A Comparative Analysis of the Empirical Validity of Two Rule Based Belief Languages

Shimon Schocken

Yu-Ming Wang

July 31, 1992

Department of Information Systems
Leonard N. Stern School of Business
New York University
40 West 4th Street
New York, NY 10003

Working Paper Series

STERN IS-92-24

A Comparative Analysis of the Empirical Validity of Two Rule Based Belief Languages

Rule based expert systems deal with inexact reasoning through a variety of quasi-probabilistic methods, including the widely used certainty factors (CF) and subjective Bayesian (SB) models, versions of which are implemented in many commercial expert system shells. Previous research established that under certain independence assumptions, SB and CF are *ordinally compatible*: when used to compute the beliefs in several hypotheses of interest under the same set of circumstances, the hypothesis that will attain the highest posterior probability will also attain the highest certainty factor, etc. This is very relevant to the expert systems field, where most inference engines and explanation facilities are designed to utilize the *relative* scales of posterior beliefs, making little or no use of their *absolute* magnitudes. The objective of this research is to explore empirically whether the compatibility of SB and CF extends to the field, where subjective degrees of belief and different elicitation procedures might bias the mathematical kinship of the two belief languages. In particular, we seek to know (i) whether this bias is random or systematic; and (ii) what the bias reveals about SB and CF as two alternative means to elicit and revise beliefs in a rule based system.

Key words: Belief revision, inexact reasoning, certainty factors, uncertainty in artificial intelligence.

1 Introduction

Many inference problems that occur in science and industry share a common logical structure: given a set of observed pieces of evidence, a domain expert is called upon to identify the one or more prospective hypotheses that provide the best explanation to the evidence at hand. For example, the evidence may be medical symptoms, mechanical malfunctions, or financial indicators, and the hypotheses may be diseases, machine disorders, or economic trends, respectively. In some cases, the expert's inference rationale can be described through a structured collection of `IF evidence THEN conclusion` rules, with several layers of propositions separating the immediate evidence from the prospective hypotheses. However, in many inference problems the relationship between evidence and hypotheses is non-categorical, forcing the expert to rely on a variety of causal, diagnostic, or simply correlated, reasoning chains which are inherently inexact. One way to model this uncertainty is to parameterize non-categorical rules by numeric degrees of belief which, broadly speaking, provide a measure of the degree to which the evidence supports the hypothesis. The basic argument, which goes back to Leibnitz, circa 1705 [5], is as follows: if we can elicit, represent, and manipulate, inference rules and degrees of belief in a credible way, we may be able to build an autonomous inference system that simulates, if not improves, the inferential ability of human experts.

Rule based expert systems – the contemporary realizations of Leibnitz ambitious program – consist of three parts: a rule base, an inference engine, and a belief calculus. The rule base is a collection of rules, representing textbook knowledge as well as subjective human judgement. The inference engine is a computer program that traverses the rules, pursuing reasoning chains from one set of propositions to another. As a side effect of this process, the inference engine uses a belief calculus to compute the degrees of belief in various propositions of interest. These numbers are used internally, to guide the inference engine to promising directions, and externally, to explain the system's reasoning to the people who consult it. Depending on the specific expert system *shell* that one uses, the belief calculus can be based on probability theory [16], subjective Bayesian inference [9], certainty factors [21], fuzzy sets [24], or the Dempster-Shafer theory of evidence [19]. The

literature offers numerous articles about any one of these methods, and the above references point to one representative article in each category.

This multitude of 'belief languages' was created during the last decade because of the realization that classical probabilistic methods are not well suited to support rule based inference under uncertainty. First, standard probabilistic algorithms require tremendous amounts of data, and their computational complexity is forbidding for most practical applications, two points which are illustrated in an appendix to this paper. Second, human experts are not Bayesian automatons; in fact, many studies indicate precisely the opposite: when left to its own devices, human judgement under uncertainty violates the axioms of probability theory in a predictable and systematic fashion [23]. Why, then, some argue, should we try to simulate the successful judgement of non-Bayesian experts with Bayesian techniques?

The search for intuitive and simple belief revision methods has led to the development of several *heuristic* belief languages, including the widely used certainty factors language (CF), and the subjective Bayesian language (SB). These languages are quite different on normative and cognitive grounds, and their validity in the context of rule based inference is still an open research question. We distinguish between two types of validity. *Descriptive validity* concerns the proximity of the system's outputs to the human judgement that it attempts to simulate. *External validity* concerns the proximity of the system's outputs to the actual state of the world. The paper describes a methodology and an experiment that compare the SB and CF models along these lines.

The plan of the paper is as follows. §2 reviews the heuristic approach to belief revision in expert systems, focusing primarily on the SB and CF models. §3 describes a general framework for comparing belief languages in terms of several independent validity criteria. Using this framework, §4 describes an experiment that pits the performance of SB and CF in a controlled experiment involving human subjects and a highly generic inference task. §5 presents the results of the experiment, and §6 analyzes their implications. §7 offers concluding remarks. The *normative* approach to belief revision in rule base systems is reviewed in a separate appendix that emphasizes the computational complexity that characterizes the general problem.

2 Heuristic Belief Revision Models

In this paper, 'rule bases' are viewed as *inference networks* in which nodes represent propositions and arcs represent rules. The specific topology of the networks are determined by the domain experts and knowledge engineers who construct them. Figure 1 depicts a simple network that is sufficiently rich to illustrate the general belief revision patterns that occur in rule base inference systems. The network consists of five propositions: two pieces of evidence (E_1 and E_2), one hypothesis (H), and two sub-hypotheses (S_1 and S_2). The directed arc that leads from proposition x to proposition y represents the rule `IF x THEN y`, and the arc's label represents the degree of belief associated with that rule.

For simplicity, we assume that the value of each proposition is either `true`, `false`, or `unknown`. Throughout the paper, uninstantiated propositions are denoted by upper-case characters like E, whereas lower-case characters like e and $\overline{e}$ denote the assertions *E is known to be true* and *E is known to be false*, respectively. With this notation, figure 1 entails eight prototypical bodies of evidence: $\{e_1\}$, $\{\overline{e_1}\}$, $\{e_2\}$, $\{\overline{e_2}\}$, $\{e_1, e_2\}$, $\{\overline{e_1}, e_2\}$, $\{e_1, \overline{e_2}\}$, $\{\overline{e_1}, \overline{e_2}\}$, and $\emptyset$ – the case of no evidence at all.

The degrees of belief that parameterize rules of the form $E \xrightarrow{d} H$ are directed, and different belief revision languages have different directionality. Some languages set $d = bel(H|E)$, the predictive support that the piece of evidence E renders to the hypothesis H. Other languages set $d = bel(E|H)$, the diagnostic impact of H on E, or the likelihood of observing the evidence E given that H is true. This information is then used to compute the posterior belief in H, using a variety of 'forward' and 'backward' reasoning techniques. If the structure of the rule base is consistent with a set of simplifying assumptions about the joint distribution function which characterizes the evidence/hypotheses space, then there exist efficient belief update algorithms which are consistent with the axioms of probability theory, e.g. [16] and [7]. This *normative* approach to belief revision in rule based systems is described at the end of the paper, in a separate appendix.

Many practitioners, though, construct rule bases and elicit degrees of belief with little or no attention to the probabilistic backdrop of the problem.

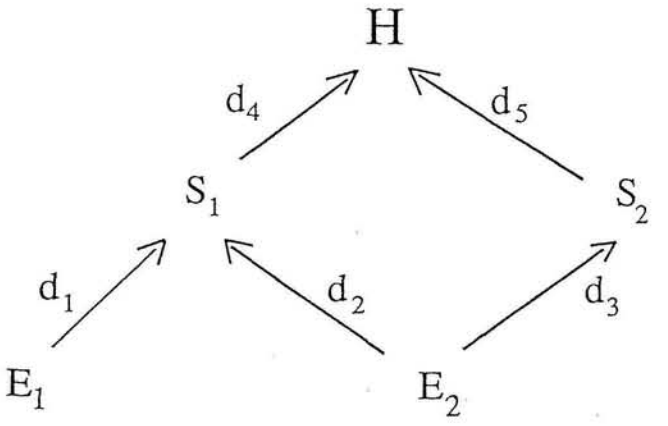

Figure 1

Furthermore, most expert system shells use *heuristic* belief calculi that are inspired not by probability theory, but by non-categorical logic. That is, the belief calculus is viewed as an extension of the inference engine, designed to compute degrees of belief on the fly, as a side effect of the standard reasoning process. As a result, the notion a joint distribution function, which is central in normative models, is either absent or implicit in heuristic models.

