Reasoning with Belief Functions: An Analysis of Compatibility

Judea Pearl
Computer Science Department, UCLA, Los Angeles, California

ABSTRACT

This paper examines the applicability of belief functions methodology in three reasoning tasks: (1) representation of incomplete knowledge, (2) belief updating, and (3) evidence pooling. We find that belief functions have difficulties representing incomplete knowledge, primarily knowledge expressed in conditional sentences. In this context, we also show that the prevailing practices of encoding if-then rules as belief function expressions are inadequate, as they lead to counterintuitive conclusions under chaining, contraposition, and reasoning by cases. Next, we examine the role of belief functions in updating states of belief and find that, if partial knowledge is encoded and updated by belief function methods, the updating process violates basic patterns of plausibility and the resulting beliefs cannot serve as a basis for rational decisions. Finally, assessing their role in evidence pooling, we find that belief functions offer a rich language for describing the evidence gathered, highly compatible with the way people summarize observations. However, the methods available for integrating evidence into a coherent state of belief capable of supporting plausible decisions cannot make use of this richness and are challenged by simpler methods based on likelihood functions.

KEYWORDS: ***belief functions, Dempster–Shafer theory, knowledge and evidence, nonmonotonic reasoning, conditional information***

1. INTRODUCTION AND A BRIEF REVIEW

In earlier publications (Pearl [27, 28]) I attempted to uncover the epistemological basis of the Dempster–Shafer belief functions (BF), so as to form a meaningful assessment of their applicability vis-à-vis the Bayesian approach.

Address correspondence to Judea Pearl, Computer Science Department, UCLA, Los Angeles, CA 90024-1596.

International Journal of Approximate Reasoning 1990; 4:363–389

655 Avenue of the Americas, New York, NY 10010 0888-613X/90/$3.50

With the help of a canonical model cast in terms of "probabilities of provability," I then identified some basic issues that must be addressed if the theory is to provide a viable formalism for evidential reasoning. The purpose of this paper is to provide an updated examination of these issues in light of recent discussions I have had with colleagues and coworkers, and in light of new observations made subsequent to these discussions.

1.1. Knowledge, Belief, and Evidence

To facilitate our discussion of belief functions and their applications, I will adopt the following distinction between knowledge, belief, and evidence. *Knowledge* encodes judgments about the general tendency of things to happen, *evidence* summarizes the impact of that which actually happened, whereas *belief* combines the two—it consists of assertions about a specific situation inferred by applying generic knowledge to a set of evidence sentences.[1] For example, the sentence "Birds fly" encodes knowledge about the general tendency of birds to enjoy flying capabilities. The sentence "Tweety is a bird" provides an evidence by summarizing certain observations made on a specific object named Tweety. The sentence "Tweety flies" represents a state of belief about Tweety's capabilities, based on both generic knowledge (about birds) and a specific evidence (about Tweety).

Of course, all sentences are subjective and hence could be annotated with various shades of confidence. Note that evidence sentences and belief sentences both refer to specific situations and may rely on both generic knowledge and specific observations. The distinction between the two is that the process by which knowledge and observation were combined is made explicit in the case of belief sentences and remains implicit in the case of evidence sentences. For example, "Tweety is a bird" will be treated as an evidence sentence if it is fed as an input to a system that reasons about the habits of birds, but will be treated as a belief sentence by a system that reasons about the classification of animals from more rudimentary observations.

1.2. The Belief Function Formalism

Belief functions result from assigning probabilities to sets rather than to the individual points, with points representing specific worlds and sets reflecting

[1] These definitions of knowledge, evidence, and belief correspond to the common distinction in AI systems between "domain knowledge," "ground facts," and "inferences." The distinction is made solely for the purpose of disambiguating future discussions; I do not claim that people do not possess other forms of knowledge (e.g., procedural or normative) or that they do not occasionally cross the boundaries of these definitions. I define evidence as a *summary* of observations, so as to allow testimonies that, instead of describing raw observations, assess the impact of undisclosed observations on questions of interest, for example, "This object is very likely to be a bird."

propositions about those worlds. Given an initial probability assignment $m(\cdot)$ to a select set F of propositions (called *focal elements*), namely,

$$\sum_{B \in F} m(B) = 1, \qquad m(B) \geq 0 \tag{1}$$

every proposition in the language then acquires a pair of measures, $\text{Bel}(\cdot)$ and $\text{Pl}(\cdot)$, such that

$$\text{Bel}(A) = \sum_{B \text{ implies } A} m(B) \tag{2}$$

and

$$\text{Pl}(A) = 1 - \text{Bel}(\neg A)$$

Any measure $\text{Bel}(\cdot)$ constructed in such a manner is called a *belief function*, and its associated measure $\text{Pl}(\cdot)$ is called *plausibility*.

A necessary and sufficient condition for a function $\text{Bel}(\cdot)$ to be a belief function is that it satisfies

$$\text{Bel}(\varnothing) = 0, \qquad \text{Bel}(A \vee \neg A) = 1$$

and

$$\text{Bel}(A_1 \vee \cdots \vee A_n) \geq \sum_i \text{Bel}(A_i) - \sum_{i<j} \text{Bel}(A_i \wedge A_j) + - \cdots \tag{3}$$

$A_1, A_2, \ldots, A_n$ being any collection of propositions.

Given two belief functions Bel_1 and Bel_2, their orthogonal sum $\text{Bel}_1 \oplus \text{Bel}_2$, also known as Dempster's rule of combination, is defined by their associated probability assignments

$$(m_1 \oplus m_2)(A) = K \sum_{A_1 \wedge A_2 = A} m_1(A_1) m_2(A_2), \qquad A \neq \varnothing \tag{4}$$

where

$$K^{-1} = \sum_{A_1 \wedge A_2 \neq \varnothing} m_1(A_1) m_2(A_2) \tag{5}$$

The operator $\oplus$ is known to be commutative and associative.

As a special case of Eq. (4), if m_2 establishes the truth of proposition B, that is, $m_2(B) = 1$, then the combined belief function becomes

$$\text{Bel}_1(A \mid B) = \frac{\text{Bel}_1(A \vee \neg B) - \text{Bel}_1(\neg B)}{1 - \text{Bel}_1(\neg B)} \tag{6}$$

This formula is known as Dempster's conditioning.

A belief function is called *additive* or *Bayesian* if each of its focal elements is a singleton, that is, an elementary event or a possible world. Bayesian belief functions satisfy $\mathrm{Bel}(A) = \mathrm{Pl}(A) = 1 - \mathrm{Bel}(\neg A)$. If Bel_1 is Bayesian, then $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ is also Bayesian, and Dempster's conditioning reduces to ordinary Bayesian conditioning (Shafer [30]).

1.3. The Methodology of Belief Functions

The methodology according to which belief functions have been applied to reasoning problems consists of three separate components:

1. *Representation of incomplete knowledge*. In cases where fully specified probabilistic knowledge is not available, belief functions are often used to represent states of partial knowledge, with $\mathrm{Bel}(A)$ interpreted as a strength of arguments in favor of A. The input information is often elicited by assessing the highest probability one is willing to commit to A and to its complement $\neg A$, and encoding these assessments as $m(A)$ and $m(\neg A)$, respectively.
2. *Belief updating*. Belief functions offer a method of assimilating the impact of new evidence into a state of partial knowledge or partial belief. If the initial state is encoded as a belief function $\mathrm{Bel}(\cdot)$ and the evidence as a belief function $\mathrm{Bel}_e(\cdot)$, then the updated state of belief $\mathrm{Bel}'(\cdot)$, accounting for the impact of the new evidence, is given by Dempster's combination $\mathrm{Bel}' = \mathrm{Bel} \oplus \mathrm{Bel}_e$.
3. *Pooling of evidence*. When multiple pieces of evidence are available, the BF combination method provides a faithful summary of the information carried by all the individual pieces. The method consists of encoding each piece of evidence as a belief function and combining these functions by Dempster's rule of orthogonal sum. Dempster's rule, Eq. (4), reflects two assumptions about evidence combination:
   a. Independence. Successive pieces of evidence are independent of each other.
   b. Renormalization. Whenever two pieces of evidence impart weights to contradictory propositions [$A_1 \wedge A_2 = \varnothing$ in Eq. (4)], that weight is redistributed among the noncontradictory propositions, proportionally to their weights.

