Rejoinder to Comments on "Reasoning With Belief Functions: An Analysis of Compatibility"

Judea Pearl
Computer Science Department,
University of California, Los Angeles, California

ABSTRACT

An earlier position paper has examined the applicability of belief-functions methodology in three reasoning tasks: (1) representation of incomplete knowledge, (2) belief-updating, and (3) evidence pooling. My conclusions were that the use of belief functions encounters basic difficulties along all three tasks, and that extensive experimental and theoretical studies should be undertaken before belief functions could be applied safely. This article responds to the discussion, in this issue, of my conclusions and the degree to which they affect the applicability of belief functions in automated reasoning tasks.

KEYWORDS: ***Belief functions, Dempster-Shafer theory, knowledge representation nonmonotonic reasoning, conditional information***

1. GENERAL COMMENTS

I was happy to find most discussants in agreement with the basic conclusion of my analysis: "To use belief functions (BF) and Dempster's rule blindly can be very dangerous indeed (Smets, Sm.13).[1] I was also gratified to find most responses, especially the insightful analysis of Nic Wilson, helpful in defining sharper boundaries between the safe and the dangerous applications of belief functions. Naturally, we have many points of disagreement. Some discussants

[1] References to the articles appearing in this issue will use the author(s) initial(s) (with Sm for Smets and Sh for Shafer), followed by a section number. References to my position paper [*Int. J. Approx. Reasoning* 4(5/6),363-389, 1990] will be given using the letter P, followed by a section number, e.g., P.4.

This work was supported in part by National Science Foundation Grant #RI-88-21444 and the State of California MICRO 90-127.

International Journal of Approximate Reasoning 1992; 6:425-443
© 1992 Elsevier Science Publishing Co., Inc.
655 Avenue of the Americas, New York, NY 10010 0888-613X/92/$05.00

have expounded philosophies unrelated to the issues raised in my paper; some have labored to show that other models of uncertainty (e.g., Bayesian analysis and upper and lower probabilities) suffer from similar or other weaknesses. I cannot address these topics here, as it would take us beyond the scope of this arena; neither will I attempt to answer every critical point mentioned by the discussants. Rather, I will focus on the central issue of this discussion, the role of belief functions in reasoning systems, and I will try to assess to what degree the discussants' replies affect the conclusions of my earlier analysis.

2. SEMANTICAL ISSUES AND PROBABILITIES OF PROVABILITY

In P.2, I described the metaphor of "probability of provability," which, to me, has been a constant source of insight into the properties and capabilities of belief functions (Pearl [1]). I slightly expanded this model in [2], where belief functions are described as conclusions drawn from randomly sampled assumptions. The probability-of-provability model turned out to be useful to many researchers. Laskey and Lehner [3] used it to augment ATMS with numerical measures of uncertainty. Greg Provan used ATMS to compute belief functions (Provan [4]) and has offered a comprehensive treatment of the relation between belief functions and propositional logic (Provan [5]). Nic Wilson [6] has developed Monte Carlo methods for calculating belief functions from such a model and has outlined ways of extending probability analysis to nonpropositional logics (see also Pearl [1], sec. 9.2.3). In summary, although some readers may find alternative metaphors more comfortable, I believe that probabilities of provability will continue to serve AI researchers as the canonical model for understanding belief functions.

Smets, on the other hand, denying all linkages to probability models (Sm.2), also denies what seems to be the only linkage belief function analysis maintains to experiential reality. The loss of this linkage from Smets's philosophy is regrettable, as it could have shed valuable light on many of his arguments, valid and invalid alike. It could have served to understand, as I will show in subsequent sections, why certain limitations of belief function theory are basic while others are curable.

Ruspini, Lowrance, and Strat endorse the usefulness of the probability-of-provability interpretation (RLS.2), and point out correctly the formal equivalence of this interpretation to Ruspini's epistemic probability. They maintain, however, that probabilities of provability, as opposed to probabilities of truth, are the "best" information that is available to the analyst, with which I cannot agree. As demonstrated in my Examples 1, 2, and 3 (P.2.1), analysts usually possess valuable knowledge that is simply not expressible as probabilities of provability. This includes partial probability assignment to subsets of variables, qualitative information of dependence and independence, and judgments about the relative magnitudes of probabilities. Belief functions are not sufficiently expressive for encoding and using this extra knowledge, and, hence, more powerful languages must often be invoked.

I must also take issue with the contention of Ruspini et al. (RLS.2) that we can only measure the probability that the evidence *implies A*, rather than the probability that A is *true*, given the evidence. When it comes to measurements, I believe it is probabilities of truth that we measure (approximately), while probabilities of implications are metareasoning constructs derived from two levels of abstraction, one consisting of a logical theory (or a ''compatibility relation'') and one that randomly imposes axioms on that theory (Pearl [1], pp. 423-427). Such constructs require the intervention of a human analyst to provide the logical theories; hence they can hardly be called ''measurable,'' not even approximately.

Provan questions the utility of attaching a declarative semantics to belief functions and proposes, instead, to examine their procedural semantics in the form of systems of conflicting arguments. I believe any concrete work in this direction will be very useful, especially now, when argument-based systems are gaining respectability in AI. At the same time, we should be cautious in appealing to procedural semantics to rescue the semantical difficulties found in the BF formulation. For example, I find little comfort in Provan's promise that ''if indeed the means of obtaining results is important, then Dempster-Shafer theory will be correct for some of the cases for which Pearl shows it is incorrect.'' Such cases, if any are found, should be carefully analyzed to ensure that the incorrectness can be safely contained.

