Response to the Discussion of Belief Functions[1]

Glenn Shafer

1. Introduction

I would like to thank everyone involved in organizing and carrying out this exchange. Prakash Shenoy, Gautam Biswas, and James Bezdek worked hard to make it possible. Judea Pearl has provided a thorough review of problems in interpreting and using belief functions, and the discussants have pondered thoughtfully both Pearl's article and my own review article.

I have had several previous opportunities to respond to published discussions of belief functions (Refs. 18, 20, and 23), but those discussions were dominated by critics and skeptics. It is a pleasure, therefore, to respond to an exchange in which most of the participants are keenly interested in developing the theory and application of belief functions.

The present discussion is primarily about how to interpret belief functions, and the discussants have a wide variety of views on this topic. I am somewhat disappointed at this, both because I would like to see work on belief functions move on from interpretation to application, and also, of course, because I have been explaining my own views on the interpretation of belief functions for many years, and I had hoped to be more persuasive. Yet the diversity of views is understandable, for belief functions use probability, and there is no consensus on how to interpret probability.

I fear we will not agree on the interpretation of belief functions until we agree on the interpretation of probability. We will not even make much progress in understanding each other's interpretations of belief functions until we have more common ground in our understanding of probability. So in this response to the discussion, I will explain the constructive interpretation of probability that I favor, and I will relate this interpretation to the views and concerns of Pearl and the discussants.

After discussing the interpretation of probability, I turn to the interpretation of probability bounds. These too, I argue, require a constructive interpretation. I then turn to the interpretation of belief functions. I review what I consider the main canonical examples for belief functions: the partially reliable witness and its generalization, the randomly coded message. And I will relate my constructive approach to belief functions, based on these canonical examples, to the different interpretations advanced by Pearl, by Philippe Smets, and by Enrique Ruspini.

I will conclude by responding to Nic Wilson's comments on the Monte Carlo implementation of Dempster's rule, and by mentioning some references that I overlooked in my review article.

---

[1]To appear in *International Journal of Approximate Reasoning*, 1992.

2. The Interpretation of Probability

Contrary to the impression that Larry Wasserman took from my review article, I do not want to separate frequency and belief in the interpretation of probability. I believe that probability describes, in the first instance, a special and unusual situation where frequency and belief are unified. This special situation is the classical game of chance, involving a sequence of experiments (successive throws of a die or draws from a pack of cards, etc.) whose outcomes have known long-run frequencies. These frequencies are our only basis for prediction, and hence they define fair betting rates and rational degrees of belief, both for events involving single experiments and events involving many experiments. We observe the outcomes of the experiments as they are performed, and the outcome of each experiment changes the fair betting rates and rational degrees of belief for events that involve that experiment.

Different ways of applying probability should be thought of as different ways of using the special situation, not simply different ways of using probability distributions. The most common frequentist methods use the special situation as a model or a standard of comparison. Bayesian and belief-function methods, in contrast, draw an analogy between the situation of a person with given evidence and certain imperfect forms of the special situation. All these ways of using the special situation are constructive. The special situation seldom occurs in nature. We do not find probabilities ready-made. We construct them.

In this section, I argue for this constructive interpretation of probability by reviewing the history of the debate over the roles of frequency and belief in probability; I explain how the ordering of events in the special situation serves as a protocol for new information; and I contrast frequentist and Bayesian uses of the special situation. I use the three-prisoner problem to show how Bayesian arguments deal with the absence of a protocol for new information. I conclude the section by comparing my understanding of the Bayesian approach with the views of other participants in the discussion, especially Didier Dubois and Henri Prade, Larry Wasserman, and Judea Pearl. I hope that this will help us find more common ground in future discussions.

Readers interested in a fuller account of the constructive interpretation of probability outlined here can consult Refs. 17, 25, 26, and 28. I have discussed conditional probability more fully in Ref. 22, historical issues in Refs. 19 and 27.

