct" answer for a problem, and (2) what a human expert says is the correct answer when presented with the same information as is available to the program. It is unfortunate that for many kinds of problems with which expert systems are designed to assist, the first of these questions cannot be answered or is irrelevant."

In the MYCIN experiments, absolute truth was approximated by the opinion of a distinguished panel of internists. A similar approach was taken here. The true ranking of the ten candidates, L_T, was estimated by the pooled human rankings of the three subjects in the experiment who were professors in the decision

sciences department. These professors are actively involved in faculty recruiting and promotion decisions; In reality, they (among other professors) would be responsible for judging the actual academic performance of the ten candidates, had they been hired. Thus, given the experiment's context, this pooled ranking seemed to be a reasonable approximation of the actual tenure prospects of the ten hypothetical candidates.

The arcs in Picture 4 represent estimates of reliability and validity. In particular, the arc (L_i, L_j) represents the Spearman rank-correlation coefficient $R(L_i, L_j)$. A subject is said to be reliable if $R(L_{H1}, L_{H2})$ is close to 1, indicating that the subject did not change his preferences over a period of two months. This hypothesis was confirmed informally by inspection: in 80% of the subjects, R was greater than 0.794, the critical value above which the population correlation coefficient is significant at the 0.005 level. All the subjects had significant reliability coefficients at the 0.10 level.

Within a particular (group I) subject, the Bayesian language is said to exhibit a higher descriptive validity than the CF language if $R(L_B, L_{H1})$ is significantly greater than $R(L_{CF}, L_{H2})$ (for Group II subjects, replace L_{H1} and L_{H2}). Of course, this statement gains extra strength if the subject is highly reliable. Similarly, within a particular subject, the Bayesian language is said to exhibit a higher external validity than the CF language

if $R(L_B,L_T)>R(L_{CF},L_T)$. Similar hypotheses can be formulated regarding the AHB language. Finally, the Bayesian model is said to outperform its corresponding human subject (Group I) if $R(L_B,L_T)>R(L_{H1},L_T)$. Similar statements can be made with regard to the CF and the AHB languages.

The above hypotheses are all within-subject. In order to test their significance across the 15 subjects, a sign test was administered. For example, we say that the Bayesian language outperformed the CF language in terms of descriptive validity if the relationship $R(L_B,L_{H1})>R(L_{CF},L_{H2})$ was found to be significant in a sign-test applied to the 15 pairs $<R_i(L_B,L_{H1}),R_i(L_{CF},L_{H2})>$, i=1,15, i being the subjects index.

The results of the sign tests are given in Table 1. The notation BAYES>>CF stands for "the Bayesian language outperformed the CF language" in the category indicated by the column's heading. For example, the first entry reads as follows: in terms of descriptive validity, the Bayesian model outperformed the CF model in 60% of the subjects. The remainder of the entries read similarly.

Hypothesis	Descriptive Validity	External Validity
BAYES>>CF	60%	73% (*)
AHB>>CF	53%	80% (**)
AHB>>BAYES	43%	57%

(*): significant at p=0.059 (single-tailed)
(**): significant at p=0.017 (single-tailed)

Table 1

We see that the Bayesian and the AHB languages outperformed the CF language in terms of descriptive as well as external validity. However, only the external validity results are statistically significant. No statistical difference was detected between the Bayesian and the AHB languages. A more elaborate discussion of the experiment's results is given in the next section.

Our choice of a non-parametric test was based on a reluctance to make any assumptions regarding the underlying distribution of the experimental observations, i.e. the calculated rank-correlation coefficients. In addition, it was felt that although these coefficients were suspectable to ordering, their absolute values were quite meaningless. Under such circumstances, the sign-test is a powerful device in detecting populations differences. In the case of n=15 observations, α =0.1, and an expected large size effect, the power of a bidirectional

sign-test is 0.69, meaning that there are 69% chances of detecting a population difference, if such exists. According to Cohen (1965, p. 98), this power is consistent with the convention that Type I errors be guarded against about four times as stringently as Type II errors.