Instead of carrying out global manipulations of a joint distribution function, heuristic belief calculi compute posterior beliefs in a local fashion, using a variety of *sequential* and *parallel* combination functions, denoted here f_s and f_p. Specifically, consider the rule $E \xrightarrow{d} H$, and suppose that a certain body of evidence Q causes the system to change the belief in the rule's premise E to $bel(E|Q)$. In such a case, a *sequential combination function* is used to revise the rule's original degree of belief, an operation which we denote $d' = f_s(d, bel(E|Q))$. When two or more rules $E_1 \xrightarrow{d_1} H$ and $E_2 \xrightarrow{d_2} H$ render support to the same proposition H, the posterior belief in H is updated through a *parallel combination function*, an operation which is denoted $bel(h|E_1, E_2) = f_p(d_1, d_2)$. The exact definitions of f_s and f_p vary from one shell to another, but their principle operation is the same: the posterior belief

in any proposition H is computed by applying f_s and f_p recursively to all the degrees of belief encountered along the reasoning chains that ultimately imply H.

To illustrate, refer to figure 1, and suppose that a certain body of evidence, say $E = \{e_1, e_2\}$, would cause the inference engine to infer s_1 and s_2. As a side effect of this inference, the belief in s_1 will be updated (through the parallel combination function) to $f_p(d_1, d_2)$, and the belief in s_2 will become d_3. Now, S_1 and S_2 play a dual role in this rule base. Till now, they were treated as *conclusions* of rules. However, once we move forward in the reasoning chain, S_1 and S_2 become *premises* of higher level rules. Since the beliefs in these premises were just updated, the degrees of belief of the rules that emanate from them should also be updated. Hence, the original degrees of belief d_4 and d_5 change to $d'_4 = f_s(d_4, f_p(d_1, d_2))$ and to $d'_5 = f_s(d_5, d_3)$. Finally, the parallel combination function is applied once again to revise the belief in H, which becomes $f_p(d'_4, d'_5)$.

The parallel combination function is typically assumed to be commutative and associative, and its extension to n rather than two degrees of belief is straightforward. The sequential combination function is assumed to be monotonically increasing in both of its arguments. Hence, it is sometimes referred to as an *attenuation function*, designed to update the rule's strength when the rule's premise becomes more or less certain. In order to define a belief language, then, one must specify three things: (i) the mathematical domain of the degrees of belief (the d's); (ii) the sequential combination function f_s; and (iii) the parallel combination function f_p. Some shells offer fixed implementations of f_s and f_p, whereas other treat the belief calculus as an external parameter, supplied by the system's designer [17].

This paper focuses on the empirical validity of two widely used *parallel combination* rules – the formulae that support inference patterns like the one depicted in figure 2. Note that due to the locality of heuristic calculi, it doesn't matter if figure 2 represents a stand-alone network or a subset of a network with lower levels of inference. That is, if it is assumed that the degrees of belief $d_1, \ldots, d_n$ were already attenuated by a sequential combination function, it is no longer necessary to consider the lower level propositions that established the body of evidence $E_1, \ldots, E_n$. Thanks to the local na-

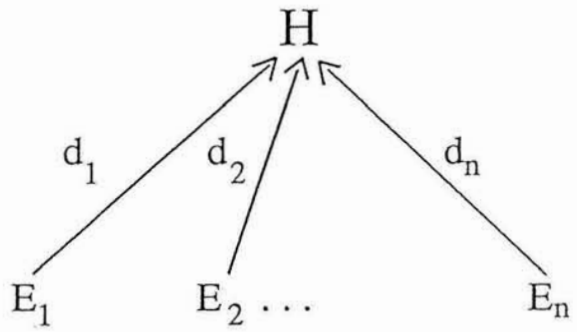

Figure 2

ture of heuristic calculi, one can assume that the evidential impact of these propositions on H was already absorbed through attenuation in the degrees of belief that support H directly.

Referring to figure 2 as a common point of departure, the remainder of this section describes the parallel combination functions of the subjective Bayesian language and the certainty factors language. In both descriptions, $bel(H|E)$ is used to denote the posterior belief in H in light of the body of evidence $E = \{E_1, \ldots, E_n\}$.

2.1 The Subjective Bayesian Language

Many expert systems (most notably, Prospector, [9]) employ a subjective Bayesian language (SB) whose parallel combination function is essentially a heuristic version of Bayes rule. In the SB language, the posterior belief $bel(H|E)$ is typically expressed through the likelihood ratio $P(H|E)/P(\overline{H}|E)$, and the degrees of belief associated with rules of the form $E_i \xrightarrow{d_i} H$ are expressed through the conditional likelihood ratios $d_i = P(E_i|H)/P(E_i|\overline{H})$.

When the premise E_i supports h, $d_i > 1$. When E_i supports $\overline{h}$, $d_i < 1$. When E_i has no relevance to H, $d_i = 1$.

Before any evidence is taken into consideration, the posterior belief in H is initialized to the prior belief $P(H)/P(\overline{H})$, when such information is available in the knowledge base (when priors are not known most systems assume that all the hypotheses are equally likely a-priori). When new information is brought to bear, the posterior belief is updated through the rules that pertain to that information. For example, suppose that the rule base includes two rules of the form $E_1 \xrightarrow{d_1} H$ and $E_2 \xrightarrow{d_2} H$, with $d_1 = P(E_1|H)/P(E_1|\overline{H})$ and $d_2 = P(E_2|H)/P(E_2|\overline{H})$. If, at a certain point of time, the inference engine establishes the evidence E_1 and E_2, the posterior belief in H is revised through the following parallel combination function:

$$bel(H|E_1, E_2) = \frac{P(H|E_1, E_2)}{P(\overline{H}|E_1, E_2)} = \frac{P(E_1|H)}{P(E_1|\overline{H})} \cdot \frac{P(E_2|H)}{P(E_2|\overline{H})} \cdot \frac{P(H)}{P(\overline{H})} \qquad (1)$$

This is a special case of the likelihood ratio version of Bayes rule:

$$\frac{P(H|E_1, E_2)}{P(\overline{H}|E_1, E_2)} = \frac{P(E_1, E_2|h)}{P(E_1, E_2|\overline{H})} \cdot \frac{P(H)}{P(\overline{H})} \qquad (2)$$

Under the following 'ratio form conditional independence' assumption:

$$\frac{P(E_1, E_2|H)}{P(E_1, E_2|\overline{H})} = \frac{P(E_1|H)}{P(E_1|\overline{H})} \cdot \frac{P(E_2|H)}{P(E_2|\overline{H})} \cdot \qquad (3)$$

Since (1) is commutative and associative with respect to E_1 and E_2, the SB parallel combination function for $n > 2$ rules is a straightforward extension of the binary case.

2.2 The Certainty Factors Language

The Certainty Factors language (CF) was first formalized and implemented in the seminal *MYCIN* project [2]. It was subsequently implemented in many expert system shells, e.g. *EMYCIN*, *M1*, and *VP-Expert*. In the *CF* terminology, which is reminiscent of Carnap's theory of confirmation [3], prior and posterior beliefs are expressed through certainty factors that vary from -1 to 1. The degree of belief associated with the rule $E_i \xrightarrow{d_i} H$ is called the conditional *certainty factor* $CF(H|E_i)$. This number also varies from -1 (E_i confirms $\overline{h}$ with certainty) to 1 (E_i confirms h with certainty). The case of E_i being irrelevant to H is modeled through $CF(H|E_i) = 0$.