2.1. How to Reunify Frequency and Belief

Probabilists have debated the competing claims of frequency and belief for over a hundred years. Long-run frequency and rational belief had been seen as complementary aspects of probability during the eighteenth and early nineteenth centuries, but in the mid-nineteenth century, rational belief was challenged by empiricists who regarded it as hopelessly metaphysical. Long-run frequency, they felt, was the only properly empirical foundation for probability theory. Nineteenth-century frequentism still dominates most people's thinking about probability, but it has been challenged in the twentieth century by a resurgent subjectivism, led by Bruno de Finetti [4] and L.J. Savage [15]. The twentieth-century subjectivists, or Bayesians, are as empiricist as the frequentists. They share the frequentists' disdain for the alleged rational degrees of belief and fair betting rates of classical probability. But rather than substitute long-run frequency for these classical ideas, they substitute personal degrees

of belief and personal betting rates, which they consider behaviorally defined and hence empirically respectable.

Most of us are comfortable with the duality of frequency and belief in the fair games of chance that were the original domain of the theory of probability. But probabilists have always wanted to extend the scope of their theory; it was this ambition, together with empiricism, that tore frequency and belief apart. Nineteenth century empiricism insisted that every term of a scientific theory be defined in terms of observables. So in order to apply probability theory to domains outside games of chance, probabilists had to find observable quantities in those domains that could be called probabilities. In most domains, there is nothing that is simultaneously a frequency and a belief. But there are frequencies, and there are beliefs. It seemed necessary to make a choice.

I do not believe that this choice is still necessary. Empiricism is still the dominant philosophy of science, but today's empiricism is more flexible in its understanding of the relation between theory and practice. We can go back to the original integrated picture of probability, which still describes only the special situation of fair games of chance, and we can think of the different uses of probability as different uses of this integrated picture. It can be used as a model, as a standard of comparison, as a tool for deliberate randomization in experiments and surveys, or as a canonical example for arguments by analogy.

Figures 1 and 2 illustrate the change I am suggesting in the way we think about probability. Figure 1 illustrates the status quo, while Figure 2 illustrates the approach I advocate.

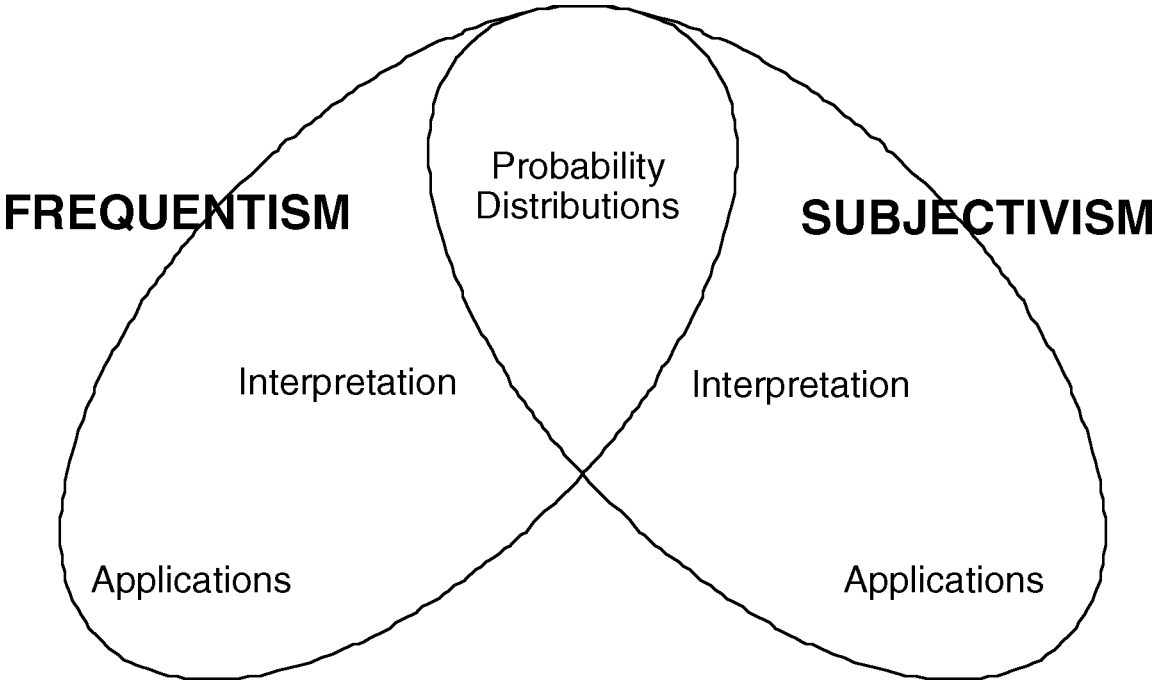

Figure 1. The agreement to disagree.