Most discussants agree with my position that belief functions do not represent ''probabilistic ignorance by an interval of possible values.'' On the other hand, Ruspini et al. (RLS.2) and Dubois and Prade (DP.2) view belief functions as a computationally efficient tool for *approximating* manipulations on probability intervals. This may be a new promising role for belief functions, and it would be interesting to determine when the gain in simplicity would justify the loss in precision. The vanishing interval in the three prisoners example (see Section 7) illustrates that the loss in precision might occasionally become detrimental. Thus, I tend to agree with Shafer's conclusion that the family of probability distributions that are bounded by belief functions ''is a purely mathematical construct, with no conceptional significance'' (Sh.4.2).

Shafer's uneasiness with probability of provability, ''because it gives the impression that compatibility relations are always permanent'' (Sh.4.3), evokes two comments. First, I do not see why anyone will get an impression of permanence when the proofs under discussion are always predicated upon the evidence actually observed. In the case of Betty the witness (Sh.4.1), for example, we do not go through life trying to establish a proof for ''a tree fell on my car'' by sampling witnesses from Betty's reference class. Rather, we try to establish such a proof given Betty's specific testimony and using one of two randomly drawn assumptions: that Betty is reliable, or that she is not. The establishment of such a proof, admittedly, is not a random event whose frequency can be determined by direct measurement unless we have a large pool of witnesses sharing Betty's testimony and reliability rating. It is nevertheless a bona fide random event whose frequency can be determined analytically from that of another random event: Betty's reliability.

Second, the distinction between permanent and transitory compatibility relations has not, to the best of my knowledge, been required before in the literature on belief functions. It seems to me that in order to rescue Dempster's rule from the difficulties of the three prisoners puzzle (see Section 7) Shafer is introducing a new and dangerous twist into the theory, one that would take belief functions farther away from practical applications.

Shafer considers this distinction to be crucial. In his words, "In order to compare this with the conditions for Dempster's rule, . . . we have to interpret the multivalued mappings as transitory or permanent compatibility relations" (Sh.4.7), or, "It will certainly not be appropriate to combine a group of belief functions by Dempster's rule if more than one of them is based on a permanent compatibility relation" (Sh.4.5). Thus, the entire methodology of belief functions seems to hinge on this test for transitoriness of compatibility relations, a concept that I find hard to define and hard to test in practice.

If by a transitory compatibility relation we mean one that is sensitive to the content of the evidence reported, not merely to the properties of the evidence source, then most of my examples, including the three prisoners story, do induce transitory compatibility relations. If, however, transitoriness means sensitivity to samples drawn at random from the available probability distributions, then it is very easy to incorporate such sensitivity in my examples. For instance, we can imagine that the guard in the three prisoners story is only partly honest, which would make the compatibility between "the guard named *B*" and "*B* will be executed" a random relation, as desired by transitoriness. Yet this type of transitoriness will not rectify the paradoxical behavior of belief functions in the three prisoners problem. If, as a third possibility, a transitory relation must be one that lacks *any* permanent component whatsoever, then I submit that such relations are either nonexistent or extremely rare. Even the simple case of Betty (the witness with uncertain reliability, Sh.4.1) contains a permanent component: The outcome "unreliable" remains compatible with "tree fell on my car" regardless of what Betty says, and the evidence "Betty says a tree fell on my car" remains compatible with "a tree actually fell on my car" regardless of whether Betty is reliable or not.

How, then, are we to distinguish between permanent and transitory compatibility relations, a distinction that Shafer considers crucial?

In Sh.4.7 (last paragraph) we find a clue for yet another interpretation of "permanence." Here Shafer connects permanence with additivity—each of the three prisoners has an a priori 1/3 chance of being executed, as if "the victim is chosen at random." I fail to see how the method by which the victim is chosen has anything to do with compatibility, which is a relation between the identity of the victims and the identity of whoever is named by the guard. Would the compatibility relation turn transitory if the additivity of the belief function were not the result of a random choice but, say, of obtaining three incriminating items of evidence, one for each prisoner? If so, this would legitimize the applicability of Dempster's rule, and lead to the same paradoxical behavior as before.

I invite Glenn to provide us a precise test for permanence, one that forbids the application of Dempster's rule in the three prisoners puzzle yet does not exclude the use of belief functions in practical reasoning systems.

3. ON DECISION SUPPORT

All discussants have acknowledged problems in using BF analysis as a guide to decision making, and most agree that such a guide should exhibit more of the properties of additive probability functions. The question remains how to construct such functions from states of incomplete information.

Smets tells us about a pigmistic transformation that takes a belief function and maps it into an additive probability function (Sm.1). An alternative approach would be to convert belief functions into likelihood functions, and combine them with reasonably assessed prior probabilities (P.4). What is not clear in both approaches is whether it is worth carrying belief functions in their full glorious details, given that eventually they must be converted to additive functions. A convincing demonstration is needed to show that these details are crucial and cannot be approximated away by likelihood functions. A second question is whether decisions based on the pigmistic transformation would carry any guarantees of good behavior. We know, for example, that they are not guarded from violating the sure-thing principle (see Section 9), or from acting strangely in the three prisoners story (Section 7). It is important therefore to explicate what doctrines of rationality such decisions are expected to uphold.