5. Discussion

Any empirical study involving the elicitation of subjective degrees of belief is prone to the problems of unreliability and inconsistency. As Fischhoff, Slovic, and Lichtenstein (1980) indicate, it would be inappropriate to think of a person's opinion about a set of events as existing within that person in a precise, fixed fashion, just waiting to be measured. Unreliable humans exhibit temporal changes in their beliefs with no apparent reason. Consequently, reliability can be measured in terms of correlations between two different encodings of the same set of events by the same subject at different times. Needless to say, this test is based on the premise that the subject did not have any (non-noise) reason to change his preferences and beliefs over this time period.

Inconsistency occurs when elicited degrees of belief do not respect the axioms (or "grammar") of the underlying language. One major source of unreliability and inconsistency is low motivation: subjects are often eager to "get done" with the experiment, and, as a result, the input that they provide does

not necessarily reflect their true preferences and beliefs. This attitude is quite distracting when students assume the role of domain experts. While the former are typically indifferent to the task at hand, the latter are highly motivated and genuinely concerned about the accuracy and validity of their inferential procedures.

Subject Reliability: Subject's reliability was controlled by asking the subjects to rank-order the same set of ten resumes twice, with the second ranking being done two months after the first. From a cognitive standpoint, the faculty selection problem was rather challenging: the academic credentials of the ten candidates were arranged in such a way that there were neither clear cut winners nor downright losers. Thèrefore, the subjects had to deal with a multi-attribute choice problem with no dominating alternatives. During the human rankings stage, the subjects employed a variety of heuristic compensatory decision rules as well as sheer intuitive judgment. The quickest and slowest (or most diligent) subjects required 30 and 85 minutes to complete the ranking, respectively. Average completion time was 55 minutes.

With that in mind, it was rather encouraging to find that the two human rankings of most of the subjects were highly correlated. Although no subject succeeded to reproduce his first ranking entirely, 80% of the subjects exhibited a highly

significant correlation (R>0.745) between the two human rankings. This high degree of reliability might be attributed to two factors. First, the subjects were promised a substantial financial award (as much as $55) that was proportional to the correlations found in the experiment. This also generated a positive contest spirit that motivated the subjects to outperform their peers.

Second, the experiment's context, which was directly related to the subjects' career interests, proved to be rather lively. Any senior Ph.D. student has a strong position regarding the relative importance of teaching, research, and service in promoting professors. The experiment gave the subjects an opportunity to formulate these preferences and express them in a systematic fashion. This provoked some interesting responses. For example, one subject argued that evidence of good teaching skills discounts the likelihood of a promotion, as devoted teachers are likely to spend less time on research, which is far more important in tenure decisions. Indeed, the three machine-based rankings of this subject tended to favor candidates with no teaching credentials, as did his two human rankings.

Subject Consistency: Consistency checks may be introduced in order to adjust subject's response to a certain standard. For example, a Bayesian elicitation procedure might force subjects to revise their judgment once a violation of the axioms of

subjective probabilities is detected. This practice was avoided here for two reasons. First, the merit of forced normalizations is still under debate. For example, Robinson and Hastie (1985) demonstrated that when subjects were forced to normalize their responses, the responses became more error-prone. Second, it is unclear how to perform consistency checks on certainty factors, short of translating them to probabilities and normalizing their Bayesian images. Consequently, any attempt to employ a language-dependant normalization scheme would introduce an unfair advantage or disadvantage to this particular "treatment." Since the experiment was concerned primarily with the relative, rather than the absolute, performance of various belief languages, it was felt that such practice should be avoided.

Descriptive validity: There is by now an overwhelming body of psychological evidence indicating that human judgment under uncertainty is not governed by, and often violates, the axioms of subjective probability. This realization was partially responsible for the original development of the CF calculus, which was supposed to be a better descriptive model than the Bayesian belief-update procedure. Nonetheless, the superior descriptive power of the CF language, if it indeed exists, did not manifest itself in this experiment. In particular, the sign tests detected no significant differences between the descriptive validity of the Bayesian and the CF languages, with the former outperforming the latter in 60% of the subjects. In

addition, no significant differences between the CF and the AHB and the AHB and the Bayesian languages were detected in terms of descriptive validity.