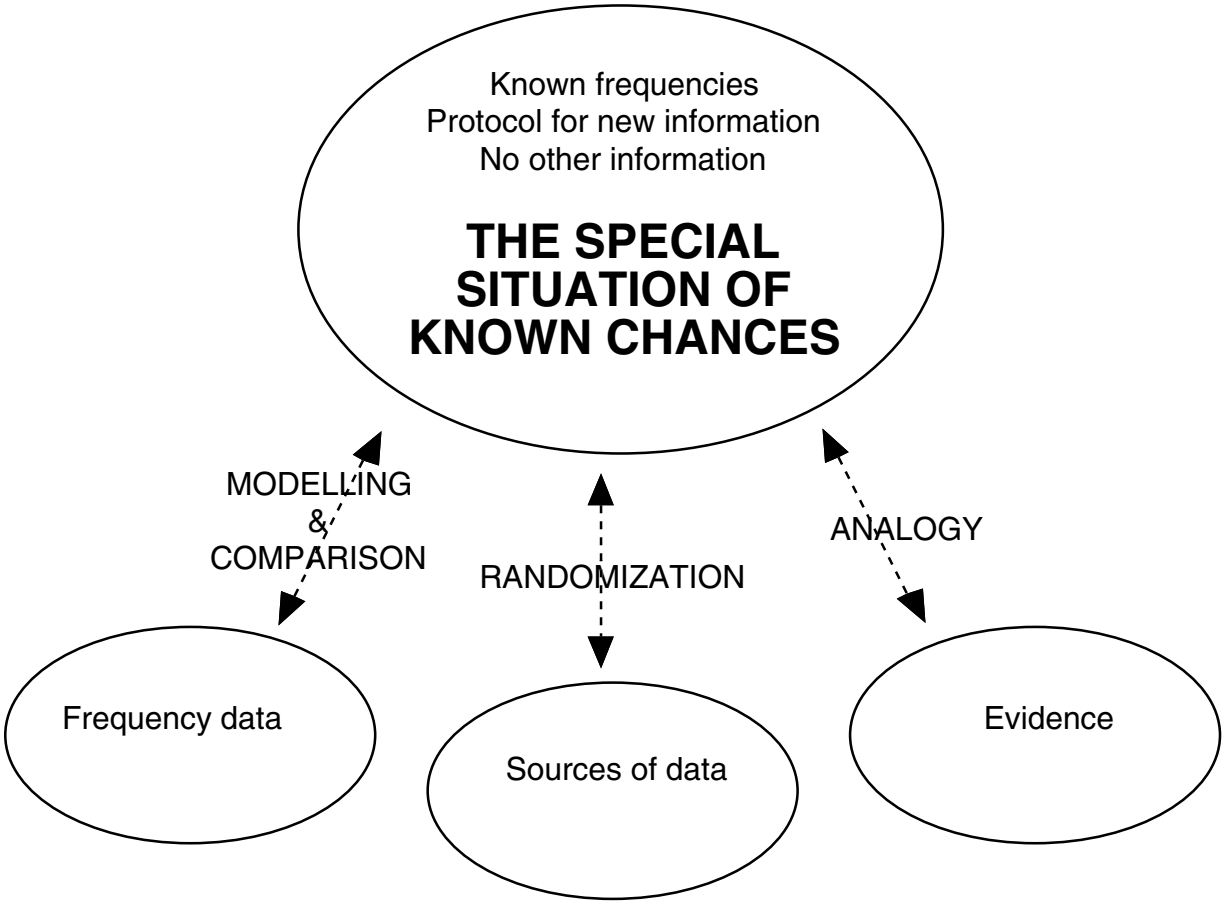

Figure 2. The constructive interpretation of probability.

The status quo is an agreement to disagree. The frequentists and subjectivists agree that their only common ground is the mathematics of probability distributions. Because they base their applications of probability on different interpretations of these distributions, their applications are incomparable. There is no common language in which we can assess whether an application using one interpretation is better or worse than an application using the other. Every attempt at such assessment becomes an argument about the interpretations. As Figure 1 emphasizes, the applications are separated from each other by the interpretations.