Ruspini, Lowrance, and Strat (RLS.3) advocate the translation of belief functions into utility intervals. This invites problems similar to those of the pigmistic transformation: If BF analysis is capable of producing overly narrow probability intervals, it will also produce overly narrow utility intervals. Moreover, having failed to find a meaningful interpretation to the probability intervals produced by the BF analysis, I see no reason why utility intervals will fare any better and why the intervals computed by Ruspini et al. have anything to do with "intervals of possible utility values." Ruspini et al. maintain that the paradoxical intervals produced in problems such as the three prisoners story (Section 7) reflect a misuse of BF theory and can be cured by more sophisticated handling of dependent evidence. As I explain in Section 7, the problem has nothing to do with dependent evidence. Rather, it stems from the basic nature of belief functions as probabilities of provability, which Ruspini et al. accept. Therefore, my concern remains that the poorly understood "utility intervals" in the RLS analysis might lead to strange, if not harmful, decisions, unacceptable even to BF analysts.

4. LIKELIHOOD FUNCTIONS AS SUBSTITUTE FOR BELIEF FUNCTIONS

I was somewhat disappointed by the lack of response to my proposal to use belief functions as a means of translating subjective probability assessments to likelihood functions (P.4). Provan, for example, portraying me as "antipluralist" fanatic, has failed to notice my earnest and concrete attempt to follow his objective of seeking what he calls "the form of uncertainty best modeled by Dempster-Shafer theory." Dubois and Prade (DP.8) were busy tracing the historical origins of Eq. (22) (P.4) and are fascinated by its mathematical similarities with operations in possibility theory. Consequently, they too have failed to answer my basic question: What is lost by the transformation defined in Eq. (22), and, if not much is lost, is this a reasonable use for belief functions?

Smets (Sm.13) dismisses my proposal as "not so enlightening" (compared with the "more interesting" relations he once found to possibility theory) but, again, stops short of telling us why it is not *useful*. I believe that the applicability of belief functions deserves to be judged, not by their mathematical interest but by a down-to-earth analysis of compatibility.

5. EXPERIMENTS WITH INCOMPLETE KNOWLEDGE BASES

My second source of disappointment is connected with the response of BF practitioners. In (P.5) I raised the hopes that experimental studies will be undertaken to assess how vulnerable BF systems are to the weaknesses identified in my paper. I have hoped that researchers having extensive experience with such systems will shed some light on this question.

Indeed, Ruspini et al. (RLS.6) report difficulties with the "manipulation of conditional and dependent evidence," which they attribute to "evidential combination falling outside the scope of its representational capabilities." Strangely, they deny any relation between these difficulties and those identified in my paper, and insist on directing the blame at what they call the "more worrisome methodological limitation" of BF theory, one that they hope to overcome by devising new extensions to the theory. I do not wish to quibble with Ruspini et al. over the correct diagnosis of the difficulties observed in their experiments; I would simply invite them to make these observations public and let the readers judge whether the symptoms observed are basically different from those predicted in my analysis—the cures will emerge eventually.

6. REPRESENTING GENERIC KNOWLEDGE AND CONDITIONAL SENTENCES

Some of the respondents agree with my position that incomplete domain knowledge cannot be adequately represented by a single belief function (P.5). Dubois and Prade (DP.2) make an interesting suggestion that belief functions may still be used to *approximate* such knowledge. They analyze the feasibility of treating rules as constraints over a privileged set of belief functions and show that, contrary to my first concern regarding such treatment (P.5), if-then rules are indeed better behaved than those encoded as single belief functions. These results are very encouraging, but they bring to mind my second concern regarding the treatment of rules as constraints on belief functions: Will the abandonment of Dempster's rule result in overly cumbersome computations?

Smets dismisses difficulties of encoding domain knowledge as problems of "poorly known probabilities" (Sm.2), a relic from the province of upper and lower probabilities that belief functions are not mandated to handle. In his words: "I developed the TBM as a model totally unlinked to any underlying probability model just to avoid any such criticism." In other words, belief functions are exempt a priori from properly expressing any information to which one can find some probabilistic interpretation, and this, I am afraid, would include the bulk of human knowledge. Smets's position is convenient but also dangerous, because, as we have seen through several examples (P.2.2, P.3.1), maintaining some linkage to underlying probability models, even when such do not seem to "exist,"[2] often protects us from forgetting common sense altogether. It is important to recognize that the partial information reflected in these examples (which comprise qualitative information about independence relations and ignorance about dependencies) is an important component in human reasoning regardless of whether one postulates "the existence" or "the nonexistence" of probabilities.

In Sm.10, Smets attempts to resolve the difficulties associated with conditional sentences using the special predicates "typical birds" (*TB*) and "non-

[2] I am not aware that probabilities can literally "exist"; for me, probabilities are merely conceptual constructs that one chooses to reason with, not physical objects to be given existence tests. See Section 8 for elaboration.

typical birds'' (*NTB*), which are similar to McCarthy's abnormality (*ab*) predicate. Had he taken seriously the semantics of probabilities of provability, much of this effort could have been saved, because this semantics predicts that all the problems encountered in the logical approaches to default reasoning are bound to surface in Smets's or any other BF approach to default reasoning, as long as rules are treated as individual belief functions. This was shown more formally by Wilson [7].