The normative Bayesian calculus is based on a mechanical integration of probabilities which, in our judgment, appears to have little if any descriptive appeal. Therefore, we expected ex-ante that the Bayesian language will perform poorly on descriptive grounds. The fact that the Bayesian language performed as well as the CF and AHB languages in this regard may be attributed to lack of statistical power. At the same time, this is indeed a preliminary indication that the Bayesian approach to rule-based inference should not be written off on the basis of a weak descriptive appeal.

External Validity: the hypotheses on external validity were based on the premise that the pooled ranking, L_T, can be credibly viewed as a measure of the ex-post, actual academic performance of the ten hypothetical candidates. In what follows, we defend this assumption and argue that it is indeed reasonable under the experiment's circumstances.

First, recall that the subjects were explicitly asked to asses the likelihood that the candidates will be offered tenure not on the basis of an abstract measure of divine academic justice, but, rather, on the basis of their expected performance in a specific

decision sciences department which the subjects knew very well. Tenure recommendations in this department are normally made by a group of professors. Three members of this group participated in the experiment as subjects. Hence, the opinions of these professors reflect closely what would have taken place if the candidates were actually evaluated by decision sciences faculty members.

Second, the literature on encoding subjective probabilities, an area which is closely related to our present concern, includes several examples in which synthetically (mathematically) determined consensus groups did much better than the average individual in terms of external validity (Huber, 1974). This finding was also reported by Stael Von Holstein (1972) who had groups of financial experts estimate the next 14-days stock prices. In a similar vein, Winkler (1968) had reported that consensus groups outperformed almost all individuals, regardless of the various weighing schemes used to generate their pooled judgments.

Given that the pooled ranking, L_T, is indeed a reasonable yardstick for actual performance, it was hypothesized that its correlation with the Bayesian-based ranking will be greater than the corresponding correlation with the ad-hoc CF-based ranking. Indeed, this relationship held for 73% of the subjects, a significant dominance at the (single-tailed) 0.059 level. This

result makes sense: notwithstanding the poor intuitive appeal of the Bayesian calculus, one would hope that its normative rigor would produce more accurate predictions than its ad-hoc counterparts. Incidentally, the high external validity of the Bayesian calculus was demonstrated in a number of other, unrelated empirical studies. For example, Gustafson (1969) had physicians assess Bayesian likelihood-ratios regarding various clues that explain the length of hospital stay of potential patients. After aggregating these assessments using Bayes rule, he found that the resulting estimates were far closer to the truth than the predictions made by a linear regression model employing actuarial data.

One peculiar result of the present study is the strong external validity of the ad-hoc Bayesian (AHB) language. A possible explanation might be that the AHB syntax, consisting of conditional probabilities, is identical to the normative Bayesian syntax. At the same time, the naive AHB sequential combination function (used to propagate posterior beliefs "upwards" the network) is at least as ad-hoc as the CF sequential combination function. Therefore, one would expect, ex ante, that the AHB language, like the CF language, would perform poorly in terms of external validity. In practice, though, the AHB and the Bayesian rankings turned to be very similar.

In the Yadrick et al (1988) simulation study, the AHB language (restricted to what they called "independent rule sets") was generally very accurate, with an overall average error (the absolute value of PROSPECTOR's estimate minus the true probability) as low as 0.014. At the same time, the inference trees that Yadrick et al simulated were single-leveled, meaning that the problematic AHB sequential combination function was not put to a test. It is therefore encouraging to report that the AHB language performs well in a two-level network, such as the one used in the present experiment. Whether or not the AHB language is externally valid in more complicated inference networks is remained to be seen in future research.