When two rules lend credence to H simultaneously with degrees of belief $CF(H|E_1)$ and $CF(H|E_2)$, the belief in H in light of the body of evidence $\{E_1, E_2\}$ is updated by the following parallel combination function:

$$bel(H|E_1, E_2) =$$

$$\begin{cases} CF(H|E_1) + CF(H|E_2) \cdot (1 - CF(H|E_1)) & \text{if both } CF\text{'s are positive} \\ -(|CF(H|E_1)| + |CF(H|E_2)| \cdot (1 - |CF(H|E_1)|) & \text{if both } CF\text{'s are negative} \\ \frac{CF(H|E_1)+CF(H|E_2)}{1-min\{|CF(H|E_1)|,|CF(H|E_2)|\}} & \text{if the } CF\text{'s have mixed signs} \end{cases} \tag{4}$$

The parallel combination function (4) is commutative and associative, and it can be applied recursively to compute the posterior belief in H in light of any number $n > 2$ of rules. The expert systems literature offers several descriptive and normative justifications of this function, and the interested reader is referred to [2] for a collection of key papers on the subject.

3 Comparative Analyses

The mathematical relationship that links certainty factors to probabilities was studied extensively, e.g. [1], [12], [11], [13], and [18]. Most of these

analyses focused on the normative validity and implicit proximity of (1) and (4). Parallel combination formulae have captured a great deal of attention because they are widely used in many systems that carry out rule based inference under uncertainty. Furthermore, they entail alternative models of the classical Bayesian learning process, in which an agent (either a human or an automaton) revises his or her beliefs in certain propositions as new information is brought to bear [10].

The key finding that emerged from this line of research is that assumption (3), along with certain transformations that map CF's on probabilities, imply that the two parallel combination formulae (1) and (4) are isomorphic to each other. However, the analytic approach has been controversial because of its fundamental reliance on inter-scale mappings of questionable validity. In this research we sidestep the problem completely by investigating a weaker linking property, denoted *ordinal compatibility*: two belief languages are said to be *ordinally compatible* if their respective calculi generate *monotonic* scales of degrees of belief.

For example, suppose that CF and SB were used to compute the posterior beliefs associated with the same set of hypotheses. The resulting two sets of numbers may not be identical, or even related to each other in any close algebraic form, but they may still be perfectly calibrated in terms of order. That is, the hypothesis which attains the highest posterior probability may also attain the highest certainty factor, and so forth. This property has important practical implications on the joint validity of the two languages. To illustrate, consider a medical expert system that presents its final prognosis as a list of several diseases sorted by decreasing certainty factors. Suppose now that the CF language employed by the system were replaced by an *SB* language. If all other things were held equal, including the physician's simulated expertise and the patient's data, is it possible that the system would switch its prognosis from one disease to another? Such a result would indicate that at least one of the languages under consideration is inconsistent with either the expert, the actual state of the world, or both.

The following thought experiment might help formalize our approach. Let $E = \{E_1, \ldots, E_n\}$ and $H = \{H_1, \ldots, H_m\}$ be two exhaustive sets of pieces of evidence and hypotheses, respectively. Without loss of generality, suppose

that the body of evidence $\{E_1, E_2\}$ supports the hypothesis H_1. We observe that the posterior belief in H_1 in light of $\{E_1, E_2\}$ can be derived, at least in theory, in three different and possibly inconsistent ways. First, the joint distribution function $P(E_1, \ldots, E_n, H_1, \ldots, H_m)$ – the function that governs the occurrence frequency of tuples in the propositional space $E \times H$ – could be used to compute $P(H_1|E_1, E_2)$ (the exact derivation is given in the appendix). This number, which constitutes the *actual* posterior belief in H_1 in light of $\{E_1, E_2\}$, is denoted hereafter S_e. Alternatively, a human expert in the domain $E \times H$ may be asked to assess his or her *subjective* belief in H_1 in light of $\{E_1, E_2\}$, denoted hereafter S_h. Finally, a rule based system that simulates the reasoning of the very same expert could be fed with the fact base $\{E_1, E_2\}$. The system, which employs a certain belief language L, would go on to produce a machine generated posterior belief in H_1, denoted hereafter S_L.

In order to proceed in the thought experiment, we now have to make a rather heroic assumption, as follows. We assume that the three numbers S_e, S_h, and S_L lie on the same interval scale of measurement [22], and therefore, that they are comparable. Delaying a discussion of the validity of this assumption to the end of this section, we define three performance criteria, as follows:

$$
\begin{array}{lcl}
|S_h - S_e| & : & \texttt{expert's external validity} \\
|S_L - S_e| & : & \texttt{system's external validity} \\
|S_L - S_h| & : & \texttt{system's descriptive validity}
\end{array}
$$

Note that the expert's external validity is an intrinsic property of the expert. At the same time, the external and the descriptive validity of the *system* depend on the belief language that the system employs. Till now, the thought experiment involved only one such language, denoted L. However, alternative languages may be considered, with the provision that all languages operate in the context of the same expert, the same system, and the same bodies of evidence throughout. By comparing the various posterior beliefs that these languages produce to S_e and to S_h (which are fixed), one can make statements about the *relative* validity of the languages. Ideally, these statements should withstand the test of statistical significance. This is the crux of our approach.

In view of the fact that S_e, S_h, and S_L represent, respectively, an objective

probability, a subjective measure of human belief, and, say, a certainty factor, the assumption that these numbers are comparable is indeed disturbing, unless one is willing to base one's analysis on arbitrary inter-scale mappings. However, our ordinal perspective on the problem avoids the assumption by focusing not on the absolute magnitudes of these measures, but on the way they rank-order a fixed set of hypotheses in terms of support. To clarify, let x and y be two alternative 'methods' to compute the posterior beliefs associated with a certain set of hypotheses, denoted S. In the experiment that is described below, S_x and S_y don't represent scaler measures, as they did in the thought experiment, but rather two alternative permutations of S, i.e. two different ways to order the same set of hypotheses in terms of support. Likewise, the term $|S_x - S_y|$ does not represent a scaler distance, but the vectorial correlation of the two rankings S_x and S_y. This methodology, which can be used to analyze the comparative validity of any two belief languages, is discussed in detail in the next section.

4 The Experiment

In order to compare the empirical validity of SB and CF on a level ground, we constructed an experiment in which the two languages were supposed to deliver, at least in theory, compatible results. The context of the experiment was a data-driven *credit rating* task that required no formal financial knowledge. The subjects in the experiment were 28 undergraduate business school students, enrolled in an information systems course. The credit rating skills of the subjects were built and then simulated in four stages, as follows. The first part of the experiment consisted of training. During a period of six weeks, the subjects were given many examples of company profiles, on the one hand, and historical loan repayment and default records, on the other. The subjects were asked to study the data set and try to detect patterns that might be used to derive credit rules, although no specific instructions were given beyond these broad guidelines. In the second part of the experiment, the subjects were asked to predict the repay/default likelihoods of eight prototypical company profiles, based on what they've learned during training. In the third part of the experiment, the subjects' credit rating rationale was elicited using both the SB language and the CF language. In

the final stage of the experiment, the two languages were used to rank order the same company profiles that the subjects ranked before. The pair-wise similarities of these rankings were then used to test several hypotheses about subjects' reliability, descriptive validity, and external validity.

The task: Most credit rating models rely on a variety of financial ratios as well as industry, managerial, and other subjective criteria. However, in order to minimize biases of partial or ill conceived domain knowledge, the experiment focused on company attributes that, ex ante, would appear to be completely neutral from a credit rating perspective. The subjects learned the predictive power of these attributes in an inductive way, using many examples that came from a controlled distribution. Hence, the credit rating problem was used only as an interesting context, designed to bring the data to life and inject a sense of competition among the subjects. At the beginning of the experiment, the subjects were told that the goal of the training stage was to prepare them for a predictive credit analysis task that will take place toward the end of the semester. Next, we announced that the subjects whose predictions will come closest to the actual solution will receive monetary awards of \$50, \$30, and \$20, respectively.

The chief objective of the training program was to endow the subjects with a certain degree of expertise in the narrow problem domain that we have constructed. In order to achieve this goal, the subjects were exposed to many company profiles that either succeeded or failed to repay previous loans. Each company profile, or 'case,' consisted of four binary attributes. E_1: whether the company is `unionized` or `non-unionized`; E_2: whether it is `private` or `public`; E_3: whether it sells `consumer` or `industrial` products; and H: whether it `repaid` or `defaulted` on its previous loan. The E_i's were construed as pieces of evidence, and H was construed as an hypothesis, although neither this interpretation nor the attribute labels were given to the subjects. Instead, the subjects were presented with textual descriptions, e.g. "company 17, which is non-unionized, industrial-oriented, and private, defaulted on its loan" or "Company 11, which repaid its loan, was a unionized and privately-held producer of consumer goods." Both the order of the companies and the order of the attributes within the company profiles were randomized, to avoid potential presentation biases.