In Figure 2, in contrast, the different ways of using probability do not represent irreconcilable philosophies. They merely represent different ways of using the special situation that probability theory describes. I place both Bayesian and belief-function arguments in the lower right-hand corner of this figure. They both draw analogies to the special situation.

2.2. A Closer Look at the Special Situation

In previous articles (especially Ref. 24), I have described the special situation in great detail. Without repeating that description here, I will outline my conclusions.

First, the special situation necessarily involves a sequence of events. Such a sequence is needed in order to talk about long-run frequency. It is also needed to talk

about fair betting rates and rational degrees of belief. The betting rates are fair because they break even in the long run, and the degrees of belief are rational because they are fair betting rates.

The events to which we assign probabilities include events defined by the outcomes of single experiments, but they also include events that involve many experiments—the event that exactly half of the first thousand die throws will come up even, the event that Peter will get double sixes before Paul, and so on. All these probabilities are fair betting rates; if you bet at these rates, you can expect to break even in the long run, you cannot make money for sure, and you have no reasonable prospect of multiplying your initial stakes substantially. Not all these probabilities are long-run frequencies. Only the probabilities for outcomes of individual experiments (if the same experiment is repeated over and over) or long-run average probabilities for outcomes of successive experiments have a frequency interpretation. This point is very important when we use the special situation as a model for time series or spatial data, where there is no prospect of repetition of the entire process of observation, and hence no frequency interpretation for global probability statements (Matheron [11]).

The ordering of events in the special situation justifies changes in belief. We learn of events as they happen—we move through the sequence of events with nature—and our probabilities change accordingly. The rules that determine the sequence in which the experiments are performed (the next experiment to be performed may depend on the outcomes of the preceding experiments) therefore determine the possibilities for what we learn. What we learn is always an event; but only events that fully specify the outcomes of all the experiments up to some point and say nothing about the outcomes of the experiments after that point can play this role. We learn such an event whenever an experiment is performed, and we change our betting rates and our beliefs to reflect the outcomes specified and the long-run frequencies for the experiments that remain. This is justified because when we change our betting rates in this way, we can still expect to break even in the long run, we still cannot make money for sure, and so on.

When the sequence of experiments is thought of as a set of rules for what we will learn, I call it a protocol for new information.

2.3. Frequentist and Bayesian Probability Arguments

Both frequentist and Bayesian probability arguments are constructive, in the sense that they do not find probabilities ready-made in nature. They construct probabilities, and the result of the construction is not independent of the process of construction. Like belief-function arguments, they have a process-dependent semantics.

The most straightforward way to use probability is to create an actual situation that approximates the special situation. This means creating a situation where we know frequencies, and where the things we learn are outcomes of planned observations or experiments. Statistical work is often thought of in this way, but expert systems with large databases provide an even better example. If we use a large expert system repeatedly under reasonably constant conditions, we can build up a database that gives us frequencies for the population defined by these conditions. If we have a protocol for the use of the system—a set of rules for the conditions under which we make various tests or ask various questions—then frequencies conditioned on the progress of our knowledge of the individual case are rational degrees of belief for that individual case.

This way of using the special situation is usually considered a frequentist use of probability. I call it "modelling" in the lower left corner of Figure 2.

There are other frequentist ways of using the special situation. Most statistical testing uses the special situation more as a standard of comparison than as a model. For example, we often assess the value of a prediction equation derived from statistical data by comparing its performance to what we could expect from a similar prediction equation in an instance of the special situation where there is no real correlation between a predictor and what is predicted (Freedman and Lane [5]). Randomized experiments and sample surveys involve another frequentist use of the special situation (this is the middle circle at the bottom of Figure 2). Here we create an instance of the special situation by generating random numbers, and then we use the random numbers to plan our experiment or survey.

Most statistical teaching is concerned with parametric estimation, where observations or experiments are planned, but the frequencies of outcomes depend on unknown parameters. Since we do not fully know the frequencies in this situation, the special situation serves as an imperfect model. The Neyman-Pearson theory of confidence intervals shows that the frequencies we do know, together with the observations, can still provide guidance about the values of the parameters. In many cases, the frequencies given the parameter values are largely hypothetical, so that they might better be called subjective judgments. We are really drawing an analogy between our evidence and a situation where we know such frequencies. Bayesians strengthen this to a fuller analogy with the special situation by positing frequencies for the parameter values as well.