In subsequent discussions I will deal separately with these three components of BF methodology. In Section 2, I will examine whether the representational scheme offered by BF adequately represents common types of incomplete knowledge. In Section 3, I will focus on those cases where the BF representation is adequate, and will evaluate its belief-updating component. My conclusions (Section 4) are that belief functions have serious difficulties serving in these two functions, and their potential of serving in the third function, evidence pooling, is limited to cases where the evidence thus pooled will ultimately be used to update an ordinary (i.e., additive) probability measure or a purely categorical knowledge state. First (Section 1.3), I would like to recount an interpretation of belief functions that I have found helpful in understanding the workings of the theory and its potential applications.

1.4. A Probabilistic Interpretation of Belief Functions

Historically, belief functions were first interpreted as lower and upper probabilities induced by a special family of probability distributions (Dempster [7]). Each member of the family represents a redistribution of the sets' probabilities $m(\cdot)$ to individual points within those sets, and each (Bel(A), Pl(A)) pair provides bounds on the "true" probability of A. This interpretation, still a favorite of many researchers (Kyburg [20], Fagin and Halpern [11]), is unsatisfactory when it comes to evidence combination; the correspondence between the constituent families of probabilities and the family resulting from Dempster's rule defies a natural explanation (see Section 3.1). Shafer's [30] interpretation of Dempster's theory abandons the idea that belief functions arise as lower bounds of a family of ordinary probability distributions; rather, it views belief functions as a natural representation of English words such as "belief," "doubt," "evidence," and "support,"' and Dempster's rule of combination as the fundamental step in reasoning about uncertain evidence. In Shafer's view, "belief" in a hypothesis does not measure the chance that it is true, but rather the strength of the arguments we have in favor of the hypothesis.

In translating Shafer's view into the language of logic, using the notions of proofs and constraints (or axioms), it becomes clear that belief functions are none other than *probabilities of provability* (Pearl [27], p. 423). In other words, Bel(A) stands for the probability that the constraints imposed by the available evidence, together with the constraints that normally govern the domain, will be sufficient to compel the truth of A and exclude its negation. Formally, we are given a collection of logical theories, $T_1,\ldots,T_n$, each logical theory is characterized by a set of formulas called axioms, and each theory is assigned a probability P_i such that the probabilities sum to 1. The belief in a formula A is the sum of the probabilities of the theories from which A follows as a logical consequence.[2] (I am indebted to Fagin and Halpern [11] for this succinct wording.) This interpretation has the advantage that Bel(A) no longer represents properties of some nebulous family of probability functions but rather an ordinary probability of a bona fide event—the existence of a log-

[2] Each theory corresponds to one focal element B in (1), and its probability P_i corresponds to $m(B)$. If one associates with each theory the set of worlds (or models) consistent with its defining axioms, the resulting theorems can be described as a system of random sets (Nguyen [25]). Our notion of "provability" is semantical, resting on logical consequences, to be distinguished from syntactical proofs, which are normally embedded in a specific axiomatization.

ical proof for A. The reason that this event is random is that the set of axioms from which the proof is to be assembled is not known with certainty, only with probability. Note that both a formula and its negation might have belief 0, since neither might follow from any of the theories.

This interpretation of belief functions now provides a simple semantics for Dempster's rule of evidence combination, Eq. (4). Each piece of evidence, say e_1 and e_2, defines a probability mass over a collection of theories, and the combined evidence $e_1 \oplus e_2$ likewise defines a probability mass over a collection of joint theories. Each joint theory is made up of a union of two axiom sets, one taken from e_1 and one from e_2. The mass assigned to such a union is the product of the individual masses (thus reflecting evidence independence), while the mass attributed to any contradictory theory is redistributed among the noncontradictory theories in proportion to their weights. Thus, the belief function resulting from this combination rule is simply the *conditional probability of provability*, given that the two pieces of evidence are noncontradictory.

The advantage of this interpretation is that it enables us to discuss in meaningful terms whether Dempster's rule is applicable: namely, whether in the context of a specific domain it is meaningful to condition beliefs on the evidence being noncontradictory. Shafer's canonical model of random codes (Shafer [32]) is an example of a context where the assumption of noncontradictory evidence is a reasonable one to make. Here messages were presumed to originate from a consistent source, and so the event of two deciphered messages yielding two incompatible conclusions can be ruled out as a physical impossibility.

What remains unjustified in this example is why we should concern ourselves with "the probability that the evidence implies A" (Shafer [32]) rather than the probability that A is true given the evidence. Motivating this computation is one unresolved difficulty with the semantics of belief functions. For example, in circuit diagnosis we normally ask for the probability that a particular component *is* faulty, not that it *must be* faulty for lack of another explanation. Taking the latter as a basis for decisions might yield undesirable policies of testing or replacing faulty components. The lack of justification for this component in Shafer's random code example was also criticized by Good [15] and Levi [22], and more of its consequences will be discussed in Section 3.2.

However, there are situations where it is indeed the probability of necessity and possibility that we wish to ascertain, rather than that of truth, and it is within such situations that we should seek canonical examples for the construct of belief functions. Suppose, for example, we are scheduling a construction project, and our body of evidence consists of a set of estimates $m_1, m_2, \ldots, m_n$, where each m_i stands for the chance that resource r_i will become unavailable during the construction. In this context, it is not uncommon to inquire as to the probability that a certain decision (say changing a design option) will *have to* be made out of a compelling necessity, for lack of any viable alternative, rather than the probability that it will *actually* be made, as an option of choice or convenience.

Unfortunately, evidence inconsistency in this context means the impossibility of meeting all scheduling constraints, not a physical impossibility, and cannot be ruled out a priori as in the diagnostic problems discussed earlier. Thus, it would not be justified to apply Dempster's rule of combination in such applications.

My attempts to find natural applications for belief functions have encountered the following pattern. In tasks of synthesis and prediction, where we sometimes have reasons for computing the probability of necessity (e.g., the scheduling problem), we normally find no justification for assuming evidence consistency. In tasks of diagnosis or abduction, where we normally find justification for assuming evidence consistency (e.g., Shafer's random code or the Three Prisoners puzzle, Section 3.1), we can find little justification for computing the probability of necessity. The only example I found that motivates both components together and thus reflects precisely the quantity computed by Bel is the retrospective version of the scheduling problem, where, years after the completion of the project, one asks: "What is the probability that a certain decision had to be made out of absolute necessity?" or, alternatively, "What is the probability that the person in charge *knew* that a certain event must have occurred, out of logical necessity?"

2. DO BELIEF FUNCTIONS FACILITATE AN ADEQUATE REPRESENTATION OF INCOMPLETE KNOWLEDGE?

By "incomplete knowledge" I mean knowledge that is insufficient for fully specifying a probability distribution over the set of possible worlds. Indeed, the most appealing feature of BF vis-à-vis the Bayesian approach has been that the latter requires the full specification of a joint distribution function over the variables in the domain while "using belief functions, one need not estimate any probabilities that are not readily available" (Strat [36]). Thus, if only some probabilities are known, but not all, one obvious semantics of our knowledge would be an implicit *family* of probability functions, a family that contains all functions that comply with what we know but makes no commitment relative to what we do not know. Alternative semantics are also possible (e.g., the maximum-entropy approach), according to which the missing probabilities can be recovered. In this section I will discuss the relationships between belief functions and families of distributions that reflect naturally occurring cases of incomplete knowledge. The discussion should also be relevant to those who disavow any relation between belief functions and uncertain probabilities (e.g., Shafer [31]), because the emphasis will not be on whether belief functions can represent such families with full numerical precision but on whether the two representations can adequately capture the meaning intended by the knowledge provider.