To exemplify these difficulties, let us examine Smets's solution to the Tweety problem. In the default reasoning literature, the added sentence saying ''penguins are abnormal birds'' is called a cancellation axiom. This apparatus has the following weaknesses: Say we have another default, asserting ''Things that look like penguins are penguins,'' written $LP \rightarrow P$. We would naturally wish to conclude that if Tim looks like a penguin then Tim is a penguin, and so he does not fly. Unfortunately, the cancellation axiom cure now leads to undesirable effects; not only can we not conclude that Tim does not fly, we cannot even conclude that Tim is a penguin.[3] The reason is that the a priori high mass given to typical birds (*TB*) tends to deplete (through the contrapositive form of $P \rightarrow NTB$) the mass given to P by LP and $LP \rightarrow P$.[4]

Wilson (W.5) takes the position that ''belief functions do represent a very natural form of ''incomplete knowledge.'' Following Shafer [9], he identifies the circumstances when this form will become natural as ''when we have a probability function on a space Ω and wish to extend this via a compatibility relation to another space Θ.'' Unfortunately, the three prisoners story (see Section 7) teaches us that merely wishing to extend a probability function over a compatibility relation does not make the extension natural.

Wilson then goes to great length to illustrate that default rules interpreted as standard belief functions have some desirable properties, properties not shared by conditional probability interpretations of defaults unless fortified by independence assumptions.[5] This is to be expected because, by encoding rules as randomized material implications, we invite all the power of standard monotonic logic, including chaining, contraposition, and irrelevant information (W.5.3), and we compromise on specificity and reasoning by cases. Indeed, the conditional and implicational approaches represent two extreme points of the spectrum. However, although substantial progress has been achieved in

[3] The ills of cancellation axioms were pointed out to me by Hector Geffner and were conveyed to Smets and Hsia in my review of their paper (Smets and Hsia [8]). I was therefore surprised to read Smets's statement, ''The paradox then disappears.''

[4] This is the same contraposition problem that Smets dismisses so lightly at the end of Section 10.

[5] Wilson (W.5.3) admits that it is easier to incorporate such independence assumptions (or conventions) in the lower probability approach than with the BF approach, where they are inflexibly constrained by Dempster's rule.

going from the conditional toward the implicational (see, e.q., Geffner [10], Goldszmidt et al. [11], Goldszmidt and Pearl [12], very little progress is reported in the opposite direction, that is, equipping implication-based systems with features that properly handle specificity and reasoning by cases.

Glenn Shafer objects to the very distinction between "knowledge" and "evidence," apparently because such distinction, if authenticated, would legitimize the asymmetry that Bayesians perceive between old evidence (which determines prior and conditional probabilities) and new evidence (on which we condition). Glenn goes so far as to endow me with a credit I do not deserve: the creation of the philosophical vocabulary that distinguishes "knowledge" from "evidence." The fact is, this distinction has been recognized by many philosophers before. The need to separate necessary from incidental relations, or background from contingent knowledge, had been the primary motivation behind the development of modal logics and possible-worlds semantics. Necessary relations, like the tendency of birds to fly, are attributed to *all* possible worlds, while the flying ability exhibited by a given observed group of birds (as well as Glenn being in Maine and wearing a white shirt) is attributed to our *particular* world.

In fact, the knowledge-vs.-evidence distinction is manifested so clearly in our language that one hardly needs philosophers to document its legitimacy. The word *if*, for example, always marks a relation that belongs to one's background knowledge; it never connects specific observations. It appears that "if" serves to warn the listener that a sentence such as "if Fido is a dog then Fido barks" conveys general knowledge about dogs—it should not be mistaken for an observational fact (to be phrased "Fido is not a nonbarking dog") and should not, therefore, be used to condition one's state of knowledge upon.

The most glaring manifestation of the separation between knowledge and evidence can be found in the use of counterfactuals. For example, believing that "all blocks on this table are green" does not entitle us to conclude that "had this blue block been on this table it would have been green." At the same time, believing that "all boxes painted in this shop are green" does entitle us to conclude "had this blue box been painted in this shop it would have been green." To me, the fact that we uniformly agree on which counterfactuals are legitimate means that we uniformly agree on what constitutes a "tendency of things to happen" as opposed to "things that actually happened." And if we agree on this distinction, I should not feel too guilty for defining "knowledge" and "evidence" the way I did in my paper, even if I were the first to do so (which is counterfactual).

I do not claim, of course, that the distinction between old evidence and new evidence is "handed to us by Nature" (Sh.2.3). What I do claim is that the human race has found it useful to summarize and organize some of the old evidence in a special way and to elevate these summaries to a special status called "knowledge," a status unshared by new observations. I also claim that, in order to be compatible with human-style reasoning, theories of evidence must take into account this peculiar organization of the mind and, accordingly, must accommodate the distinction between knowledge and evidence.

7. THE THREE PRISONERS PROBLEM

In the three prisoners problem (P.3.1),[6] the discussants are divided into two groups, the first approving of the BF result of $\mathrm{Bel}(A) = 1/2$, the second attempting to modify it within the BF framework. Smets accepts the increase of belief from $\mathrm{Bel}(A_1) = 1/3$ to $\mathrm{Bel}(A_1) = 1/2$ as perfectly reasonable, apparently because in his analysis $\mathrm{Bel}(x)$ can be interpreted as the probability that x occurs ''whatever Mother Nature (however hostile) will do'' (see Sm.7.1), and this most hostile strategy occurs when the guard is determined to avoid naming prisoner 3 whenever possible.

Dubois and Prade, too, find nothing strange in the answer $P(A_1) = 1/2$. They consult a ''purely logical'' analysis (DP.6) and find that the information provided by the guard warrants an increase of belief from 1/3 to 1/2. After all, so the argument goes, prisoner 1 started with three equally likely possibilities and ended up with only two, so ''Why would prisoner 1 conclude that prisoner 3 has twice as many chances as himself to be executed given that the guard said prisoner 2 was to be saved?''