The task of artificial intelligence and the ambition of Bayesian theory both go beyond statistical problems. In statistical problems, we have a plan for observations or experiments that defines a protocol for new information, and frequencies for these experiments are partially known. But in the problems of everyday life, we constantly encounter unexpected information—information for which we have no protocol and no frequency experience. How can Bayesian methods deal with such problems?

Outside the realm of planned experiments and surveys, Bayesian methods are not really implemented by conditioning on new information as it comes along. The distinction between old evidence (which determines prior probabilities) and new evidence (on which we condition) is not handed to us by nature. The Bayesian makes it up. The Bayesian deliberately partitions her evidence into a part on which she can conveniently base prior probabilities and a part she can conveniently use to condition these prior probabilities, or perhaps a part on which to base likelihoods (Shafer and Tversky [29]). This can be confirmed by examining almost any serious Bayesian study, where the evidence on which to base prior probabilities is searched out after the "new" evidence is obtained.

The absence of a protocol for new evidence in Bayesian analyses of everyday evidence means that the Bayesian analogy between that evidence and the special situation is imperfect. But analogies are never perfect. A Bayesian's imperfect analogy can be persuasive if the Bayesian can convince us that her choice of the information on which to condition is not systematically biased, and if we are also willing to assume that

the mechanism or people who gave her the information did not have any systematic bias. In order to explain this point, let me turn to the three prisoners puzzle.

2.4. The Bayesian Three Prisoners Problem

Though Pearl and several of the discussants treat the three prisoners problem as a puzzle about belief functions, it originated, as Enrique Ruspini, John Lowrance, and Thomas Strat note, as a puzzle about probability (Mosteller [13]). It is one of many puzzles, some going back many decades, that have been used to illustrate the need for protocols in strict probability reasoning (Shafer [22]). In this section, I discuss it as a probability puzzle. In Section 4.7, I will discuss it as a belief-function puzzle.

Art, who is a Bayesian, knows that one of the inmates in his prison has been singled out for execution. In fact, he knows that the unfortunate inmate is either himself, Bob, or Carl. He knows that one of the three will be executed, and the other two will be released, but he does not know who will be executed and who will be released. He has some evidence, which he represents by assigning probability one-third to each possibility:

$$\text{P(Art will be executed)} = \frac{1}{3} \ ,$$

$$\text{P(Bob will be executed)} = \frac{1}{3} \ ,$$

$$\text{P(Carl will be executed)} = \frac{1}{3} \ .$$

The guard knows who will be executed, but he is forbidden to reveal to Art whether he will be executed. Within this constraint, he can tell Art the name of one prisoner who will be released, and Art convinces him to do so. He says Carl will be released. So Art revises his probabilities by conditioning, obtaining

$$\text{P(Art will be executed | Carl will be released)} = \frac{1}{2} \ ,$$

$$\text{P(Bob will be executed | Carl will be released)} = \frac{1}{2} \ ,$$

$$\text{P(Carl will be executed | Carl will be released)} = 0.$$

Had the guard said Bob will be released, Art would have revised his probabilities by conditioning on that information, obtaining

$$\text{P(Art will be executed | Bob will be released)} = \frac{1}{2} \ ,$$

$$\text{P(Bob will be executed | Bob will be released)} = 0,$$

$$\text{P(Carl will be executed | Bob will be released)} = \frac{1}{2} \ .$$

So merely by convincing the guard to tell him the name of a prisoner who will be released, Art seems to have guaranteed that his probability for his own survival will go down from two-thirds to one-half. This hardly seems appropriate.

There are many ways to explain what is wrong. One way is to say that Art has not conditioned on all his information. He should take into account, and condition on, everything he knows. He should condition not on the event

A = Carl will be released,

but on the event

B = the guard says Carl will be released after I convince him to tell me the name of a prisoner that will be released, within the constraint that he not reveal whether I will be executed.

In order to condition on B, Art must extend his probability model so that B is in it. His new probabilities will depend on how he does this.

Suppose Art extends his model as follows. First he assumes that it was a sure thing that he would convince the guard to tell him the name of a prisoner to be released, so that conditioning on B amounts to conditioning on B', where

B' = the guard says Carl will be released.