2.1. The Problem of Representing Unknown Probabilities

It is well known (Dempster [7], Shafer [31], Kyburg [20]) that although every belief function represents a (convex) family of probability distributions, the converse is not true; not every family of distributions is compatible with a coherent system of [Bel, Pl] intervals. I will now argue that the more common state of incomplete knowledge corresponds to an implicit family of distributions of the latter kind, whereas states of knowledge representable as belief functions are rather rare.

EXAMPLE 1 (After N. Dalkey) Consider a set of three mutually exclusive and exhaustive events, E_1, E_2, and E_3, and assume we do not know the exact probabilities of these events. We know only that each event has a probability smaller than 1/2, that is,

$$0 \leq P(E_i) \leq 1/2, \qquad i = 1, 2, 3 \tag{7}$$

It can be shown that no belief function can induce the bounds

$$\text{Bel}(E_i) = 0 \qquad \text{and} \qquad \text{Pl}(E_i) = 1/2, \qquad i = 1, 2, 3 \tag{8}$$

as dictated by Eq. (7). In order words, any assignment Bel$(\cdot)$ that satisfies (8) must violate the basic definition of belief functions given in (3).

Admittedly, probability bounds such as those in (7) do not represent naturally occurring types of incomplete knowledge, because one can seldom be precise about probability intervals. Nevertheless, the example shows how easy it is for a family of distributions to escape BF representation.

EXAMPLE 2 We have two events, E_1 and E_2. We know their marginal probabilities,

$$P(E_1) = P(E_2) = 1/2 \tag{9}$$

but we know nothing about the joint probabilities. This state of partial knowledge is more common, because we often begin thinking about a problem through isolated frames, paying no attention to interdependencies. This case, too, does not admit belief function representation. The information given in (9) permits the probability of each of the four joint events $\{E_1 \wedge E_2, E_1 \wedge \neg E_2, \neg E_1 \wedge E_2, \neg E_1 \wedge \neg E_2\}$ to range between 0 and 1/2, which translates to a Bel measure of zero. However, no belief function can assign a zero belief to four individual points and, simultaneously, a belief of 1/2 to four pairs of these points [as required by (9)].

Note that the problem in these examples lies not with Dempster's rule of combination but rather with the inherently limited expressiveness of the BF language. In terms of our probability of provability model, the second example means that there is simply no way to assign probabilities to axioms such that four mutually exclusive and exhaustive propositions will each have zero chance of being proved true while each of four pairwise disjunctions retains a 50% chance of being proved.

EXAMPLE 3 We have two independent events, E_1 and E_2. We know the marginal probability of E_1,

$$P(E_1) = 1/2$$

but we know nothing about E_2 (except its being independent of E_1). The reader might expect that this state of ignorance would be ideal for BF representation, since E_2 is merely superfluous and irrelevant. However, the knowledge we possess about E_1 dictates definite probabilities on certain formulas involving E_2. For example,

$$P[(E_1 \wedge E_2) \vee (\neg E_1 \wedge \neg E_2)] = 1/2$$

and

$$P[(E_1 \wedge \neg E_2) \vee (\neg E_1 \wedge E_2)] = 1/2$$

The former expresses the probability that the true values of E_1 and E_2 are the same, the latter that they are different. However, no belief function exists that is compatible with these equalities while simultaneously reflecting our state of ignorance about E_2, namely, $\mathrm{Bel}(E_2) = 0$. Example 5 (Section 3.2) describes a situation where failing to represent $\mathrm{Bel}[(E_1 \wedge E_2) \vee (\neg E_1 \wedge \neg E_2)] = 1/2$ leads to a major clash with intuition.

2.2. The Problem of Representing Conditional Information

EXAMPLE 4 Consider the very common case where we are given the conditional probabilities

$$P(A|B) = p \qquad \text{and} \qquad P(A|\neg B) = q, \qquad 0 < 1 - q \leq p \leq q < 1 \quad (10)$$

but we are not given any of the prior probabilities.[3] The information given in

[3] This form of incomplete knowledge was discussed by Shafer and others under the heading "generalization of Bayesian parametric inference" (see Shafer [32]). The word "parametric" might create the impression that the difficulties originate with one's insistence on fitting frequency information by precise numeric parameters. Section 2.3 will demonstrate that these translate to fundamental qualitative difficulties in handling partial knowledge of any sort, not merely statistical models with missing parameters.

(10) induces several constraints on the probabilities of other propositions in this frame, including, for example,

$$0 \leq P(A \wedge B) \leq p$$

$$p \leq P(A) \leq q$$

$$1 - q \leq P[(A \wedge B) \vee (\neg A \wedge \neg B)] \leq p$$

Again, no belief function exists that matches these upper and lower probabilities without violating the basic conditions in (3).

I consider this latter example to be a major shortcoming of BF theory. Conditional information is, in my opinion, the basic building block in the organization of human knowledge. It is this information that enables us to sort experiences and store them in the right context, with the conditioning propositions (*context* or *reference classes* as they are sometimes called) serving to identify the conditions under which the future use of this experience would be appropriate. It is for that reason that conditional sentences[4] are so common in natural discourse and if-then rules are so popular in the design of AI reasoning systems. Therefore, it is a common practice to find an expert feeling very strongly about conditional probabilities (e.g., that a given symptom will accompany a given disease) while having no feeling at all about the priors. Unfortunately, this type of incomplete knowledge cannot be captured by belief functions.

The source of this weakness lies deep in the very nature of BF theory. Recall that the theory rests on assigning probabilities to sets of logical formulas (i.e., theories). Thus BF is essentially a theory of randomized monotonic logic, and, like every monotonic logic, it lacks a mechanism equivalent to conditioning with which context dependencies can be encoded. Indeed, it is well known that the information conveyed in a conditional probability statement

$$P(A|B) = p$$

cannot be represented by assigning probabilities to some Boolean function of A and B or to any set of Boolean formulas (Lewis [23], Goodman [16]).

Example 4 demonstrates that if our knowledge consists of conditional probabilities and we are missing the necessary priors, this knowledge cannot be adequately represented as a belief function. However, even if we are given the

[4] By conditional sentences I mean utterances such as "Birds fly," "Fire causes smoke," "Smoke suggests fire," "When it rains, it pours," and many others, that do not concern frequency information as in "Most of the people in Lawrence are white." Although the word "if" is not mentioned explicitly, the common denominator is that the assertions made by such sentences are context-dependent; they are applicable to a narrow context circumscribed by the conditioning phrase. It is the encoding of such sentences that we examine in this example.

priors, the problem of inadequate representation might arise as soon as some conditional probabilities are missing from the model. The reason is that if our knowledge is organized hierarchically, then the conditional probabilities at one level of the hierarchy determine the priors of the next level, where the problem demonstrated by Example 4 might recur.

Other types of partial knowledge that render BF encoding cumbersome, if not impossible, or comparative ratings of probabilities and qualitative conditional independence relationships. For example, suppose we are given the information that $P(A \wedge B)$ is at least twice as large as $P(A \wedge C)$ and that A and B are independent given C. Can this information be encoded as belief functions? We know of no effective procedure for deciding such questions. Currently, the only procedure for determining the existence of belief function representation is the inequality of Eq. (3), which is computationally infeasible because it requires the enumeration of all bounds that follow from the database. Can the information be approximated by belief functions while preserving its conceptual intent? I know of no research attempting to answer such questions.