What both Smets and Dubois and Prade fail to notice is that the remaining possibilities are no longer equally likely. Prisoner 3, unlike 1, *could* have been named by the jailer, and the fact that he was not named renders him suspect of being ''unnamable,'' that is, the one who will be executed. To convince the reader that such an argument should play a central role in plausible reasoning, consider the 1000 jailers version of the three prisoners story, as I used in [13] and [1, revised second printing, Exercise 9.5, p. 466]. Imagine prisoner 1 repeating the experiment with 1000 different jailers, none knowing about the others, and all naming 2 as one who will be released. Intuitively, such a strange coincidence should call for some explanation as to why 3 was not named by any of the jailers, strongly suggesting that he was not in fact available for naming, being the victim himself. This, then, is the reason that prisoner 1 should conclude that prisoner 3 has a much greater chance than himself to be executed.

This commonsense consideration cannot be brushed off as a by-product of

[6] For the sake of uniformity, I will use the same notation as in P.3.1; that is, one of three prisoners 1, 2, 3 is to be executed; A_i stands for the event that prisoner i is the unfortunate one; and the jailer (or guard) named prisoner 2 as one who is to be released. The strategy by which the jailer selects names (in case there is a choice) is *not* known [see P.3.1, Eq.(15)].

one's "postulating the existence" of an additive probabilistic function (Sm.2); they are an integral part of human intuition, and if postulating the existence of probability models is our only way of capturing this intuition, then it is certainly an exercise worth pursuing. This intuition is not reflected in the "purely logical" approach advocated by Smets and Dubois and Prade, in which there will always remain two possibilities (either 1 or 3 will be executed) regardless of how many jailers are involved. Those committed to the use of belief functions in reasoning systems must then find other ways of embracing this intuition within the formalism.

Superficially, the BF analysis may seem to adapt the explanation that prisoner 3 was not named because each and every jailer (without colluding with the other) was determined to avoid naming 3 whenever possible, as if BF analysis is always concerned with Nature's most hostile strategy, as suggested by Smets (Sm.7.1). However, if Nature's most hostile strategy yields $\mathrm{Bel}(x) = 1/2$, then, presumably, Nature's "most benevolent" strategy should be used in calculating $\mathrm{Bel}(\neg x) = 1 - \mathrm{Pl}(x)$. But, alas, the "most benevolent" strategy consists of the jailers conspiring to name prisoner 2 whenever possible, and this strategy yields $P(A_1) = 0$, not $P(A_1) = 1/2$ as in the BF analysis. In fact, the BF analysis does not concern itself with strategies at all; the result $\mathrm{Bel}(\mathrm{A}_1) = 1/2$ is a direct product of invoking the unusual notion of probability of provability (instead of probability of truth) in problems of diagnosis (see P.1 and Pearl [2]).

Ruspini, Lowrance, and Stratt present a totally new defense for the three prisoners story (RLS.5.1). They simply claim that Dempster's rule of updating is inapplicable in this case because the jailer's answer is not independent of the process by which guilt and innocence was decided. This suggests a refreshing new requirement of evidence independence that unfortunately, would limit the application of Dempster's rule to trivial cases where it is hardly needed.

For an item of evidence to be of any value, it must depend in some way on the state of nature for which it is an evidence (e.g., the identity of the victim in the three prisoners story). Moreover, if we have several items of evidence, each depending on the state of nature, these items of evidence should also depend on each other. This kind of dependency is not a nuisance but a necessary bliss; no evidential reasoning would otherwise be possible. This is precisely the kind of dependency that Dempster's rule was designed to handle (see Shafer's random code example). Thus, I do not see the point in the effort of Ruspini et al. "to extend the original theory to produce and utilize conditional belief information that incorporates known dependencies between evidential bodies." If this statement means that one should refrain from using Dempster's rule (and BF analysis?) until their extension is completed, then Ruspini et al. have a much more pessimistic view of the applicability of belief functions than that expressed in my paper.

Glenn Shafer, too, takes the position that Dempster's rule cannot be applied to the three prisoners problem, albeit for a different reason: the compatibility relation being nontransitory. I have questioned the viability of the ''transitoriness'' criterion in Section 2; here I will only comment on Shafer's treatment of ''protocols.''

Contrary to Shafer, I doubt that the three prisoners story has ever been a serious puzzle for Bayesians. It might have served as a curious riddle for testing theory against unaided intuition, but not as a puzzle that threatens to undermine the foundations of the theory. Deeply entrenched in the Bayesian analysis is the principle that we should condition only on the evidence actually observed, not on conclusions derived from observations. For this reason, Bayesians never deny the need to consider what Shafer calls a ''protocol,'' namely, the method by which the evidence was obtained. On the contrary, Bayesians consider protocols to be useful tools for interpreting and assessing the impact of evidence, not a constraining nuisance as suggested by Shafer's analysis.

Indeed, any theory of evidence that is insensitive to protocols is destined to clash with common intuition. The 1000 guards variant of the three prisoners story (third paragraph, this section) illustrates this point quite clearly. Here we have a situation where the bare evidence is identical to that of the original story (namely, prisoner 2 is sure to be released) but the protocol is different: 1000 guards (instead of one) now point to prisoner 2, and none points to prisoner 3. Surely, human intuition dictates that evidence obtained under these circumstances should have a significantly different impact on the fate of prisoner 1 than that obtained in the original story. We expect such a difference in impact to prevail even in cases where we have no idea how the guard(s) would choose between naming 2 and naming 3. Yet any theory that attempts to ignore protocols, belief functions included, will remain oblivious to these intuitive demands. I thus wonder: How many guards are needed before we acknowledge that protocol analysis is inevitable and that merely assuming any reasonable protocol (or several reasonable ones) is better than pretending that a protocol does not exist?