Next, he assigns probabilities to B' under various hypotheses. His assumptions require him to assign the probabilities

$$P(\text{ B' | Carl will be executed}) = 0 ,$$
$$P(\text{ B' | Bob will be executed}) = 1 ,$$

but these assumptions say nothing about how the guard will choose between naming Bob and naming Carl if Art is to be executed. Suppose Art gives probability one-half to both possibilities:

$$P(\text{ B' | Art will be executed}) = \frac{1}{2} . \qquad (2.4.1)$$

Then conditioning on B' produces the probabilities

$$P(\text{Art will be executed | B'}) = \frac{1}{3} ,$$
$$P(\text{Bob will be executed | B'}) = \frac{2}{3} .$$

The guard's testimony does not change Art's probabilities for his own survival after all.

Unfortunately, Art may not have any evidence on which to base the probability (2.4.1). He may not have any idea how the guard would choose between naming Bob and naming Carl. In this case, Art has no way of making a good Bayesian argument.

The primary message of this puzzle is that conditioning on new information is strictly legitimate within probability theory only if our probability model includes a protocol specifying conditions under which we will acquire different possible items of new information. If our model does not include a protocol for new information, then a policy of conditioning on any information that comes along leaves us open to paradox, and perhaps even to manipulation by those supplying the information.

Fortunately, the constructive understanding of Bayesian arguments softens the conclusion that a protocol is absolutely necessary. The imperfect analogy to the special situation made by a Bayesian argument can be persuasive even if the process by which information is acquired or chosen as conditioning information is not part of the probability model, provided that the choice of conditioning information is not systematically biased. The guard in the three prisoners problem violates this assumption; he systematically supplies a subset B that not only always contains the truth (the name of person to be executed) but also always contains Art. But it is often reasonable to hope that a Bayesian can avoid this kind of systematic bias when she chooses information on which to condition.

I would like to be able to articulate more precisely the requirement that there be no systematic bias in the choice of the information B on which we condition. But the only

obvious way to be more precise is to specify a probabilistic mechanism for the choice of B. (For example, we might determine B by drawing a partition Π from some probability distribution for partitions of Θ and letting B be the element of Π that contains the truth.) And this would simply recreate the protocol that we have admitted does not exist.

2.5. Updating and Focusing

Didier Dubois and Henri Prade, in their contribution to the discussion, distinguish between "updating" and "focusing." They use this terminology effectively to discuss limits on conditional probability and complications that arise when we consider families of distributions. I prefer to avoid the term "updating," however, because it supports the fiction that Bayesians condition on information as it comes along.

2.6. De Finetti's Foundations for Bayesian Methods

Larry Wasserman reviews de Finetti's foundations for the Bayesian approach to probability, and he asks, in a sympathetic way, whether analogous foundations are possible for belief functions. Is there a theory of "coherence" for belief functions? Can belief functions subordinate the idea of frequency to the idea of belief as de Finetti does using the concept of exchangeability?

Underlying these questions, I believe, is a desire to relate belief functions, and every other application of probability, as closely as possible to the special situation where frequency and belief are unified. I am sympathetic to this desire, and I will try to respond to it in Section 4. I must first, however, emphasize my opposition to de Finetti's approach to Bayesian foundations.

Wasserman states de Finetti's coherence argument concisely. Suppose, he says, that I interpret P(A) as the price at which I would buy and sell tickets that are worth one unit if A is true and are worthless otherwise. Such offers can lead to a sure loss for me unless the numbers P(A) obey the rules of probability. So this interpretation is a justification for using the rules of probability: "...to be coherent, one *must* use probability." The betting scenario, Wasserman says, is more than just a canonical example: "it serves as a way of testing the theory of probability to see if it makes sense."

After pondering this argument for many years, I am still unable to see in it more than a series of non sequiturs. Why I should set a price at which I would both buy and sell a ticket on A? Is it incoherent for me to demand a buy-sell spread? (As Wasserman mentions, allowing buy-sell spreads leads to lower instead of additive probabilities.) Is it incoherent for me to refuse to gamble at all? Or to gamble only in a situation where I know at least as much as those who might bet with me? Or to gamble only when I know the betting rates are fair?