2.3. The Material Approximation

The reader may wonder at this point how BF practitioners, lacking a conditioning facility with which to express inexact if-then rules, have nevertheless managed to construct rule-based expert systems. It turns out that the prevailing practice in these systems has been to represent the rule "If A then B" by the material implication formula $A \supset B$ $(= \neg A \vee B)$ and assign to this formula some weight w that purports to measure the strength of the rule or its validity, thus converting the rule into a bona fide belief function satisfying $m(A \supset B) = w$. This practice is not entirely without merit. For example, combining the resulting belief function with the evidence A = *true* does give the expected result $\text{Bel}(B|A) = w$. Moreover, if we are given a full specification of a joint probability, say in the form of a causal network (Pearl [27]), we can replace every conditional probability $P(x|\text{causes-of-}x) = q$ by its material implication counterpart $m(\text{causes-of-}x \supset x) = q$ and combine these $m(\cdot)$ functions using Dempster's rule, and the result would be a belief function that is equivalent to the original probability model. The problems begin when the probability model is incomplete and some of the conditional probabilities (or the priors) are missing. In such cases, the material implication scheme may yield very undesirable effects, three examples of which are shown next.[5]

[5] Another method of encoding conditional rules was proposed by Smets [35] and named *conditional embedding*. The idea is to represent each given rule by the least committed belief function Bel_0 that satisfies $\text{Bel}(A|B) = w$ and then combine the Bel_0 functions by Dempster's rule. By "least committed," he means that of all possible belief functions, Bel_0 allocates the least possible value for every proposition. This approximation, as well as others (Eddy and Pei [10]), have apparently not been as popular among practitioners as the material implication. Nevertheless, the difficulties demonstrated in this section apply to both.

CHAINING The facility of chaining inferential links, sometimes known as *transitivity*, is inherent to rules expressed as material implication formulas; from $A \supset B$ and $B \supset C$ there follows $A \supset C$. When uncertainties are attached to such rules, the conclusion $A \supset C$ is weakened but still obtains a positive measure of support (equal to the product of the supports given to the individual rules). On many occasions, however, rule transitivity must be totally suppressed, not merely weakened, or strange results will surface. One such occasion occurs in property inheritance, where subclass specificity should totally override superclass properties. The rule "Typically penguins don't fly" should not be weakened by adding the two rules "Penguins are birds" and "Typically birds fly," regardless of how strongly we believe in the latter two. Another occasion occurs in causal reasoning, where predictions should not trigger explanations; for example, "Sprinkler was on" predicts "Ground is wet," "Ground is wet" suggests "It rained," yet "Sprinkler was on" should not suggest "It rained." In such cases, softening the rules by probability assignments and combining them by BF methods does not produce the expected behavior (Pearl [27], pp. 447–450]; it weakens the flow of inference through the rule chain but does not bring it to a dead halt as it should.

CONTRAPOSITION The formula $\neg B \supset \neg A$ is logically equivalent to $A \supset B$, so the two will acquire the same Bel measure in any formulation based on belief functions. Yet, in many cases we are ready to accept the rule "If A then B" but are unwilling to accept "If *not-B* then *not-A*." For example, in a world that is made up primarily of birds (some of which might be sick), I am willing to accept the rule "Typically birds fly" but unwilling to accept "Typically nonflying things are non-birds," because nonflying things in this world are likely to be sick birds.

The symmetrical nature of the material implication may lead to counterintuitive inferences,[6] especially when the negation of the consequent is established by other rules and the rules convey causal relationships. Consider the rules

$I(x) \rightarrow P(x)$: If a person is intelligent, then that person is popular.

$F(x) \rightarrow \neg P(x)$: If a person is fat, then that person is unpopular.

together with our cultural bias that intelligence is independent of physical appearance. Now assume that each of the two rules has a strength m and we learn that Joe is fat. Formulating this information in BF terminology (using the material implication) produces the strange result that Joe is believed to be not intelligent with strength m^2. Most people would agree that it is reasonable to believe (with degree m) that Joe is unpopular, yet it is absurd to chain this

[6] These are further magnified when we consider criteria by which conditional rules are confirmed by empirical data (see Lewis [23] and Goodman [17]).

belief with the contraposition of the first rule and conclude (with belief m^2) that Joe is not intelligent.

REASONING BY CASES Suppose we are given the following two rules:

$$\begin{aligned} &\text{If } A \text{ then } B, \text{ with certainty } 0.9.\\ &\text{If } notA \text{ then } B, \text{ with certainty } 0.7. \end{aligned} \tag{11}$$

Common sense dictates that even if we do not have any information about A we should still believe in B to a degree at least 0.7. The BF formulation does not support this intuition. If we interpret (11) as statements of conditional probabilities, then this information cannot be represented in terms of belief functions (see Section 2.2, Example 4). If we try to force a BF encoding by interpreting the rules as material implications, we obtain $\text{Bel}(B) = 0.63$, in clear violation of common sense.

To appreciate the danger of casting rules as graded material implications, consider the following (hypothetical) database, rating the commitments of three religions toward observing the biblical prohibition on eating pork:

$$\begin{aligned} \text{Jew}(x) &\rightarrow \text{Observe}(x) \quad (0.70)\\ \text{Muslim}(x) &\rightarrow \text{Observe}(x) \quad (0.90)\\ \text{Christian}(x) &\rightarrow \text{Observe}(x) \quad (0.001) \end{aligned} \tag{12}$$

Suppose we know that Joe is either a Jew or a Muslim; does he refrain from eating pork? The answer we get from the material approximation is rather strange: yes, with belief 0.63. Further yet, imagine that instead of the first rule in (12), we have a more detailed account of Jews' commitment to dietary laws, as given by the table[7]

$$\begin{aligned} \text{Orthodox-Jew}(x) &\rightarrow \text{Observe}(x) \quad (0.999)\\ \text{Conservative-Jew}(x) &\rightarrow \text{Observe}(x) \quad (0.80)\\ \text{Reformed-Jew}(x) &\rightarrow \text{Observe}(x) \quad (0.40)\\ \text{Nonaffiliated-Jew}(x) &\rightarrow \text{Observe}(x) \quad (0.20) \end{aligned}$$

How strongly are we now willing to believe that Joe would refrain from eating pork? The answer produced by the material approximation is equal to the product of all the probabilities listed in Joe's category, giving a minute belief of

[7] The numbers in this table should not be construed as reflecting statistical data but, rather, subjective assessments of the degree to which each antecedent provides an argument for the consequence.

only 0.0575. Thus, the more we know about Jewish customs, the less willing we are to endow Joe with any trait of Jewish tradition.[8]

CAN THE MATERIAL APPROXIMATION BE IMPROVED? Strangely, despite the weaknesses demonstrated in the examples above, the material implication formulation of if-then rules still permeates the practice of BF systems (Ginsberg [14], Baldwin [2], Falkenhainer [13], D'Ambrosio [5], Lowrance et al. [24], Zarley et al. [37], Biswas and Anand [3], Laskey et al. [21], Hsia and Shenoy [19]). In some systems (e.g., Laskey et al. [21]), special provisions are installed to prevent undesirable chaining. In others (e.g., Ginsberg [14]), metarules are proposed to enforce specificity. This brings into question whether a theory requiring the protection of special programming fixes provides an adequate account of evidential reasoning. The difficulty is further compounded by the fact that we do not know in advance when such fixes are necessary. Given a partially specified knowledge base, we do not have any guidance for deciding whether it can be represented by belief functions or whether the material implication formulation will provide a reasonable approximation of the knowledge available.

A natural question to ask is whether the material approximation can be improved. All indications are that every BF encoding is bound to encounter difficulties similar to those shown above; the very idea of encoding if-then rules as individual belief functions to be combined by a uniform principle seems to be incompatible with the information conveyed by the rules. This can be seen from the penguin example of Section 2.3.1. Suppose we know that Tweety is both a penguin and a bird, and we wish to assess her flying capability. Any method based on belief function combination will remain oblivious to the second rule, stating that all penguins are birds. Knowing that Tweety is both a penguin and a bird renders Bel(*Tweety flies*) solely a function of rule 1, "Typically penguins do not fly," and rule 3, "Typically birds fly," regardless of whether penguins are a subclass of birds or birds are a subclass of penguins. This stands contrary to common discourse, where people expect class properties of be overriden by properties of more specific subclasses.

In conclusion, whereas it is true that "with belief functions, one need not estimate any probabilities that are not readily available" (Strat [36]), it is equally true that with belief functions, one cannot use some probabilities that are available, especially those embodied in if-then rules.