8. HOW DOES JUDEA PEARL INTERPRET PROBABILITY? (SH.2.7)

Shafer disapproves of the way I have been avoiding the historical debates concerning the frequency-vs.-belief interpretation of probability (Sh.2.7). My position on the relevance of these debates to AI is expounded elsewhere (Shafer and Pearl [14], pp. 341–343). Briefly, I always interpret probabilities to mean degrees of belief, and I attribute their conformity with the calculus of frequency (in most judgmental tasks, not only in gambling situations) to their being summaries of mental experience, and mental experience is rich with frequencies. In this section, I would like to stress the similarity of my interpretation of probability to Shafer's general idea of constructivism.

In my mind, ''a willingness to talk about the probability of something without asking whether there is such a probability'' (Sh.2.7), a strategy I called ''model completion'' in Shafer and Pearl [14], lies at the very heart of constructivism. To me, constructivism means that we do not stop to question whether an object still exists each time we turn our head; rather, we construct a convenient coherent picture of the world assuming that the object remains intact throughout the head-turning experience. In the physical sciences, constructivism permits us to imagine that fields exist even in a vacuum, where there are no particles to attest to their presence. It means that when we fail to estimate the mass of an object, we do not suddenly question whether it has mass in the first place. Drawing the analogy closer to home, constructivism means that the practice and benefits of consulting probability models when we have enough evidence to construct them do not stop abruptly just because we happened to miss a measurement, an argument, or a reference class; they simply require constructive ways of compensating for the missing information.

Shafer claims that his constructive theory of belief functions is ''a way of making some probability judgments without creating so many fictional probabilities that speculation about them swamps what little evidence we have.'' I believe the few examples in my paper show that this overly cautious strategy often produces precisely the opposite results: A failure to construct just one or two judgments can cause us to abandon or misrepresent valuable information that is well grounded in experience, and to totally shift our interest away from our original intent. In the three prisoners problem, for example, the BF strategy prevents us from expressing the valuable information that our evidence suggests that the guard has no reason to prefer naming one prisoner over another, or that we strongly believe that the 1000 guards acted independently of each other. Additionally, instead of asking for the probability that Art will be executed, we are now forced to settle for the probability that his execution is inevitable, a question we might not care to ask.

By contrast, my interpretation views probability as a mental construct that we impose on reality, whose legitimacy lies in the computational and psychological advantages it produces. In the same way that it is legitimate to attribute particle-like properties to entities like electrons, holes, quasars, and waves, so it is also legitimate to attribute probability-like properties to states of uncertainty, part of which, admittedly, are convenient extrapolations at best.

I believe that the Peter, Paul, and Mary example (P.3.2) demonstrates quite clearly what can be gained by pretending that probabilities exist even where their existence could be questioned by the zealous philosopher. It demonstrates that certain useful information about probabilities, about independence, and about causality (see Section 9, paragraph 5) can best be represented as part of a hypothetical probability model, and that this information will be lost if we are prevented from constructing such a model.

To summarize, the advantages of the "model completion" strategy include

1. Proper utilization of those probabilities that do have evidential basis, especially judgments about independence and about causal relationships
2. A single calculus, facilitating conditionalization, for handling both partial and complete probability models
3. Coherent behavior and protection from paradoxical reasoning patterns like those demonstrated in my examples
4. Indications of where ignorance lies and where information is missing
5. Sensitivity to protocols and multiple evidence

9. THE SPOILED SANDWICH EFFECT (P.3.2)

Smets (Sm.7) launches an all-out effort to show that the sandwich principle [perhaps because it hurts the promotion of his Eq. (10)] should not be taken as a guiding principle for commonsense reasoning or for decision making. He gives several examples in which our intuition should presumably favor the spoiling of the sandwich, none of which is really convincing.

EXAMPLE 1 (WHETHER X IS YOUNG) The weak point in this example is premise (5), $\mathrm{Bel}_{12}(Y) < \mathrm{Bel}_{1M}(Y)$, which I find to be rather questionable. Imagine, for the sake of argument, that our doubt in the full reliability of the witnesses W_1 and W_2 originates from the possibility that the person they both saw was not really X but X', who is known to be an old woman. Now, from testimony E_2 we can logically infer that the person observed was X, not X'. So our belief $\mathrm{Bel}_{12}(Y)$ is 1. In contrast, learning that X is a male (M) still leaves room for suspecting that the person seen was X', not X, so Bel_{12} $(Y) > \mathrm{Bel}_{1M}(Y)$. True, the story does not tell us that X' is an old woman. But not being told about a possibility does not mean we should not account for it before rushing into agreement with Smets's premise (5).

EXAMPLE 2 (WHETHER THE KILLER WILL USE A GUN) Suppose you have three potential killers, A, B, and C. I shall select one of them, but you will not know how. Each killer selects his weapon by an independent random process with $P(gun) = 0.2$ and $P(knife) = 0.8$. What is your "belief" that the killer will use a gun? The BF solution is $\mathrm{Bel}(gun) = 0.2 \times 0.2 \times 0.2 = 0.008$, which Smets attempts to defend. The solution offered by the sandwich principle is 0.2, which Smets insists on attacking.