With three binary pieces of evidence, the 'case base' consisted of varying repay/default histories of eight prototypical company profiles. The distribution of these profiles was structured in such a way that the *actual* posterior repay/default probability of each company could be computed from the data. For example, to determine the posterior probability that a unionized, public, and consumer goods company will make good on its loan, we computed the the likelihood ratio $P(h|e_1, \overline{e_2}, e_3)/P(h|e_1, \overline{e_2}, e_3)$ through Bayes rule, operating on the data-driven probabilities $p(e_1|h)$, $p(e_1|\overline{h})$, $p(\overline{e_2}|h)$, $p(\overline{e_2}|\overline{h})$, and $p(e_3|h)$, $p(e_3|\overline{h})$. This provided a 'gold standard' ranking of the eight companies against which other rankings could be compared.

In order to facilitate these computations, and, at the same time, create an interesting inference task, the structure of the case base had to satisfy many simultaneous constraints. First, the distribution of the four attributes was made to be consistent with the assumption of ratio form conditional independence (3). Second, we wanted the three explaining attributes to have a large, medium, and small impact on the posterior distribution of the outcome attribute, in terms of conditional likelihood ratios. Third, we wanted the posterior repay probabilities of the eight companies to be nicely spread apart, to allow for distinguishable differences. Fourth, the eight company profiles had to appear roughly the same time in the case base, so that individual profiles will not dominate or skew the distribution. Fifth, the number of `repay` and `default` cases (viz, the prior probabilities) had to be roughly the same, to control for potential representativeness biases during the elicitation procedure[1]. Finally, the total number of cases had to be kept within a reasonable range, to avoid cognitive strain from the subjects perspective.

Training: Following several simulations and a pilot study, a case base of 66 companies was constructed, satisfying all the above constraints, with the possible exception of the last one. Several steps were taken to make the

[1]The representativeness bias [23] occurs when human experts overestimate the predictive power of observed evidence, say a recommendation letter or a job interview, and underestimate the power of objective information, e.g. the base-rate probability that *any* person will succeed in a certain job. However, when the prior information is neutral, or, in our case, $P(h) = P(\overline{h})$, representativeness is in fact a good heuristic, because the evidential data E is the *only* relevant information for updating the belief in either h or not $\overline{h}$. In terms of formula (1), this is the situation when $\frac{P(H|E_1,E_2)}{P(\overline{H}|E_1,E_2)} = \frac{P(E_1|H)}{P(E_1|\overline{H})} \cdot \frac{P(E_2|H)}{P(E_2|\overline{H})}$.

experimental task manageable and interesting. First, the case base was randomly divided into three data sets comprising of 16, 16, and 32 company profiles. The subjects were asked to analyze these data sets in three different spreadsheet modeling exercises that spanned a period of 6 weeks. In each exercise, the subjects were asked to encode, sort, and process, the company data in any way they thought fit to detect useful credit rating rules. With each exercise, two weeks were given to hand in the following items: (i) a spreadsheet model of all the data accumulated thus far; (ii) an executive summary that proposes rules for identifying good and bad credit; and (iii) a data-driven justification of the proposed rules. The homeworks were scrutinized for errors and task miscomprehension, but no feedback on the quality of the reports was given. For example, if a student wrote that 70% of the defaulting companies were private and unionized, and the actual figure was 30%, the error was corrected, but no comment on the potential usefulness of such a rule was given. Needless to say, some subjects worked harder than others, a fact which was clearly reflected in their subsequent prediction ability, as we'll see shortly.

The Questionnaire: After the subjects handed in their third and final exercise, an in-class questionnaire was administered. In the first part of the questionnaire, English descriptions of the eight prototypical $< E_1, E_2, E_3 >$ company profiles were handed out, and the subjects were asked to rank them in terms of decreasing order of predicted ability to repay their debt. Next, the subjects were told that two computer based inference models will also be used to rank-order the same companies, and that the accuracy of these models hinges on certain parameters that they have to specify. This set the stage for an elicitation procedure that focused on assessing the degrees of belief associated with 6 rules of the form $E_i \xrightarrow{d} H$, $i = 1, 2, 3$. Each subject underwent two elicitation 'treatments,' using both the SB and the CF languages. The order of the two treatments was randomized across the subjects.

Bayesian Elicitation: The SB elicitation procedure involved three pairs of 'backward inference' questions. For example, in the case of E_1 = `the company is private/public`, the subject was asked to answer the following two questions: "Assume that a certain company **repaid** its loan. Given what you've learned during training, what is your belief that the company is either **public** or **private**?" The subject was asked to answer this question

by placing an X marker on the following graphical scale:

```
PUBLIC                                                          PRIVATE
  |-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
100    90    80    70    60    50    60    70    80    90   100
```

The subject's answer was then used to determine the pair of subjective probabilities $P(e_1|h)$ and $P(\overline{e_1}|h)$. Next, the same question was repeated under the alternative background assumption that the company has defaulted, leading to the pair of probabilities $P(e_1|\overline{h})$ and $P(\overline{e_1}|\overline{h})$. Three such pairs of questions were asked, one for each piece of evidence E_1, E_2 and E_3.

Certainty Factors Elicitation: The CF elicitation procedure consisted of 'forward inference' questions. For example, in the case of $E =$ `the company is private/public`, the question was "Assume that a certain company is **private**. Given what you've learned during training, what is your belief that the company will either **repay** or **default**?"

```
REPAY                                                           DEFAULT
  |-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
100    90    80    70    60    50    60    70    80    90   100
```

The subject's answer was then used to determine the pair of certainty factors $CF(h|e_1)$ and $CF(\overline{h}|e_1)$. Next, the same question was repeated under the alternative assumption that the company is `public`, leading to the pair of certainty factors $CF(h|\overline{e_1})$ and $CF(\overline{h}|\overline{e_1})$. Three such pairs of questions were asked, one for each piece of evidence E_1, E_2 and E_3.

After the two elicitation procedures were completed, the subject's degrees of belief were plugged into their respective SB and CF belief calculi. Formula (1) was used to produce a ranking of the eight companies in terms of posterior probabilities, and formula (4) was used to produce a ranking of the same set of companies in terms of posterior certainty factors. This completed the data-gathering part of the experiment.

5 Results

To recapitulate, each subject produced three rankings of the same set of eight companies: human ranking, SB-based ranking, and CF-based ranking. These rankings are denoted S_h, S_b, and S_c, respectively. In addition, the actual ranking of the same companies, denoted S_e, was available in terms of posterior repay probabilities computed from the data. Technically speaking, each of the four rankings was a permutation of the same set of eight company numbers. If all rankings were the same for a certain subject, the subject would be a perfect predictor, and the SB and CF languages would be equally capable of capturing his or her prediction rationale. In reality, the rankings exhibited various degrees of similarity across the subjects, and these similarities were used to test hypotheses about subject reliability and language validity.

Specifically, the pair-wise similarities of the various rankings were estimated through the Spearman rank correlation coefficient, denoted hereafter R. This statistic varies from -1 (reversed rankings) to 1 (identical rankings) through 0 (no correlation). With 8 items in each rankings, the critical value above which the two rankings are said to be significantly correlated is $|R| > 0.643$ ($p = 0.05$, bi-directional test) [15]. The specific rankings and correlations that were studied in this experiment are depicted in figure 3 as nodes and arcs, respectively. In the figure and throughout the paper, $R(x, y)$ denotes the Spearman rank correlation coefficient between the rankings S_x and S_y.