The betting interpretation of probability did not originate with de Finetti; it began with the invention of probability theory in the seventeenth century. In the original betting interpretation, one bought and sold a ticket on an event at a single price because this price was fair. One aspect of the price's fairness, first articulated in the eighteenth century, was the guarantee it provided of breaking even in the long run. De Finetti, writing in the 1920s and 1930s, wanted to liberate the betting interpretation from any reliance on long-run frequency, so he substituted personal prices for fair prices. He

demanded, with no justification that I have ever been able to see, that for every event a person should have a single price at which he or she will both buy or sell a ticket.

De Finetti and his Bayesian followers have failed to look critically at their own rhetoric because they have been preoccupied with refuting the errors of frequentism. In order to be useful weapon against the frequentists, the "coherence" argument only needs to make as much sense as the excessive claims of frequentism. Now this battle has been won, and the time has come to be more critical.

My disagreement with Wasserman about de Finetti's contribution extends to de Finetti's theory of exchangeability, which subordinates frequency to belief. Wasserman sees this as a success story; I see it as a mistake. If we do not have frequencies in the first place to justify the additivity of probability, then we cannot follow de Finetti in getting frequencies out of additivity plus exchangeability. In my view, we should subordinate neither frequency nor belief to the other; they are equally essential aspects of the special situation. A mathematical description of the special situation can begin with rules or assumptions about either aspect, but justifying these rules will then take us back to the other aspect. To justify fairness for betting rates and additivity for beliefs, we need long-run frequency. To justify our knowledge of the long run, we need odds and beliefs.

These strictures on interpretation aside, I share Wasserman's curiosity about the possibility of extending de Finetti's mathematical results to belief functions. I have not followed up on the work I published along these lines in 1982, but I still consider the question intriguing.

2.7. How Does Judea Pearl Interpret Probability?

Although Judea Pearl does not discuss the interpretation of probability explicitly, many of his comments are incompatible with my constructive interpretation.

The most glaring point of difference between us is Pearl's ability to take the concept of probability for granted. He is willing to talk about the probability of something without first asking whether there is such a probability, or by what argument or analogy it is to be constructed. At the end of Section 1 of his article, he writes that we should be more interested in "the probability that A is true given the evidence" than in belief-function calculations of "the probability that the evidence implies A." He spends several paragraphs driving this point home, without stopping to consider the claim, made in almost every exposition of belief functions, that "the probability that A is true given the evidence" sometimes simply does not exist.

The reason for being interested in belief functions is not that we prefer them to direct probabilities, but that sometimes we cannot construct convincing direct probabilities. Nic Wilson makes this point very effectively.

A willingness to talk uncritically about *the* probability of A is common, especially among those uninterested in philosophical issues. But Pearl goes beyond the common culture to create his own philosophical vocabulary. He draws a novel distinction between "knowledge" and "evidence." Knowledge, he writes, "encodes judgments about the general tendency of things to happen," while evidence "summarizes the impact of that which actually happened" (page 364). This distinction is designed to

make the Bayesian formalism sound like common sense. Knowledge is supposed to give us prior probabilities, and evidence is supposed to give us likelihoods. But the distinction is contrary to common usage and to fact. Most of my knowledge has nothing to do with general tendencies or probabilities (my shirt is white; I am in Maine, etc.), and my evidence often does involve frequencies The distinction is also deeply misleading. It wrongly suggests that nature tells us what should go into the prior probability part of a Bayesian analysis and what should go into the likelihood part.

Pearl's emphasis on frequency knowledge suggests that he is primarily interested in frequentist rather than Bayesian applications of probability. I applaud this interest. As I said in my review article, a few solid bits of frequency information are often much more useful than extensive subjective judgment (Dawes et al. [1]). And the further development of frequentist expert systems is an important area of endeavor for both AI and statistics. But I believe that both Bayesian and belief-function arguments belong in a different arena. They belong in the arena of everyday and commonsense reasoning, in which we constantly grapple with unexpected evidence and questions for which we have not compiled frequency information.

3. The Interpretation of Probability Bounds

Much of the present discussion is concerned with whether belief functions express bounds on probabilities. I find this disappointing, because I gave belief functions their name in order to distinguish them from probability bounds.[2] But the debate goes on. Enrique Ruspini, John Lowrance, and Thomas Strat, unconvinced by my arguments or Judea Pearl's examples, still contend that belief-function degrees of belief are bounds.

This continuing disagreement may be due in part to the need to interpret probability bounds themselves. Are they bounds on frequencies or bounds on beliefs? How can we use them in a constructive theory of evidence?