6. Conclusion

The major findings obtained in the limited context of this experiment are as follows. First, contrary to certain claims, the CF language is not a better descriptive model than the Bayesian language. Second, in terms of external validity, the Bayesian and the ad-hoc Bayesian languages dominate the CF language. The reader is encouraged to qualify these results with the fact that the experiment consisted of 15 subjects, and, consequently, the descriptive power of the CF language, if it indeed exists, may have gone undetected due to lack of statistical power.

So the question remains -- which belief language should a knowledge engineer employ in the next expert system that he or she is developing? It seems safe to suggest that, in spite of its vast popularity, the CF language scores low in all respects, namely normative foundation, descriptive validity, and external validity. The AHB language performed quite well in our experiment, but there seems to be no clear explanation why. The Bayesian language thus emerges as the only language of choice in non-deterministic, rule-based, expert systems.

It is important to remember, though, that Bayesian inference in complex belief networks is generally NP-hard (Cooper, 1987). If, however, the joint distribution function underlying the rule-base obeys certain assumptions of conditional independence, one can credibly employ the new Bayesian inference algorithms developed by Pearl (1986) and his colleagues. Computational complexity was not a problem in our experiment, due to the relatively small networks that we have used. More complicated networks might require restructuring and addition of extra nodes in order to remove dependencies and make the underlying joint distribution function amenable to efficient Bayesian algorithms.

7. References

Adams, J. B., "Probabilistic Reasoning and Certainty Factors' in Rule-Based Expert Systems," in: Buchanan, B. G., and Shortliffe, E. H. (Eds.), Rule-Based Expert Systems, Addison-Wesley, 1984, pp. 263-271.

Bettman, J. R., An Information Processing Theory of Consumer Choice, Reading, MA: Addison Wesley, 1979.

Buchanan, B. G. and Shortliffe, E. H., "Uncertainty and Evidential Support," in: Buchanan, B. G., and Shortliffe, E. H. (Eds.), Rule-Based Expert Systems, Addison-Wesley, 1984, pp. 217-219.

Carnap, R., Logical Foundations of Probability, Chicago: University of Chicago Press, 1954.

Charniak, E., "The Bayesian Basis of Common Sense Medical Diagnosis," Proceedings of the National Conference in Artificial Intelligence, 1983, 3, pp. 70-73.

Cohen, J., "Some Statistical Issues in Psychological Research," in B. B. Woleman (ed.) Handbook of Clinical Psychology, McGraw-Hill, New York, 1965, pp. 95-121.

Cooper, G.F., "Probabilistic Inference Using Belief Networks is NP-Hard," Tech. Report KSL-87-27, Medical Computer Science Group, Knowledge Systems Laboratory, Stanford University, Stanford, CA, May, 1987.

Dawes, R. M. and Corrigan, B., "Linear Models in Decision Making," Psychological Bulletin (1974).

Duda, R. O., Hart, P. E., and Nilsson, N. J., "Development of a Computer-Based Consultant for Mineral Exploration," SRI International Projects 5821 and 6415, october 1977.

Einhorn, H. J. and Hogarth, R. M., "Adaptation and Inertia in Belief Updating: the Contrast-Inertia Model," University of Chicago School of Business working paper, October, 1987.

Fischhoff, B., Slovic, P., and Lichtenstein, S., "Knowing What You Want: Measuring Labile Values," in T.S. Wallsten (ed.), Cognitive Processes in Choice and Decision Behavior, Erlbaum Associates, Hillsdale, New Jersey, 1980.

Grosof, B. N., "Evidential Information as Transformed Probability," in Lemmer, J. F., and Kanal, L. (eds.), Uncertainty in Artificial Intelligence. North Holland, 1986.

Gustaffson, D. H., "Evaluation of Probabilistic Information

Processing in Medical Decision Making." Organizational Behavior and Human Performance, 1969, 4, 20-34.

Heckerman, D. E., "Probabilistic Interpretation for MYCIN's Certainty Factors," in Lemmer, J. F., and Kanal, L. (eds.), Uncertainty in Artificial Intelligence. North Holland, 1986.