For each subject, $R(h, e)$ was used to estimate the proximity of the subject's *human* ranking, S_h, to the *actual* ranking, S_e. This coefficient provided a measure of the task expertise that the subject has gained during the preliminary training stage. If the subject learned the features of the data well and was capable of synthesizing his or her knowledge into accurate predictions, the subject's $R(h, e)$ score was close to 1. $R(h, e)$ scores less than 1 but greater than 0.643 characterized subjects whose predictions where better than random, indicating varying degrees of gained expertise. Finally, low or negative $R(h, e)$ scores characterized subjects who failed the prediction task, either because of poor training or because they were unable to translate what they've learned into accurate predictions. To give a sense of the difficulty

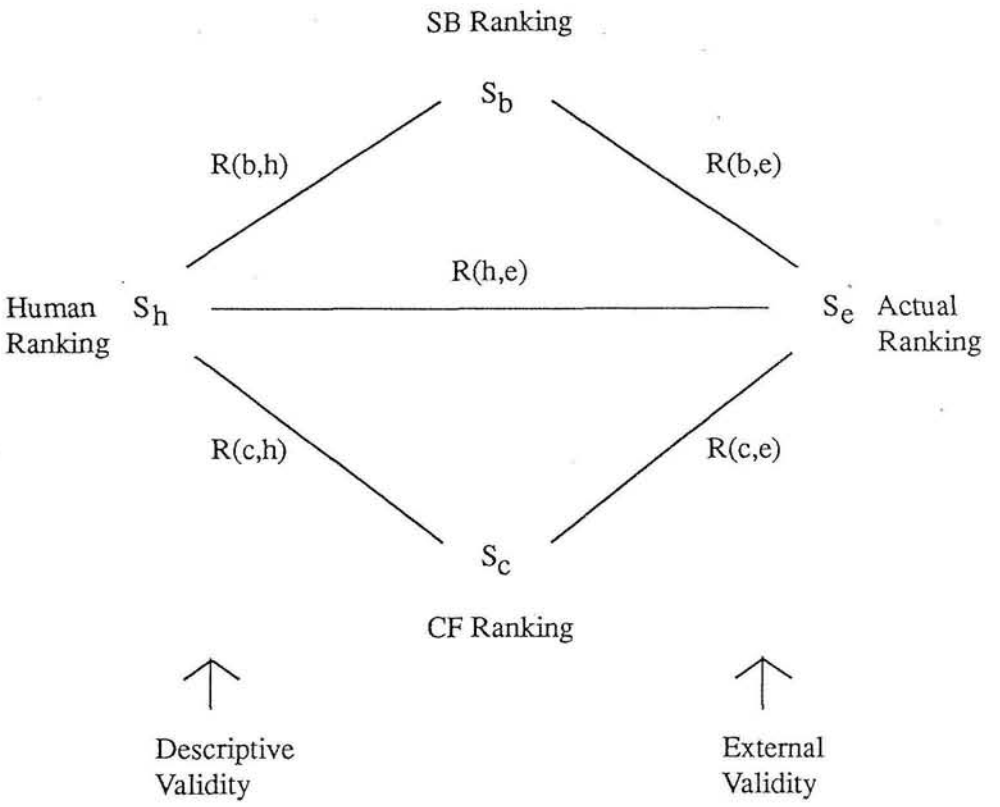

Figure 3

that the inference task posed to the subjects, note that the act of specifying a ranking (either by human or by machine) amounted to guessing the correct order of 8 entities. Therefore, the probability of getting a reliability score of 1.0 by chance alone was $1/!8 = 0.000024$. Hence, a reliability score greater than 0.9 (to pick an arbitrary cutoff value) indicates that the subject became quite an expert in the limited context of this experiment.

Of the 35 subjects who began training, 28 had positive $R(h, e)$ scores, of which 24 were greater than 0.643, and 7 had negative $R(h, e)$ scores. The latter 7 subjects were considered unreliable in the context of this study. Because the experiment was designed to simulate a certain degree of task expertise, and because we wished to avoid analyzing rankings that were pulled out of a hat, the 7 unreliable subjects were subsequently eliminated from the experiment.

For each of the remaining 28 subjects, the *descriptive validity* of SB and CF were estimated through the coefficients $R(b, h)$ and $R(c, h)$, respectively. These coefficients measured the degree of agreement between the subject's human ranking and the subject's simulated SB and CF rankings. The *ex-*

ternal validity of SB and CF were estimated through the coefficients $R(b, e)$ and $R(c, e)$, which measured the degree of agreement between the subject's simulated SB and CF rankings and the actual ranking. The various coefficients of all the subjects are listed in table 1, which is sorted in decreasing reliability scores, or gained expertise.

With the exception of the last four subjects in table 1, all subjects had significant reliability scores, indicating a better than random ability to assign credit ratings consistent with the correct solution. For 10 out of the 28 subjects, the SB and CF rankings were precisely the same, resulting in tied descriptive and external validity measures (subjects 4,6,8,10,13,14,17,20,26, and 27). For the remaining 18 subjects, the simulated rankings were in various degrees of agreement. For example, consider subject 16, an 'average' performer with a reliability score of 0.82. For this subject, CF performed better than SB in terms of external validity ($0.98 > 0.93$) as well as descriptive validity ($0.77 > 0.62$). For this subject, both SB and CF scored very high in terms of external validity, providing better predictions than the subject's human performance ($0.93, 0.98 > 0.82$). This is an example of a 'bootstrapping effect' in which a simulated model of an expert outperforms the expert's own predictions, typically because the model is more robust and consistent than the human [8].

The most interesting pattern in table 1 was the relative performance of SB and CF as a function of the subject's human performance. In roughly the top two thirds of the table – subjects 1 to 17 – the CF language performed either as well as, or better than, the SB language in terms of *both* external and descriptive validity. Within the 11 subjects at the bottom of the table, no language dominated the other categorically. To test the statistical significance of this pattern across the entire subject population, we ran a series of Wilcoxon tests, as follows. For each subject, we say that language L_1 dominates language L_2 in terms of external validity if, for this particular subject, $R(L_1, e) > R(L_2, e)$. Likewise, L_1 is said to dominate L_2 in terms of descriptive validity if $R(L_1, h) > R(L_2, h)$. Finally, we say that language L outperformed the subject's own predictions if $R(L, e) > R(h, e)$. To test if such significant language effects persisted in the subjects population, four bidirectional Wilcoxon signed-rank tests were applied to pairs of R columns from table 1, with the basis hypotheses of no significant differences between

	subject reliability	external validity		descriptive validity	
N#	R(h,e)	R(b,e)	R(c,e)	R(b,h)	R(c,h)
1	0.99	0.99	1.00	0.94	0.98
2	0.98	0.81	0.94	0.73	0.88
3	0.98	0.59	0.98	0.60	0.99
4	0.97	1.00	1.00	0.97	0.97
5	0.97	0.78	0.93	0.86	0.95
6	0.96	0.87	0.87	0.88	0.88
7	0.95	0.88	0.94	0.89	0.94
8	0.93	0.87	0.87	0.94	0.94
9	0.92	0.52	0.90	0.53	0.96
10	0.91	0.88	0.88	0.97	0.97
11	0.90	0.74	0.82	0.92	0.96
12	0.90	-0.52	1.00	-0.28	0.90
13	0.90	0.94	0.94	0.94	0.94
14	0.88	0.88	0.88	1.00	1.00
15	0.83	0.68	0.90	0.91	0.95
16	0.82	0.93	0.98	0.62	0.77
17	0.80	1.00	1.00	0.80	0.80
18	0.80	0.81	0.94	0.94	0.88
19	0.79	0.87	0.98	0.89	0.80
20	0.78	0.78	0.78	1.00	1.00
21	0.75	0.88	0.79	0.74	0.99
22	0.70	0.93	0.82	0.82	0.82
23	0.67	0.59	0.67	0.93	0.92
24	0.66	0.88	0.79	0.87	0.92
25	0.63	1.00	0.90	0.63	0.37
26	0.54	0.98	0.98	0.60	0.60
27	0.43	0.48	0.48	0.99	0.99
28	0.37	0.38	0.34	0.98	0.81

Table 1

Hypothesis	Result	Z score	Sig. Level
External validity	CF dominates SB	2.02	0.05
Descriptive validity	CF dominates SB	1.67	0.10
CF bootstrapping effect	CF dominates subject	1.97	0.05
SB bootstrapping effect	SB dominates subject	0.24	non-sig.

Table 2

the R values in the compared columns. The results of the tests are given in table 2 and discussed in the next section.