3. BELIEF UPDATING

In this section we focus on those cases where partial knowledge does lend itself to BF encoding. In such cases the bounds provided by belief functions

[8] Readers preferring more "practical" examples are invited to exercise this paradox on the cholestatic jaundice example of Gordon and Shortliffe [18]. Assume that we know the tendency t_1 of hepatitis to develop a complication C and the tendency t_2 of cirrhosis to develop C, and we wish to assess the tendency t of a patient known to be definitely suffering from intrahepatic cholestasis, that is, from either hepatitis or cirrhosis. The answer produced will be $t = t_1 t_2$.

coincide exactly with the envelope of the family of distributions that mirrors the input information. We will see that this fortunate coincidence is destroyed as soon as we attempt to update knowledge using Dempster's rule.

3.1. The Mystery of the Vanished Interval

Let $\mathcal{P}$ be a family of distributions compatible with a belief function $\mathrm{Bel}(\cdot)$, and assume that we receive information that event B has occurred. The natural way of assimilating this information would be to replace each distribution P in $\mathcal{P}$ by $P(\cdot|B)$, thus obtaining a new set of bounds on the propositions of concern. On the other hand, if we treat the new information as a belief function $\mathrm{Bel}'(B) = 1$ and combine it with $\mathrm{Bel}(\cdot)$ by Dempster's rule, we get [see Eq. (6)]

$$\mathrm{Bel}(A\,|B) = \frac{\mathrm{Bel}(A \vee \neg B) - \mathrm{Bel}(\neg B)}{1 - \mathrm{Bel}(\neg B)} \tag{13}$$

It was noted by Dempster [7], Shafer [31], and Kyburg [20] that the bounds produced by Dempster conditioning [i.e., by Eq. (13)] are never wider, and are occasionally narrower, than those produced by straight conditioning. Namely,

$$\min_{P\in\mathcal{P}}[P(A\,|B)] \leq \mathrm{Bel}(A\,|B) \leq \mathrm{Pl}(A\,|B) \leq \max_{P\in\mathcal{P}}[P(A\,|B)] \tag{14}$$

What perhaps has not received sufficient emphasis is that the $\mathrm{Bel}(A\,|B)$ and $\mathrm{Pl}(A\,|B)$ bounds could become so absurdly narrow that one should seriously doubt whether they ever represent meaningful information.[9] A classical example of such extreme narrowing is the Three Prisoners puzzle, first discussed in Diaconis [8] and later in Diaconis and Zabell [9] and Pearl [27]. To summarize formally, the partial model is given by

$$P(A_i) = 1/3, \qquad i = 1, 2, 3$$

$$P(B|A_2) = 0, \qquad P(B|A_3) = 1, \qquad P(B|A_1) = \text{unknown} \tag{15}$$

A_i stands for the event that prisoner i is the one found guilty, and B stands for the event that a jailer names prisoner 2 as one who has been found innocent. The probability bounds defined by this information are

$$0 \leq P(A_1|B) \leq 1/2 \tag{16}$$

[9] One should note that the bounds produced by straight conditioning are also inadequate because, in practice, they turn out to be very broad and not very informative. They do help, however, identify sensitive areas of incompleteness in the knowledge base.

while the belief function formulation gives [taking $m(A_i) = 1/3$, $i = 1, 2, 3$]

$$\mathrm{Bel}(A_1|B) = 1/2 = \mathrm{Pl}(A_1|B)$$

$$\mathrm{Bel}(A_1|\neg B) = 1/2 = \mathrm{Pl}(A_1|\neg B) \tag{17}$$

This example demonstrates that by encoding a state of partial knowledge as a belief function and conditioning it on a certain body of evidence, the gap between Bel(·) and Pl(·) could mysteriously disappear, thus giving the false impression that Bel(·) is based on precise probabilistic information, when such information is in fact not available. Any confidence interval interpretation one attempts to ascribe to the original belief function is destined to be very short-lived; as soon as we receive a piece of evidence, the correspondence between the original family of distributions and the one resulting from Dempster's combination might forever be destroyed. Indeed, Eq. (17) shows that the family of distributions defined by the partial model of Eq. (15) shrinks to a singleton distribution:

$$P(A_1 \wedge B) = 1/2, \qquad P(A_3 \wedge B) = 1/2$$

as soon as the evidence $B = \mathit{true}$ is obtained. Why such an innocuous piece of evidence should bring about the total exclusion of all other distributions, perfectly consistent with the evidence, remains a total mystery.

3.2. The "Spoiled Sandwich" Effect

One can perhaps dismiss the strange phenomenon demonstrated by the Three Prisoners story and claim that the interval [Bel, Pl] was never meant to represent upper and lower probabilities, but rather a new mental construct that conforms to its own norms of coherence, unshared by probabilities. I would be inclined to accept such arguments had the quantity Bel computed in (17) been reflective of some distinct linguistic expression in natural usage, or had the difficulties encountered been confined to numerical disagreements with the predictions of probability theory. Unfortunately, the difficulties go much deeper, as they clash with some universal principles of plausible reasoning having nothing to do with probabilities. One such clash is demonstrated by the "spoiled sandwich" effect, which illustrates that Bel(·) does not possess qualities we normally associate with the English word "belief" and brings into question whether it can ever serve as a basis for decision making.

The "sandwich" metaphor is borrowed from Aleliunas [1] to describe the following principle of plausible reasoning:

> If two diametrically opposed assumptions yield two different degrees of belief in a proposition Q, then the unconditional degree of belief merited by Q should be somewhere between the two. (Pearl [27], p. 17)

It is by virtue of this principle that we agree, given the information in Eq. (11) (Section 2.3), that B should merit a belief somewhere between 0.9 and 0.7. In general, for any proposition A and for any assumption or evidence B, we should have

$$\mathrm{Bel}(A) \geq \min[\mathrm{Bel}(A\,|B), \mathrm{Bel}(A\,|\neg B)] \tag{18}$$

This principle, unfortunately, is not satisfied by belief functions, as we have seen in the Three Prisoners puzzle; the unconditional belief was $\mathrm{Bel}(A_1) = 1/3$ [see Eq. (15)], while the conditional beliefs $\mathrm{Bel}(A_1|B)$ and $\mathrm{Bel}(A_1|\neg B)$ were both 1/2 [see Eq. (17)].

The real culprit for this phenomenon is not Dempster's normalization but the very nature of $\mathrm{Bel}(A)$ being the probability of finding a proof for A. Under the random-theory model (Section 1.4), it is quite possible that A is not provable from any of the random theories available, thus rendering $\mathrm{Bel}(A) = 0$, but if we add either B or $\neg B$ as an axiom, then A will be provable by some (though not the same) theory, thus rendering both $\mathrm{Bel}(A|B)$ and $\mathrm{Bel}(A|\neg B)$ greater than zero. This is illustrated by the following example, where Dempster's normalization plays no role.

EXAMPLE 5 (The Peter, Paul, and Mary Sandwich) Mary challenges Peter to guess what kind of sandwich she happened to prepare for lunch that day—ham or turkey. She also promises to pay Paul \$1000 if Peter guesses correctly. Peter says that, for lack of even the slightest clue, he is going to toss a fair coin and guess "ham" if it turns up heads, "turkey" if it turns up tails. Mary asks Paul if he is not anxious to know what sandwich she actually prepared, but Paul brushes her off saying that he already had lunch and that it makes no difference to him; regardless of whether it is ham or turkey, in either case he has exactly a 50% chance of winning the \$1000.

Mary retorts that Paul is behaving like an incurable Bayesian, and that instead of considering the chances of winning he should be considering the chances that winning is *assured* by the specific evidence at hand, namely, by Peter's guessing policy. She claims that Paul's current "belief" of winning is, in fact, zero, because either outcome of the coin, heads or tails, would leave him with no assurance of winning. However, if he would only listen to her for a moment, his belief would immediately jump to 1/2, because knowing what kind of sandwich it is would give him a 50% assurance of winning.