The reason I repeat here the full description of this example is that it demonstrates so vividly what belief functions compute, why these computations yield counterintuitive results, and how the sandwich principle can guard us against such results. Smets himself notices that not all is right with his analysis. He therefore rationalizes the conclusion Bel(*gun*) = 0.008 by appealing to a bizarre selection process, whereby the killer is chosen by a knife-loving person *after* finding out what weapon each candidate is about to use. The more sensible result, Bel(*gun*) = 0.2, is dismissed, as usual, as one that applies only in those cases where the killer is chosen by a random device.

Well, here is a challenge to TBM analysts: Suppose we know with absolute certainty that the killer is to be chosen *before* the weapon is selected and, of course (to block the usual escape route), that the choice of the killer is *not* made by a random device but by some undisclosed strategy. Should we still accept the result Bel(*gun*) = 0.008 as measuring our belief that a gun will be used? In other words, suppose I happen to disbelieve in premonition and other supernatural powers, hence I am totally convinced that it is impossible to foretell which weapon will be selected for each killer (by the random process). Would the language of belief functions permit me to express this conviction as part of my knowledge base, or must TBM users be resigned to having their beliefs governed by fears of the unnatural? From Smets's analysis, it appears that the latter is the current state of affairs in TBM.

A similar challenge can be addressed to Smets's treatment of the fat-intelligent-popular example (Sm.10) that I used to illustrate why the contrapositive form of if-then rules should not be invoked indiscriminately (P.2.3). Smets promises us: "The reason these conclusions are unsatisfactory is that there are extra constraints we would like to see fulfilled... but did not include in our initial analysis. Introduce them in the TBM analysis, and the result will be satisfactory." Well, here are the extra constraints: Our culture normally considers obesity and intelligence to be independent properties of individuals. Can we encode this cultural bias within the TBM language? And when we do, can the TBM analysis guarantee that Joe would not be branded a moron if he were found to be fat?

Among all respondents, only Wilson has seriously addressed this challenge (W.5.2). He observes correctly that the assumption of "exception independence" (similar to that used in noisy OR models) is not appropriate for handling conflicting rules, and concludes, therefore, that Dempster's rule, which implicitly invokes this assumption, should not be used. Wilson errs in one point, though (W.5.2). The Bayesian interpretation of if-then rules never appeals to likelihood ratios. Likelihood ratios are reserved for the encoding of specific items of evidence, whereas if-then rules are interpreted as conveying domain knowledge. Additionally, although it is fairly easy in the Bayesian analysis to impose the constraint that two variables (e.g., obesity and intelligence) remain independent, it is not straightforward to encode this valuable piece of information in the BF language.

EXAMPLE 3 (KING DAVID AND BATH-SHEBA) Smets seems to be so immersed in King David's dilemma that he fails to notice my anticipative warning that such examples bear no relevance to the acceptability of the sure-thing principle, which reads:

> 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.

In (P.3.2) I stated:

While arguments against the universal validity of the sure-thing principle have been contrived (see, for example, [Blyth 1972]), 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*."

Smets's version of the King David and Bath-Sheba story (a variant of Simpson's paradox) falls precisely into those situations where actions influence the outcome of the experiment. Thus, nothing in that story should convince us to abandon the sure-thing principle (hence the sandwich principle) in situations where the outcome of the experiment is independent of the decision one is about to make.[7] The statistical table in the King David story shows B and C dependent, as though the kings surveyed did not make their B decisions freely but were inhibited by their charismatic standing. Alternatively, this dependence might mirror how the kings' charisma (C) was tarnished by their iniquitous acts (B). In King David's case, since he is trying to make a free-choice decision about Bath-Sheba (B), the table that should be consulted depends on whether he believes that C is a cause for B or the other way around. In the former case, he should consult the table that fits his charismatic class; in the latter case, he should rely on the population as a whole (Table 3).

But this story has nothing to do with the sure-thing principle. The principle simply claims that if David were contemplating taking a charisma test, and if he is determined to act B assuming the test comes up C and act B assuming that the test comes up $\neg C$, then he can skip the test altogether and act B immediately. The principle does not force David to act one way or the other given his charismatic status, nor does it tell David which statistical table should be consulted.

Indeed, no one insists on accepting the sure-thing principle in cases where

[7] The following paragraphs contain excerpts from my February 5, 1990, letter to Smets, in which I discussed his King David example.

the decision influences the observation. For instance, if I knew for sure whether or not you would accept my position, in either case I would go to sleep and not bother to build a stronger argument. However, not knowing your reaction, I choose to stay awake and try to convince you. This is not a violation of the sure-thing principle. On the other hand, a prisoner waiting patiently for his verdict before getting the jailer's answer who suddenly decides to escape from prison regardless of whether the answer is B or not-B, does violate the sure-thing principle.

Both Wilson (W.4) and Ruspini et al. (RLS.5.2) find comfort in the fact that upper and lower probabilities also violate the sandwich principle. This is not surprising; lower probabilities are used primarily to indicate where knowledge is missing but were not advocated as guides to decision making or as measures of degree of belief (see P.3.1, footnote 9). It is worth pointing out, though, that upper and lower probabilities do obey a weaker form of the sandwich principle:

$$\min\left[P_*(A \mid B), P_*(A \mid \neg B)\right] \leq P_*(A) \leq P^*(A) \leq \max\left[P^*(A \mid B), P^*(A \mid \neg B)\right]$$

No such bounds hold for belief functions, as illustrated by the three prisoners example. The ramification of this bound is that it is always possible to satisfy the sandwich principle by selecting values for $P(A)$, $P(A \mid B)$, and $P(A \mid \neg B)$ from the ranges defined by the upper and lower probabilities. Moreover, when these ranges become sufficiently narrow, as is the case in ϵ-semantics (Pearl [1]), the sandwich principle is safely restored.