The Wilcoxon test is a powerful non-parametric means to detect differences in paired values that come from an unspecified distribution (in this case, the R values). Because it takes into account not only the signs of the differences but also their magnitudes, the Wilcoxon test is considerably more efficient than the Sign test whenever the difference magnitudes are available. For samples larger than $n = 8$, the Wilcoxon test statistic Z has an approximate normal distribution, providing a test which is only slightly less efficient than the classical T test [20]. The test can be either directional or nondirectional, depending on the alternative hypothesis. In this study, no a-priori assumptions were made about the potential advantage of one language over the other. Therefore, all the tests were bidirectional.

A follow up investigation of table 1 focused on the last four subjects, who had non-significant reliability scores. Subjects 25 and 26 exhibited a strong bootstrapping effect. Subjects 27 and 28 had the peculiar combination of low reliability and external validity scores, on the one hand, and very high descriptive validity scores, on the other. The subject type that is consistent with these results is a student who, perhaps for lax training, didn't come close to the correct solution, yet was capable of expressing his or her poor prediction rationale very well via the elicitation procedures. This is tantamount to a credit rating expert system that provides an excellent simulation

of an incompetent human credit analyst. Three reruns of all the Wilcoxon tests without subjects 25 and 26, without subjects 27 and 28, and without subjects 25,26,27, and 28, yielded the same results as in table 2, with the only difference that in the latter two runs, the second hypothesis was significant at the $p = 0.05$ level instead of 0.1.

6 Discussion

The overall validity of a belief language hinges on the individual validity of several components. In order to conduct a rigorous study of these components, it is necessary to isolate them from the rest of the language. This study focused on how SB and CF carry out parallel combination – a fundamental belief revision pattern that comes to play whenever one caries out rule based inference under uncertainty. Our research methodology centered on the notion of *ordinal compatibility*. Specifically, if CF and SB were ordinally compatible under parallel combination, the posterior belief scales that the two languages generate (when applied to the same problem) would be perfectly correlated except for random noise, in which case the two languages would be indistinguishable in terms of validity. Yet the experiment indicated otherwise. In most cases, the SB and CF rankings of the same subject were positively correlated, but not identical. The differences were statistically significant in one direction, favoring CF on SB in terms of three different validity criteria (see table 2). The remainder of this section explores several explanations to these results.

Random differences: Let $R_i(c, h)$ and $R_i(b, h)$ be the descriptive validity scores of CF and SB for the ith subject, $i = 1, \ldots, 28$. When two compatible models operate on subjective inputs elicited from human subjects, it should be expected that the outputs of the models will also be compatible, except for random noise. Specifically, if CF and SB were ordinally compatible, the rankings S_c and S_b would be more or less the same, in which case $R_i(c, h) - R_i(b, h)$ would be randomly distributed around zero. In actuality, the Wilcoxon tests conducted in this experiment indicate that the random differences hypotheses should be rejected in favor of the alternative hypothesis that $R(c, h) - R(b, h)$ was significantly positive across the subjects

population. This dominance was displayed in three different performance criteria: descriptive validity (p=0.05), external validity (p=0.1 or p=0.05 – see end of §5), and bootstrapping ability (p=0.05). Therefore, it appears that in this experiment, the superior performance of the CF language was not a random phenomenon, but rather a manifestation of a concrete advantage, yet to be identified.

Subject's mindset: The SB and CF elicitation procedures are congruent with two opposite inference 'mindsets:' backward reasoning and forward reasoning, respectively. In the SB language, one is asked to specify one's belief that a piece of evidence E will be observed, given that the background hypothesis H is true. In contrast, the CF language requires one to specify one's belief in H occurring, given that E is observed. During the training stage, the subjects were exposed to many examples of evidence and hypotheses, and were asked to identify and articulate patterns that might be useful for prediction purposes. However, the subjects were not instructed to adopt any one particular reasoning style, and were free to analyze the data in any way they thought fit. Therefore, the strong performance of the CF language may be attributed to the possible explanation that, during training, most subjects conditioned themselves to think about the features of the data in terms of forward reasoning, consistent with the CF parlance.

In order to test this hypothesis, we went back to the students homework (the written reports that they submitted during training), and coded their reasoning style into four categories: forward, backward, combination, and neither. 'Forward' reports included a majority of observations of the form "Companies that have this or that characteristic tend to repay (or default on) their loans." 'Backward' reports were dominated by observations of the form "Companies that repay (or default on) their loans tend to have this or that characteristic." 'Combination' reports involved a mixture of these formats, whereas 'neither' reports used a variety of ad-hoc scoring methods instead of rule based reasoning.

The results of this analysis did not support the hypothesis of a pre-conditioned CF mindset. First, the number of subjects in the 'forward' and 'backward' groups was almost the same (12 and 11, respectively). Furthermore, a subsequent application of the Wilcoxon tests *within* the four groups of reasoning

styles yielded no significant mindset effects. For example, even within the group of subjects who conditioned themselves to think about the data set in terms of backward reasoning, the SB language did not perform better than the CF language. In conclusion, the reasoning style that the subjects adopted during training did not seem to influence the subsequent performance of the subject's belief models. This conclusion is qualified because the power of the Wilcoxon tests within each of the four mindset groups was low, due to the small number of subjects in each group.

Model Robustness: Even if SB and CF were compatible on mathematical grounds, they may still be different in terms of their capacity to handle inaccurate inputs, i.e. biased degrees of belief. First, the two belief languages use different elicitation procedures to obtain their inputs. Second, the two languages may be different in terms of robustness, or sensitivity to inaccurate inputs. A belief language is said to be robust if small perturbations in the degrees of belief that it operates on have little or no impact on the posterior beliefs that it generates. Robustness is a key property of belief languages, because it allows the human experts who specify the degrees of belief a certain margin of error that the language can tolerate without reversing the rankings of the hypotheses.

In constructing the experiment, we have tried to control for potential biases at the input stage by ensuring that the SB and the CF elicitation procedures would be on equal footing in terms of structure and contents. Although the *direction* of the elicitation was different, both procedures consisted of the same number of questions, and the questions themselves were designed to exert equal cognitive efforts from the subjects. Therefore, no procedure was predisposed to produce more biases than the other. However, as long as human experts are involved in specifying degrees of belief, the inputs that *any* elicitation procedure passes to its respective calculus are bound to be error prone. This was alluded to by Fischhoff, as follows: "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 [10]."

If we begin with the realistic assumption that subjective degrees of belief are bound to be biased, then a belief calculus that uses as *few* human-supplied

inputs as possible will be more robust than other calculi, all other things are held equal. With this property in mind, an inspection of SB and CF suggests that the former is markedly less robust than the latter. This is based on the simple observation that formula (1) requires twice as many degrees of belief as formula (4) to achieve precisely the same computational goal.

To illustrate, suppose the body of evidence at hand were $E = \{e_1, e_2\}$. The SB computation of $bel(h|E)$, which makes use of (1), depends on four elicited degrees of belief: $P(e_1|h)$, $P(e_1|\overline{h})$ $P(e_2|h)$, and $P(e_2|\overline{h})$. On the other hand, the CF computation of $bel(h|e)$ requires only two degrees of belief: if both $CF(h|e_1)$ and $CF(h|e_2)$ are positive (negative), the first (second) line of (4) is invoked. Otherwise, the third line of (4) is invoked. In either case, only two degrees of belief are necessary, compared to SB's four. The result generalizes to any number of pieces of evidence: CF is twice as parsimonious as SB in terms of reliance on elicited degrees of belief.

If SB is criticized for requiring too many inputs, one suggestion might be to elicit conditional likelihood ratios of the form $P(e|h)/P(e|\overline{h})$ directly, instead of pairs of probabilities $P(e|h)$ and $P(e|\overline{h})$, for each rule. However, conditional likelihood ratios are notoriously difficult to specify in a credible way. First, the expert is required to think about the relative likelihoods of observing e in light of the background hypotheses h and $\overline{h}$ *simultaneously.* Second, unlike probabilities, likelihood ratios are unbounded. Whether one uses verbal or graphical means to elicit them, it is very difficult to map subjective beliefs on a numeric $(-\infty, \infty)$ scale whose neutral point – the degree of belief associated with irrelevant evidence – is 1.