Paul answers that he gets enough assurance just thinking about Mary's sandwich: "If I have a 50% assurance assuming it is ham, and 50% assuming it is turkey, then I have a 50% assurance, period!"

Why is the sandwich principle so important for rational thinking? The answer lies, I believe, in its role as a guide to decision making and knowledge acquisition. Translated to decision situations, the sandwich principle leads to the "sure-thing principle":

> If the person would not prefer f to g, either knowing that the event B obtains, or knowing that the event $\neg B$ obtains, then he does not prefer f to g. (Savage [29])

Stated another way:

> If a person would choose the same action for every possible outcome of an experiment, then he ought to choose that action without running the experiment.

While arguments against the universal validity of the sure-thing principle have been contrived (see e.g., Blyth [4]), these usually involve intricate situations where actions influence the outcome of the experiment. If we violate the sandwich principle, we are bound to compromise the sure-thing principle even in those cases where its validity is clear and compelling. For example, a prisoner who prefers to wait patiently for his verdict before asking a question would suddenly decide to attempt an escape regardless of whether the jailer answers "B" or "not B."

In general, if actions are guided by the value Bel(A) accorded to some hypothesis A, any inference mechanism that violates the inequality in (18) would enable an agent to increase his perceived payoffs without actually observing the outcome of an experiment; merely verifying that the experiment was performed or just imagining the possible outcomes should suffice. Such an agent will spend precious resources chasing after useless information sources (like the jailer in the Three Prisoners story), or merely daydreaming about the existence of such sources. He will quickly become a victim of money pumping in any betting situation where useless information sources are offered for a price.

In view of this phenomenon, it is hard to see how belief functions could serve as a direct basis for rational decision making. It seems that to preserve rationality we must devise an updating rule that does satisfy Eq. (18). Alternatively, since the maladies associated with belief updating tend to get masked when it applies to additive probability distributions, we can go the Bayesian route and insist that before committing to any decision (or even to any belief), we first integrate the evidence into an additive probability distribution that reflects our prior state of belief. The next subsection discusses an attempt that follows the former alternative, while Section 4 will pursue the latter.

3.3. The Fagin–Halpern Conditioning

Fagin and Halpern (FH) [12] recently proposed a conditioning rule that avoids some of the pitfalls of Dempster's rule (similar rules have also been proposed by De Campos et al. [6]). The FH rule reads

$$\mathrm{Bel}(A \| B) = \frac{\mathrm{Bel}(A \wedge B)}{\mathrm{Bel}(A \wedge B) + \mathrm{Pl}(\neg A \wedge B)} \tag{19}$$

Like Dempster's rule, FH conditioning reduces to ordinary Bayesian conditioning when Bel(·) is an additive probability function, but, unlike Dempster's rule, it faithfully reflects the probability bounds that are produced by such conditioning. In other words, instead of the inequalities in Eq. (14), FH conditioning satisfies the equalities

$$\max_{P \in \mathcal{P}} [P(A|B)] = \mathrm{Pl}(A \| B) \geq \mathrm{Bel}(A \| B) = \min_{P \in \mathcal{P}} [P(A|B)]. \tag{20}$$

Unfortunately, although FH conditioning alleviates some of the problems encountered with belief functions, it creates new ones. First, FH conditioning is not commutative, and hence we cannot process multiple evidence sequentially upon arrival. In other words, in order to correctly produce the bounds created by the conjunction of two pieces of evidence, $B_1 = \mathit{true}$ and $B_2 = \mathit{true}$, we must condition Bel(·) directly on $B_1 \wedge B_2$ and compute $\mathrm{Bel}(A \| B_1 \wedge B_2)$; we cannot compute $\mathrm{Bel}(A \| B_1)$ and then condition the result on B_2.

Second, the FH rule does not specify how to integrate a body of evidence that is uncertain. If we have a belief state Bel(·) and we obtain evidence e having $m(B_1) = m_1$, $m(B_2) = 1 - m_1$, a natural way to update Bel(·) would be to take the convex mixture

$$\mathrm{Bel}(A \| e) = m_1 \,\mathrm{Bel}(A \| B_1) + (1 - m_1) \,\mathrm{Bel}(A \| B_2) \tag{21}$$

which represents the lower bound

$$\min_{P \in \mathcal{P}} [m_1 P(A|B_1) + (1 - m_1) P(A|B_2)]$$

However, it is not clear that this update captures the intended impact of the evidence. For example, a conflicting evidence with $B_1 = A$, $B_2 = \neg A$, and $m_1 = 1/2$ would always yield $\mathrm{Bel}(A \| e) = 1/2$ regardless of the prior state of belief Bel(A).

4. EVIDENCE POOLING

We have seen that belief functions are often incapable of representing partial knowledge, and in those cases where they are, they must first get converted into an additive probability function before they can absorb evidence in a plausible way, so as to support rational decisions. Thus, from the three components of BF methodology—representing partial knowledge, belief updating, and evidence pooling—we are left with evidence pooling as the only subtask in which belief functions can play a useful role.[10] This raises the question of whether it is worth invoking the entire BF machinery when conventional likelihood functions can offer an adequate solution.

Consider a body of evidence e characterized by a basic probability assignment $m(\cdot)$, and suppose we use this evidence to update a Bayesian belief function characterized by $m_0(w)$, where each w is a singleton set (i.e., a possible world). The updated belief function is governed by

$$(m_0 \oplus m)(w) = K m_0(w) \sum_{A:\, w \in A} m(A)$$

where K is a normalizing constant. Now assume that we were to update $m_0(w)$ by Bayesian methods, that is, we regard $m_0(w)$ as the prior distribution $\pi(w)$, and we represent the evidence e by the likelihood function

$$\lambda(w) \triangleq P(e|w) = \sum_{A:\, w \in A} m(A) \tag{22}$$

The posterior probability, in this case, will be identical to $(m_0 \oplus m)(w)$, since $P(w|e) = K\pi(w)\lambda(w)$. We can map any belief function into a likelihood function using (22) and be assured that, as long as the final belief is Bayesian, this Bayesian updating would yield identical results. Moreover, given that $m_1(\cdot)$ maps into $\lambda_1(w)$ and that $m_2(\cdot)$ maps into $\lambda_2(w)$, we can combine $\lambda_1(w)$ and $\lambda_2(w)$ directly into $\lambda_3(w)$ such that $m_1 \oplus m_2$ properly maps into $\lambda_3(w)$. This is done by ordinary multiplication, $\lambda_3(w) = \lambda_1(w)\lambda_2(w)$, an even simpler operation than the orthogonal sum required for m. Why, then, do we need the powerful machinery of belief functions, if we can accomplish the same task with likelihood functions?

True, belief functions convey more information than likelihood functions; two belief functions with totally different structures can be mapped into the same likelihood function, and their structures lost. But where is this structural information ever used? Clearly, since this information gets lost the moment it is absorbed into an additive probability function, the only way it can be used is prior to belief updating and decision making, namely, in the process of evidence pooling itself. But, in what way? A natural area of application would be that of evidence assessment and knowledge elicitation; an expert may feel

[10] Fagin and Halpern [12] reached the same conclusion but, intrigued by BF's ability to mimic likelihood functions, they see wider applications for belief functions than I do.

more comfortable describing the impact of an evidence in terms of weight assignment to classes rather than to individual points (see Gordon and Shortliffe [18]). However, here too, the likelihood function approach does a fairly decent job in an amazingly simple way (see Pearl [26]). Other possibilities are to use the structural information of belief functions for tracing back sources of contradiction and for identifying areas of incompleteness. Unfortunately, we have seen that the interval Pl – Bel does not reflect the degree of conflict in the evidence or the degree of ignorance in our knowledge. Thus, the challenge remains that of identifying how the rich mathematical structure of belief functions can be harnessed to facilitate the task of automated reasoning.