Henrion, M., "Should We Use Probability in Uncertain Inference Systems?" Proc. of the 8th Annual Conference of the Cognitive Sciences Society, Amherst, MA August, 1986.

Horvitz, E. J., Heckerman, D. E., Langlotz, C. P., A Framework for Comparing Alternative Formalisms for Plausible Reasoning," in Proceedings of the AAAI Conference, Philadelphia, PA, 1986, pp. 210-214.

Huber, G., P., "Methods for Quantifying Subjective Probabilities and Multi-Attribute Utilities," Decision Sci. Vol. 5 (1974), pp. 430-458.

Kopcso, D., Pipino, L., and Rybolt, W., "A Comparison of the Manipulation of Certainty Factors by Individuals and Expert System Shells," Proc. of the 21st Hawaii Int. Conf. on System Sciences, Vol III, 1988, pp. 181-188

Mitchell, D. H., "The Shape Experiment," Manuscript, Psychology Department, Northwestern University, 1986.

Pearl, J., "Fusion, Propagation, and Structuring in Belief Networks," Artificial Intelligence, September, 1986.

Robinson and Hastie, 1985: this paper is cited without reference in Mitchell, D. H., "The Shape Experiment," Manuscript, Psychology Department, Northwestern University, 1986.

Schocken, S. and Finin, T., "Prolog Meta-Interpreters for Reasoning Under Uncertainty," Center for Research on Information Systems Working Paper CRIS #165, New York University, September, 1987.

Schocken, S., Kleindorfer, P.R., "Artificial Intelligence Dialects of the Bayesian Belief Language," Center for Research on Information Systems Working Paper CRIS #160, New York University, October, 1987.

Shafer, G. and Tversky, A., "Languages and Designs for Probability Judgment," Cognitive Science 9 (1985), pp. 309-339.

Shortliffe, E. H., Computer-Based Medical Consultations: MYCIN. New York: American Elsevier, 1976.

Shortliffe, E. H., and Buchanan, B. G., "A Model of Inexact

Reasoning in Medicine," in: Buchanan, B. G., and Shortliffe, E. H. (Eds.), Rule-Based Expert Systems, Addison-Wesley, 1984, p. 242.

Stael von Holstein, C., "Probabilistic Forecasting: an Experiment Related to the Stock Market," Organizational Behavior and Human Performance, Vol. 8 No. 1 (August, 1972), pp. 139-158.

Tversky, A. and Kahneman, D., "Judgment Under Uncertainty: Heuristic and Biases," Science, 185, September, 1974, pp. 1124-1131.

van Melle, W., Shortliffe, E. H., and Buchanan, B. G., "EMYCIN: a Knowledge-Engineer's Tool for Constructing Rule-Based Expert Systems," in: Buchanan, B. G., and Shortliffe, E. H. (Eds.), Rule-Based Expert Systems, Addison-Wesley, 1984, pp. 302-313.

Wise, B. P., "Experimentally Comparing Uncertain Inference Systems in Probability," in: Lemmer, J.F. and Kanal, L.N. (Eds.), Uncertainty in Artificial Intelligence 2, North-Holland, 1988, pp. 89-102.

Winkler, R. L., "The Consensus of Subjective Probability Distributions," Management Science, Vol. 15, No. 2 (October 1968), pp. 61-75.

Yadrick, R.M., Perrin, B.M., Vaughan, D.S., Holden, P.D., and Kempf, K.G., "Evaluation of Uncertain Inference Models I: Prospector," in: Lemmer, J.F. and Kanal, L.N. (Eds.), Uncertainty in Artificial Intelligence 2, North-Holland, 1988, pp. 77-88.

Yu, V. L., et al, "An Evaluation of MYCIN's Advice," in: Buchanan, B. G., and Shortliffe, E. H. (Eds.), Rule-Based Expert Systems, Addison-Wesley, 1984, pp. 589-596.