As was mentioned earlier in the paper, previous analyses of CF and SB established that the two calculi are isomorphic to each other under certain interscale transformations and independence assumptions. This functional equivalence doesn't seem to sit well with the observation that SB requires twice as many degrees of belief as CF to achieve the same computation. However, a simple manipulation of (1) reconciles the dilemma. First, recall that SB operates on diagnostic probabilities of the type $P(E_i|H)$, whereas CF operates on predictive certainty factors of the type $CF(H|E_i)$. Now, whichever form this relationship might take, there is no doubt that $CF(H|E_i)$ is strongly related to $P(H|E_i)$, as both measure the belief in H in light of a new piece of evidence

E. Invoking Bayes rule, we note that $P(E_i|H) = P(H|E_i) \cdot P(E)/P(H)$. If this transformation is applied to all the conditional probabilities $P(E_i|H)$ and $P(E_i|\overline{H})$ throughout (1), the following combination rule emerges after a few algebraic steps:

$$bel(H|E_1, E_2) = \frac{P(H|E_1, E_2)}{P(\overline{H}|E_1, E_2)} = \frac{P(H|E_1)}{P(\overline{H}|E_1)} \cdot \frac{P(H|E_2)}{P(\overline{H}|E_2)} \cdot \frac{P(\overline{H})}{P(H)} \qquad (5)$$

Both (1) and (5) compute exactly the same posterior belief, but the latter is half as dependent on elicited inputs as the former. This is because unlike $P(E_i|H)$ and $P(E_i|\overline{H})$, which are unrelated, $P(\overline{H}|E_i)$ is the complement of $P(H|E)$, so either probability can be used to determines the other. In conclusion, a predictive calculus like (5) has an advantage on a diagnostic calculus like (1) in terms of elicitation efficiency. It should be noted that from the technical viewpoint of this explanation, whether the predictive calculus is cast in terms of conditional probabilities or certainty factors is irrelevant, as long as both calculi rely on the same number of human-supplied inputs. Of course, either calculus may have other advantages or limitations in terms of validity that are unrelated to this particular observation.

7 Conclusion

The main finding of this research is as follows: in spite of the implicit assumption that CF and SB are ordinally compatible, the CF language performed better than the SB language in a controlled experiment involving human subjects and subjective degrees of belief. Although we don't have a concrete explanation to this superiority, we observe that the CF calculus is more parsimonious than the SB calculus, and thus less prone to biased degrees of belief. It is also possible that people find it easier to think and communicate about evidence and hypotheses in a forward reasoning style, consistent with the CF language, but this explanation was neither supported nor negated by the data.

The empirical results that were reported in this paper must be qualified by

the limited scope of the research. First, among the three main features of SB and CF – elicitation procedure, parallel combination, and sequential combination – only the first two played a role in the experiment. Second, the evidence/hypotheses space was consistent with a ratio-form conditional independence assumption, so the two calculi operated under ideal circumstances. Third, the two languages were compared to each other not across a wide range of inference problems with different characteristics, but rather in the context of a single inference task. It should be emphasized, though, that this task was highly generic. That is, once the credit analysis scenario is stripped away from the task, what remains is a typical belief revision pattern that emerges in practically every setting that involves rule based inference under uncertainty. In other words, in spite of its specific *appearance*, the inference task that we used had a general *structure* that cuts across many different domains of application.

The study of belief languages will not be complete until we learn how to deal with rule bases that violate the independence assumptions that underlie such calculi as SB, CF, and the Dempster Shafer model. This research can proceed in three complementary directions. First, we can seek knowledge engineering techniques to detect dependencies and then eliminate or minimize them by adding more structure to the knowledge base [4]. Second, we can develop heuristic belief languages that deal with correlated evidence directly, although the general problem is essentially *NP*-hard [14]. Third, we can acknowledge that independence assumptions are rarely met in practice, and proceed to investigate the conditions under which some languages operate better than others when the assumptions are violated.

Once again, we advocate the use of a modular research strategy – one that isolates the studied feature from the rest of the language. If one conducts a sensitivity analysis that investigates the robustness of CF and SB under various violations of independence assumptions, there is no need to complicate the study further with human subjects and subjective degrees of belief. Instead, what is called for is a laboratory setting in which CF and SB are applied to *objective* degrees of belief that are drawn from simulated distributions whose independence (or lack thereof) is controlled and manipulated by the experimenter. This way, the ex-post performance of the languages can be attributed to one factor only – the extent to which the independence as-

sumptions were violated. The results of such simulations, combined with the results obtained from experiments with human subjects like the one reported here, can contribute to a fuller understanding of how to best utilize heuristic belief languages in the context of rule based inference under uncertainty.

Appendix: Normative Belief Revision Models

The objective of the paper was to study the empirical validity of two *heuristic* methods to revise beliefs in rule based systems. This appendix provides a brief overview of the *normative* approach to belief revision, using the inference network in figure 1 as a common example. Most normative models view such networks from a *causal perspective*, in which different hypotheses (e.g. diseases) cause different sets of pieces of evidence (e.g. symptoms). The causal relationships, which might involve several layers of propositions (e.g. syndromes) are modeled through rules of the form $H \xrightarrow{d} E$, where d is normally set to $P(E|H)$ – the probability of observing the piece of evidence E when H is known to obtain. In order to illustrate this model, we assume hereafter that the direction of the arrows in figure 1 is reversed, pointing downward throughout the network.

Note that except for the hypotheses and pieces of evidence located at the network's boundaries, the bulk of the network's topology is made up of interim propositions, or sub-hypotheses, e.g. S_1 and S_2 in figure 1. These propositions play two different roles in the construction of inference networks. Typically, they are used to represent proxy attributes, or named syndromes, that are part of the expert's terminology and inference rationale. In other cases they serve to reduce dependencies among correlated pieces of evidence, a notorious problem that challenges the validity of *all* multi-attribute inference models. Specifically, in order to construct a credible rule base, the knowledge engineer must ensure that the rules that the expert provides are as independent as possible in terms of their evidential impact on the hypotheses. If the evidence is correlated, the knowledge engineer can seek background propositions that explain out the correlation [4]. For example, it may be that E_1 and E_2 are *not* conditionally independent with respect to H, but *are* conditionally independent with respect to S_1. In the former case,

the inferential relationship between E_1, E_2, and H cannot be described accurately by two independent rules. In the latter case, the relationship between E_1, E_2, S_1 and H can be captured using three rules, as depicted in figure 1, provided that S_1 and S_2 are also conditionally independent with respect to H. Said otherwise, the topology of the network in figure 1 implies several independence assumptions, as we'll see shortly.

The remainder of this appendix illustrates a normative approach to computing the posterior belief in H in light of any body of evidence (fact base) $\{E_1, E_2\}$, subject to the network's topology in figure 1. Compared to heuristic calculi, the normative approach is unique in its strict reliance on the joint distribution function $P(H, S_1, S_2, E_1, E_2)$ – the mechanism that governs the joint occurrence of the five propositions that make up the network. In principle, one can invoke probability theory and show that the posterior belief $P(H|E_1, E_2)$ can be 'easily' obtained by conditioning and integrating $P(\cdot)$ in a certain way. The problem, of course, is that $P(\cdot)$ is typically unavailable for direct inspection, and that even if it were available, the proposed brute force computation would be exponentially inefficient. Pearl, who analyzed this problem in detail in [16], proposed a solution which is based on generating $P(\cdot)$ in a piece meal fashion, using the 'chain rule' of probability theory. For example, the above 5-place function $P(\cdot)$ can be described in 5!=120 equivalent ways, including the following expansion:

$$P(E_1, E_2, S_1, S_2, H) = \begin{array}{l} P(E_1|E_2, S_1, S_2, H)\cdot \\ P(E_2|S_1, S_2, H)\cdot \\ P(S_1|S_2, H)\cdot \\ P(S_2|H)\cdot \\ P(H) \end{array} \tag{6}$$