My personal assessment is that the most natural role for belief functions is to serve as a mechanism for processing evidential information prior to conversion into likelihood functions. This is based on the following observation: When a person says, "I am 50% sure that *B* . . .," he does not mean to sound neutral about *B* but expects to convey a positive support in favor of *B*. Even when a person says, "I think *B* is true but I am only 20% sure," he still expects the listener to increase his degree of belief in *B*. As strange as it sounds, people use absolute probabilities to communicate instructions about probability updates! In the Bayesian language, we need to convert such update statements to likelihood functions. For example, the statement "the evidence *e* supports *B* to a degree *p*" would be translated [see Eq. (22)] into the likelihood ratio

$$L = \frac{P(e|B)}{P(e|\neg B)} = \frac{1}{1-p}$$

This translation is indirect and lacks an immediate, natural justification. In the BF language, on the other hand, the statement above is encoded directly as a basic probability assignment $m(B) = p$. This suggests that there might be some phase of reasoning where people use belief functions, possibly in interpreting the impact of an observation, or during the mental exercise of estimating probabilities from a qualitative knowledge. It is possible that during such an exercise, which requires that exceptions and assumptions be summarized in terms of numbers, a person does indeed resort to random sampling of logical theories (Section 1.4) and computes (by stochastic simulation) the probability that a proof is found for *B*.

If this is so, then the BF theory could serve as a useful interface between a human probability assessor and a mechanical reasoner—a translator from colloquial assertions about probability assignments into likelihood functions. In this view, the application of belief functions in reasoning systems would parallel the role that likelihood functions have served in statistical analysis, namely, summarizing the impact of past observations (and arguments) and organizing them for absorption into a knowledge base that is compatible with one's background information, thus serving as a basis for decision making.

5. DISCUSSION

The limitations demonstrated in this paper raise some interesting issues concerning the relationships between the Bayesian and BF models. First, if the BF formulation is a generalization of the Bayesian one, how can the former suffer from limitations that escape the latter? In other words, why don't we regard the conclusions reached by the Bayesian analysis as a special case of what can be obtained by the BF analysis?

The answer is that the source of the difference between the two formalisms lies in their mode of application. While every additive probability function is indeed a special kind of a belief function, the Bayesian analysis concerns not one but a family of such functions. Interpreting a rule as a restriction on a family of ordinary probability functions gives different results than interpreting the rule as a single belief function, which is the prevailing interpretation in the BF strategy. If we give up this latter convenience and, mirroring the Bayesian strategy, we venture to treat each rule as a restriction on a *family* of belief functions, a new calculus of belief functions might emerge that might resolve some of the difficulties noted above. However, this new calculus would no longer possess the attractive features of Dempster's rule and, additionally, would still not guarantee to yield the intended results. The reason is that the space of belief functions complying with a given constraint, say $\text{Bel}(A|B) = p$, is much richer than the corresponding space of simple probability functions and might include some belief functions that exhibit undesirable features. This is a price we normally pay for generalization and needs to be further explored in the case of belief functions.

The second question relates to belief functions as a representation of evidence. In Section 4 we saw that the BF language offers a more refined representation of evidence than the ones used in the Bayesian analysis (e.g., the likelihood-ratio representation), and the question arises as to why it encounters such difficulties expressing simple items of information as if-then rules, which, presumably, also reflect and arise out of empirical evidence.

The answer lies in the distinction between knowledge and evidence and in the realization that if-then rules encode the former, not the latter. Belief functions offer a rich language for conversing about an item of evidence that has just been obtained but a rather inadequate language for describing cumulative evidence that has already been crystallized and elevated to a level of *knowledge*, that is, suitable for interpreting other items of evidence. In the words of Shafer and Srivastava [34]: "when we use the belief function formalism we always work with specific items of evidence" and "Statistical evidence is not, however, the kind of evidence for which the belief-function formalism is most simple and most natural." Translated into AI terminology, belief functions are not natural for representing "domain knowledge"—generic knowledge that an expert extracts either from statistical observations or from other experts in the field, as opposed to specific items of evidence. Conditional rules such as "Birds fly," "Fire causes smoke," "Smoke suggests fire," are the basic building blocks of expert knowledge, and it is in the representation of such rules that the BF formalism finds difficulties.

This distinction between knowledge and evidence must also be made in the Bayesian framework, and its importance in reasoning systems cannot be overemphasized, even in dealing with hard rules such as "All penguins are birds" (see Geffner and Pearl [13a]). Treating such a sentence as a generic rule that applies to all objects would yield totally different results than treating it as an item of evidence pertaining specifically to Tweety, saying "If Tweety is a penguin then she must be a bird" (equivalently, "There is indisputable evidence that Tweety is not both a penguin and non-bird").

The Bayesian formalism offers us the flexibility of distinguishing between these two interpretations; the former is treated as a restriction on a space of admissible probability functions, and the latter as an observation upon which the admissible functions should be conditioned. Given that Tweety is a penguin and a bird, the former treatment yields the expected conclusion that Tweety does not fly (see Pearl [27]), while the latter yields the vacuous range [0, 1], that is,

$$0 \leq P(\mathit{Flies}(\mathit{Tweety}) \,|\, \mathit{Penguin}(\mathit{Tweety}), \mathit{Bird}(\mathit{Tweety})) \leq 1$$

because the added information $\mathit{Penguin}(\mathit{Tweety}) \supset \mathit{Bird}(\mathit{Tweety})$ is subsumed in the available observation $\mathit{Penguin}(\mathit{Tweety}) \wedge \mathit{Bird}(\mathit{Tweety})$, and so it is unable to exploit the information that penguins are birds.

The BF approach, unfortunately, does not distinguish items of evidence regarding specific individuals from generic rules that reflect beliefs about all individuals, and it is for this reason that $\mathrm{Bel}(\mathit{Flies}(\mathit{Tweety}))$ remains oblivious to whether penguins are a subclass of birds or birds are a subclass of penguins. It should be noted that the problems associated with the representation of rules plague the BF approach only when rules are subject to exceptions. In domains where all rules are categorical, such as in strict taxonomic hierarchies, the distinction between knowledge and evidence is unnecessary and the paradoxes uncovered in Section 3 will not surface.

6. CONCLUSIONS

The current popularity of the belief function formalism stems from two factors:

1. Its willingness to admit partial knowledge, knowledge insufficient for specifying a complete probabilistic model.

2. The uniform and incremental manner in which it encodes and combines partial specifications, being reminiscent of logical deduction. Each specification (say a rule) is encoded as a belief function and then combined by a uniform principle, namely, Dempster's rule of combination.

This paper raises concerns relative to the ability of belief functions to properly handle partial knowledge. Our concerns have been twofold:

1. The BF analysis computes probabilities of logical necessity rather than probabilities of truth. The former may behave contrary to expectations (relative to the notion of "measure of belief") and may not provide the desired answers in diagnostic or abductive tasks.
2. The BF formalism encounters difficulties representing domain knowledge, especially class-property relationships, conditional assertions, and default rules. Encoding and combining such rules as belief functions may yield counterintuitive conclusions.

Shafer (personal communication) has defined the scope of BF applications as follows: "Belief functions are useful when we can formulate evidence bearing on one question in terms of probabilities, and then relate these probabilities, through a compatibility relation, to another question." In AI terminology, this scope limits the application of belief functions to cases where domain knowledge is articulated in purely deterministic terms, using a compatibility relation. Uncertainty may characterize the impact of specific items of evidence but may not enter into the stable relationships that govern entities in the domain. Such applications include strict taxonomic hierarchies, terminological definitions, and descriptions of deterministic systems (e.g., electronic circuits) but exclude domains in which the rules tolerate exceptions (e.g., medical diagnosis, default reasoning). Conditional rules, like those discussed in Section 3, do not lend themselves to direct BF analysis, and the problem of managing incomplete probabilistic knowledge remains unresolved.