Wilson (W.4), like Smets, maintains that the sandwich principle cannot be regarded as a general principle for plausible reasoning. I can follow Wilson's examples relative to measures of support (W.4.1), likelihoods (W.4.2), random probabilities (W.4.3), and lower probabilities (W.4.4), and I agree that these measures do not obey the sandwich principle. However, when it comes to measures of belief (W.4.5), I must admit to a terrible shortcoming: I cannot see how the "Philippe, Pearl, and Mary" example violates the sandwich principle nor why it has anything to do with measures of belief. I believe that the reader will also find this example hard to follow and less than convincing.

10. CONCLUDING REMARKS

This brings to mind a rather discouraging thought. If the sandwich principle is indeed as readily broken as Wilson and Smets would like us to believe, why does it take such laborious effort to contrive even one simple and natural example that invites a violation of this principle? After all, if belief functions were representative of commonly used concepts in human reasoning, we should have been able to collect such examples by just randomly sampling sentences from ordinary conversations. The paucity and contrivance of such examples suggest that the language of plausible reasoning has not endowed Bel and Pl (or lower probabilities) with the status of *basic* concepts, as it did with "degree of belief."

Let us examine this point more closely via the Peter, Paul and Mary story of Example 5 (P.3.2). According to Smets, Peter's credal state of belief in winning the 1000 dollars is zero, while his pigmistic belief is 1/2. Yet if we were to simply ask a person how strongly he/she believes in winning, the answer would be 1/2, not zero. We would need to labor really hard to explain to that person what that epistemic construct should be that attains a value zero. (It is for that reason that I had to use the phrase "a 50% assurance," which Ruspini et al. correctly detected as being ill defined, at least formally; the English language simply does not have a term corresponding to Smets's "credal state of belief.") I therefore doubt that **BV** theory formalizes an "epistemic construct " (Sm.1) as natural as "degree of belief." The closest mental construct to a belief function would be "strength of argument" or "evidential support," but even these I doubt will receive a measure of zero for "winning" and a measure of 1/2 for "heads" in the Peter, Paul, and Mary example. What evidence do we then have that people commonly use an epistemic construct possessing the properties of belief functions? The only evidence I found (P.4) is the use of the expression "I am only 20% sure that X" when we wish to convey evidence in favor of X. But, apparently, my suggestion for using belief functions to handle such expressions did not gain much support among the discussants.

Regardless of how my analysis will impact the actual use of belief functions, I believe it has served a useful purpose in focusing, clarifying, and organizing the issues involved from an automated reasoning standpoint. At the very least, I hope it helped define a common vocabulary and canonical examples for communicating new ideas. I thank all discussants for tolerating the bluntness of my inquiry and for taking the time to seriously address the issues of my concern.

References

1. Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference,* Morgan Kaufmann, Palo Alto, Calif., 1988.

2. Pearl, J., Which is more believable, the probably provable or the provably probable? *Proceedings, CSCSI-90,* Eighth Canadian Conference on Artificial Intelligence, Ottawa, May 23-25, 1990, pp. 1-7.

3. Laskey, K. B., and Lehner, P. E., Assumptions, beliefs and probabilities, *AI* **41**(1),65-77, 1989.

4. Provan, G., An analysis of ATMS-based techniques for computing Dempster-Shafer belief functions, *Proceedings International Joint Conference on AI,* 1115-1120, 1989.

5. Provan, G., A logic-based analysis of Dempster-Shafer theory, *Int. J. Approximate Reasoning,* 4,451-498, 1990.

6. Wilson, N., Justification computational efficiency and generalization of the Dempster-Shafer theory, Res. Rep. 15, Dept of Computing and Mathematical Science, Oxford Polytechnic, 1989.

7. Wilson, N., Rules, belief functions and default logic, in *Proc. 6th Conference on Uncertainty in AI,* P. Bonissone, and M. Henrion, Eds., Cambridge, Mass., August 1990.

8. Smets, P., and Hsia, Y. T., Default reasoning and the transferable belief model, *6th International Conference on Uncertainty in AI,* Cambridge, Mass., 1990.

9. Shafer, G., Probability judgment in artificial intelligence and expert systems (with discussion), *Stat. Sci.* **2**, (1), 3-44, 1987.

10. Geffner, H., Default reasoning: causal and conditional theories, UCLA Cognitive Systems Laboratory, Tech. Rep. R-137, November 1989; Ph.D. Dissertation.

11. Goldszmidt, M., Morris, P., and Pearl, J., A maximum entropy approach to nonmonotonic reasoning, *Proceedings, AAAI-90,* Boston, July-August 1990, pp. 646-652.

12. Goldszmidt, M., and Pearl, J., System-Z$^+$: a formalism for reasoning with variable-strength defaults, *Proceedings AAAI-91,* Anaheim, Calif., July 1990.

13. Pearl, J., Bayesian and belief-functions formalisms for evidential reasoning: a conceptual analysis, in *Intelligent Systems: State of the Art and Future Directions* (Z. W. Ras and M. Zemankova, Eds.), Ellis Horwood, Chicago, 1990, pp. 73-117.

14. Shafer, G., and Pearl, J., Eds., *Reading in Uncertain Reasoning,* Morgan Kaufmann, Palo Alto, Calif., 1990.