The relationship between this expression and the inference network depicted in figure 1 is subtle. Let $S(x)$ be the set of all propositions that cause x directly, i.e. $S(x) = \{y|y \rightarrow x\}$. If the network were a tree, then for all x, $S(x)$ would contain at most one node. In figure 1, however, we have $S(E_1) = \{S_1\}$, $S(E_2) = \{S_1, S_2\}$, $S(S_1) = \{H\}$, $S(S_2) = \{H\}$, and $S(H) = \emptyset$ (recall that the direction of the arcs is reversed, denoting a causal orientation). According to Pearl, if an inference network is constructed properly, then for

each node x the set $S(x)$ 'shields' x from any evidential influence emanating from nodes that are outside $S(x)$. That is, for every set of propositions $y_1, \ldots, y_n \notin S(x)$, we have $P(x|S(x), y1, \ldots, y_n) = P(x|S(x))$. The topology of the network in figure 1 translates this constraint into four conditional independence assumptions regarding $P(\cdot)$, e.g. $P(E_1|E_2, S_1, S_2, H) = P(E_1|S_1)$. When these assumptions are plugged into the right hand side of (6), the expression simplifies considerably into:

$$P(E_1, E_2, S_1, S_2, H) = \begin{array}{l} P(E_1|S_1)\cdot \\ P(E_2|S_1, S_2)\cdot \\ P(S_1|H)\cdot \\ P(S_2|H)\cdot \\ P(H) \end{array} \tag{7}$$

It is useful for future purposes to divide both sides of the equation by the prior on H, $P(H)$, yielding the expression

$$P(E_1, E_2, S_1, S_2|H) = \begin{array}{l} P(E_1|S_1)\cdot \\ P(E_2|S_1, S_2)\cdot \\ P(S_1|H)\cdot \\ P(S_2|H) \end{array} \tag{8}$$

In the general case, then, the joint probability of observing *all* the propositions 'below' a certain hypothesis is given by the product of all the degrees of belief between the hypothesis and the propositions. Several researchers, most notably Pearl [16] and Cooper [7], developed belief revision algorithms that can compute $P(H|E_1, E_2)$ under these assumptions. For example, Cooper's technique is based on computing $P(E_1, E_2|h)$, and then applying Bayes rule to reverse the two sides of the conditioning bar[2]. First, $P(E_1, E_2|h)$ is derived by integrating the left hand side of (8) as follows:

[2]Recall that upper case notation, e.g. H, stands for the propositional variable whose name is H, whereas h and $\overline{h}$ stand for the constant propositions 'H is known to be true' and 'H is known to be false', respectively.

$$P(E_1, E_2|h) = \begin{array}{l} P(E_1, E_2, s_1, s_2|h) + \\ P(E_1, E_2, \overline{s_1}, s_2|h) + \\ P(E_1, E_2, s_1, \overline{s_2}|h) + \\ P(E_1, E_2, \overline{s_1}, \overline{s_2}|h) \end{array} \quad (9)$$

At that point, the method takes advantage of the assumptions that are implicit in the network's topology. In particular, each of the summands of (9) is expanded and computed using (8). After summing up these probabilities, $P(E_1, E_2|h)$ is obtained. Following a similar procedure, $P(E_1, E_2|\overline{h})$ is also computed. Given that the prior probability of H is know, its posterior probability in light of any body of evidence $\{E_1, E_2\}$ can be computed through Bayes rule, as follows:

$$\frac{P(h|E_1, E_2)}{P(\overline{h}|E_1, E_2)} = \frac{P(E_1, E_2|h)}{P(E_1, E_2|\overline{h})} \cdot \frac{P(h)}{P(\overline{h})} \quad (10)$$

To sum up, the computation of (10) boils down through (9) and (8) to many elementary manipulations of the degrees of belief that parameterize the rules in figure 1. The procedure is simple, general, and, most importantly, consistent with probability theory. At the same time, it can be applied only to relatively small inference problems. This is because the number of summands in (9) is 2^m, m being the number of internal nodes (sub-hypotheses) in the network. In fact, Cooper has shown that the problem of updating probabilities in a general network is *NP*-hard [6]. In the case of singly connected networks, however, there exists a belief revision algorithm whose run time is polynomial with the size of the network [16].

In addition to its inherent computational limitations, normative Bayesian techniques also require massive amounts of data, leading to severe knowledge elicitation and storage problems. For example, consider the relationship between E_2, S_1, and S_2 in figure 1. In order to fully specify the probabilistic nature of this triplet, one must elicit four probabilities: $P(e_2|s_1, s_2)$, $P(e_2|\overline{s_1}, s_2)$, $P(e_2|s_1, \overline{s_2})$, and $P(e_2|\overline{s_1}, \overline{s_2})$ (the probabilities $P(\overline{s_2}|S_1, S_2)$ can be derived from $1 - P(\overline{s_2}|S_1, S_2)$). In general, the evidential relationship between a single node and its n parent nodes requires the elicitation of 2^n

conditional probabilities. These limitations, along with the fact that human belief revision methods are often inconsistent with Bayesian inference, have led many to consider the use of heuristic, rather than normative, belief revision models in expert systems.

References

[1] J.B. Adams. Probabilistic reasoning and certainty factors. In B.G. Buchanan and E.H. Shortliffe, editors, *Rule-Based Expert Systems*, pages 263–271, Addison-Wesley, 1984.

[2] B.G. Buchanan and E.H. Shortliffe, editors. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.* Addison-Wesley, 1984.

[3] R. Carnap. *Logical Foundations of Probability.* University of Chicago Press, 1954.

[4] E. Charniak. The Bayesian basis of common sense medical diagnosis. In *Proc. of the National Conference in Artificial Intelligence*, pages 70–73, 1983.

[5] C. W. Churchman. *The Design of Inquiring Systems.* Basic Books, 1971.

[6] G.F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(2-3):393–402, March 1990.

[7] G.F. Cooper. A diagnostic method that uses causal knowledge and linear programming in the application of Bayes' formula. *Computer Methods and Programs in Biomedicine*, 22:223–237, 1986.

[8] R.M. Dawes and B. Corrigan. Linear models in decision making. *Psychological Bulletin*, 1974.

[9] R.O. Duda, P.E. Hart, and N.J. Nilsson. *Development of a Computer-Based Consultant for Mineral Exploration.* Technical Report, SRI, 1977. International Projects 5821 and 6415.

[10] B. Fischhoff and R. Beyth-Marom. Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, 90(3), 1983.

[11] B.N. Grosof. Evidential information as transformed probability. In L.N. Kanal and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 272–294, North Holland, 1986.

[12] D.E. Heckerman. Probabilistic interpretation for mycin's certainty factors. In L.N. Kanal and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, North Holland, 1986.

[13] E.J. Horvitz, D.E. Heckerman, and C.P. Langlotz. A framework for comparing alternative formalisms for plausible reasoning. In *Proc. of the AAAI Conference, Philadelphia, PA*, pages 210–214, 1986.

[14] Lauritzen and Spiegelhater. Local computation with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society*, B50(2):157–224, 1988.

[15] W. Mendenhall, R.L. Scheaffer, and D.D. Wackerly. *Mathematical Statistics With Applications.* Boston, MA: Duxbury Press, 1981.

[16] J. Pearl. Fusion, propagation and structuring in belief networks. *Artificial Intelligence*, September, 1986.

[17] S. Schocken and T.W. Finin. Meta-interpreters for rule-based inference under uncertainty. *Decision Support Systems*, 6, 1990.

[18] S. Schocken and P. R. Kleindorfer. Artificial intelligence dialects of the Bayesian belief revision language. *IEEE Transactions on Systems, Man, and Cybernetics*, 19:1106–1121, 1989.

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

[20] D. Sheskin. *Statistical Tests and Experimental Design: A Guide Book.* New York, NY: Gardner Press, 1984.

[21] E.H. Shortliffe and B.G. Buchanan. A model of inexact reasoning in medicine. In B.G. Buchanan and E.H. Shortliffe, editors, *Rule-Based Expert Systems*, page 242, Addison-Wesley, 1984.

[22] S.S. Stevens. Measurement, psychophysics, and utility. In C.W. Churchman and P. Ratoosh, editors, *Measurements, Definitions and Theory*, Wiley, New York, 1959.

[23] A. Tversky and D. Khaneman. Judgement under uncertainty: heuristics and biases. *Science*, 185:1124–1131, 1974.

[24] L.A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning (iii). *Information Sciences*, 9:43–80, 1975.