But even limiting ourselves to applications where domain knowledge is expressible in categorical terms, serious thought must still be given as to whether the probabilities of provability computed by belief functions would be adequate for our purpose. Circuit diagnosis is one typical example where categorical relationships adequately describe the relations between the input and the output. However, shouldn't we really be concerned with the likelihood that a component is faulty as opposed to the likelihood that it can be proven faulty? If circuits are diagnosed by BF methods and components are replaced on the basis of the magnitude of Bel and Pl, we are back at the mercy of the spoiled sandwich effect, which in diagnosis problems might lead to unreasonable test-and-replacement strategies. Extensive experimental and theoretical studies should be undertaken to assess the behavior of such strategies, and to understand where and how they can be implemented safely.

This paper is written in the hope that by identifying areas where BF theory should not be applied, we will improve our ability to search and find admissible applications for the theory. As DeGroot once remarked relative to the Bayesian approach, "Like all powerful weapons, it must be used only with the utmost care and in accordance with the highest ethical standards" (see discussion following Shafer [33]). In order to use our weapons with care, we must understand their powers, dangers, and limitations.

ACKNOWLEDGMENTS

I have benefited from discussions with many colleagues, including J. Baldwin, N. Dalkey, A. Dempster, D. Dubois, H. Geffner, J. Halpern, H. Kyburg, I. Levi, M. McLeish, H. Nguyen, G. Shafer, B. Shay, P. Smets, P. Suppes, P. Walley, and N. Wilson. This work was supported in part by National Science Foundation grant IRI-88-21444 and Naval Research Laboratory grant N00014-89-J-2007.

References

1. Aleliunas, R. A., A new normative theory of probabilistic logic, *Proceedings of the 7th Biennial Conference of the Canadian Society for Computational Studies of Intelligence* (CSCSI-88), Edmonton, Alta., 1988, pp. 67–74.
2. Baldwin, J. F., Evidential support logic programming, *Fuzzy Sets Syst.* 24, 1–26, 1987.
3. Biswas, G., and Anand, T. S., Using the Dempster-Shafer scheme in a mixed-initiative expert system shell, in *Uncertainty in Artificial Intelligence*, Vol. 3 (L. N. Kanal et al., Eds.), North-Holland, Amsterdam, 1989, pp. 223-240, 1989.
4. Blyth, C. R., On Simpson's paradox and the sure-thing principle, *J. Am. Stat. Assoc.* 67, 364–366, 1972.
5. D'Ambrosio, B., Truth maintenance with numeric certainty estimates, *Proceedings of the 3rd IEEE Conference on AI Applications*, Orlando, Fla., 1987, pp. 244–249.
6. De Campos, L. M., Lamata, M. T., and Moral, S., The concept of conditional fuzzy measure, *Int. J. Intell. Syst.*, 1989.
7. Dempster, A. P., Upper and lower probabilities induced by a multivalued mapping, *Ann. Math. Stat.* 38, 325–339, 1967.
8. Diaconis, P., Review of "A mathematical theory of evidence," *J. Am. Stat. Assoc.* 78, 677–678, 1978.
9. Diaconis, P., and Zabell, S. L., Some alternatives to Bayes's rule, in *Information Pooling and Group Decision Making* (B. Grofman and O. Guillermo, Eds.), JAI Press, Greenwich, Conn., 1986, pp. 25–38.
10. Eddy, W. F., and Pei, G. P., Structures of rule-based belief functions, *IBM J. Res. Develop.* 30(1), 93–101, 1986.

11. Fagin, R., and Halpern, J. Y., Uncertainty, belief and probability, Research Report RP. 619/(60901), IBM, Almaden Research Center. Short version in *Proceedings, International Joint Conference in AI (IJCAI-89)*, Detroit, 1989, pp. 1161–1167.

12. Fagin, R., and Halpern, J. Y., Updating beliefs vs. combining beliefs, Unpublished report, 1989.

13. Falkenhainer, B., Towards a general purpose belief maintenance system, *Proceedings, 2nd Workshop on Uncertainty in AI*, Philadelphia, 1986, pp. 71–76.

13a. Geffner, H., and Pearl, J., A framework for reasoning with defaults, in *Knowledge Representation and Defeasible Reasoning* (H. Kyburg, R. Loui, and G. Carlson, Eds.), Kluer Academic Publishers, 1990, pp. 69–87.

14. Ginsberg, M. L., Non-monotonic reasoning using Dempster's rule, *Proceedings, 3rd National Conference on AI (AAAI-84)*, Austin, 1984, pp. 126–129.

15. Good, I. J., Discussion of "Lindley's paradox," by Glenn Shafer, *J. Am. Stat. Assoc.* 77(378), 325–351, 1982.

16. Goodman, I. R., A measure-free approach to conditioning, *Proceedings, 3rd Workshop on Uncertainty in AI*, Seattle, 1987, pp. 270–277.

17. Goodman, N., *Fact, Fiction and Forecast*, Athlon, London, 1954.

18. Gordon, J., and Shortliffe, E. H., A method of managing evidential reasoning in a hierarchical hypothesis space, *AI* 26, 323–357, 1985.

19. Hsia, Y. T., and Shenoy, P. P., An evidential language for expert systems, in *Methodologies for Intelligence Systems*, vol. 4 (Z. W. Ras, Ed.), Elsevier, New York, 1989, pp. 9–16.

20. Kyburg, H. E., Bayesian and non-Bayesian evidential updating, *AI* 31, 271–294, 1987.

21. Laskey, K. B., Cohen, M. S., and Martin, A. W., Representing and eliciting knowledge about uncertain evidence and its implications, *IEEE Trans. Syst., Man Cybern.* 19(3), 536–545, 1989.

22. Levi, I., Consonance, dissonance and evidentiary mechanisms, in *Evidentiary Value: Philosophical, Judicial, and Psychological Aspects of a Theory* (P. Gardenfors, B. Hansson, and N. Sahlin, Eds.), Library of Theoria No. 15, C. W. K. Gleerups, Lund, Sweden, 1983, pp. 27–43.

23. Lewis, D., Probabilities of conditionals and conditional probabilities, *Phil. Rev.* 85(3), 297–315, 1976.

24. Lowrance, J. D., Garvey, T. D., and Strat, T. M., A framework for evidential-reasoning systems, *Proceedings, 5th National Conf. on AI (AAAI-86)*, Philadelphia, 1986, pp. 896–901.

25. Nguyen, H. T., On random sets and belief functions, *J. Math. Anal. Appl.* 65, 539–542, 1978.

26. Pearl, J., On evidential reasoning in a hierarchy of hypotheses, *J. AI*, 28(1), 9–15, 1986.

27. Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Mateo, Cal., 1988.

28. Pearl, J., On probability intervals, *Int. J. Approximate Reasoning*, 2(3), 211-216, 1988.

29. Savage, L. J., *The Foundations of Statistics*, Wiley, New York, 1954.

30. Shafer, G., *A Mathematical Theory of Evidence*, Princeton Univ. Press, Princeton, N.J., 1976.

31. Shafer, G., Constructive probability, *Synthese* 48, 1-60, 1981.

32. Shafer, G., Belief functions and parametric models, *J. Roy. Stat. Soc. B* 44(3), 322-352, 1982.

33. Shafer, G., Lindley's paradox, *J. Am. Stat. Assoc.* 77(378), 325-351, 1982.

34. Shafer, G., and Srivastava, R., The Bayesian and belief-function formalisms: a general perspective for auditing, *Auditing, J. Pract. Theory*, vol. 9, supplement, 1990.

35. Smets, P., Un modéle mathématico-statistique simulant le processus du diagnostic médical, Doctoral dissertation, Free University of Brussels, Presses Universitaires de Bruxelles. (Cited and discussed in Shafer [32].)

36. Strat, T. M., Making decisions with belief functions, *Proceedings, 5th Workshop on Uncertainty in AI*, Windsor, 1989, pp. 351-360.

37. Zarley, D., Hsia, Y.-T., and Shafer, G., Evidential reasoning using DELIEF, *Proceedings, 7th National Conference on AI* (*AAAI-88*), St. Paul, 1988, pp. 205-209.