I begin this section by commenting on the distinction between bounds on distributions and families of distributions, and on the possible role of Bayesian conditioning in a constructive theory based on bounds or families. Then I discuss several distinct ways of basing a constructive theory of evidence on bounds or families.

3.1. Bounds and Families

Let me underline a point made, in different ways, by Pearl and Wasserman. Any family of probability distributions on a frame Θ determines lower bounds for the probabilities of subsets of Θ. We call this system of lower bounds the family's lower probability function. But several families can have the same lower probability function. Only one of these families, the one that includes *all* probability distributions bounded

[2] The first published use of “belief function” in its present meaning was in my 1976 book, *A Mathematical Theory of Evidence* (Ref. 16). The term had been used earlier, by Richard Jeffrey [7] and perhaps others as well, to mean a Bayesian probability function. Since I have always emphasized that I adopted “belief function” as a new name for Dempster's systems of “lower probabilities” in order to forestall misinterpretation in terms of probability bounds, I am disconcerted by Pearl's assertion that “belief functions were first interpreted as lower and upper probabilities induced by a special family of probability distributions” (page 367).

below by the lower probability function, is specified simply by specifying the lower probability function. Thus bounds are inadequate to represent families of distributions. The idea of a family of distributions is more general than the idea of a lower probability function.

The fact that bounds are less general than families does not establish the inadequacy of a theory of evidence based on bounds alone. Greater generality is not always needed, and greater representational power always has its price in computational difficulty. Is the greater representational power of families of distributions needed? Different ways of using bounds to represent evidence—different constructive theories that use bounds—may lead to different answers to this question.

3.2. Conditioning Bounds

A related issue is the role to be played by Bayesian conditioning. Bayesian conditioning cannot be applied to lower bounds, but it can be applied to the distributions bounded by them. If we have represented our evidence by a lower probability function and we learn that the truth is in a certain subset, then we can change the lower probability distribution by conditioning each distribution it bounds.

As Pearl points out in Section 3.3 of his article, this procedure for conditioning lower probability functions is not commutative.[3] If we use families of distribution as our representation of evidence, then conditioning is commutative; conditioning all the distributions in a family P first on B and then on C is the same as conditioning them all first on C and then on B. But if we use lower probability functions as our representation, and we think of conditioning the distributions bounded by a lower probability function merely as a background device, then this commutativity is lost. This is because a second conditioning will involve all the distributions bounded by the lower probability function obtained by the first conditioning, not merely the distributions obtained by the first conditioning. If P_* is the lower probability function with which we begin, P is all the distributions bounded by P_*, P_B is the family obtained by conditioning the distributions in P on B, and $P_{*|B}$ is the lower probability function for P_B, then P_B may not include all the distributions bounded by $P_{*|B}$. If we regard $P_{*|B}$ as the new representation of our evidence, then in order to condition on another event C, we will repeat the whole procedure, which means conditioning all the distributions bounded by $P_{*|B}$, not merely the ones in P_B.

It is not clear that conditioning should be commutative. As we have seen, conditioning is strictly justifiable for probability only in accordance with a protocol, which specifies the order in which information can be learned. The order in a protocol cannot be reversed. If the protocol allows B to be all we have learned at one point, and C to be all we have learned at a later point, then C is a subset of B. So the formal commutativity of conditioning for probability (and belief functions) seems more a curious accident than a basic principle.

It is also not clear why we should condition a lower probability function by conditioning the distributions it bounds, or why we should condition a family of

[3] Pearl emphasizes the case of lower probability functions that happen to qualify mathematically as belief functions. In this case, this kind of conditioning can be implemented by the formula he cites.

distributions by conditioning its individual distributions. To justify these procedures, we need a semantics for bounds or families.

3.3. What do Bounds or Families Mean?

The unity of belief and frequency that characterizes the special situation described by probability theory is possible because the long-run frequencies in that situation are known. This unity cannot survive a generalization to bounds or families of distributions. If we know only bounds on probabilities, then the probabilities are not known frequencies. So we must choose, to some extent at least, between interpreting them as beliefs and interpreting them as frequencies.

There are several ways of making the choice, each of which leads to a different interpretation of bounds on probabilities and possibly to a different constructive theory of evidence. Here I will sketch four possibilities: we can adopt a strict belief interpretation of families, we can interpret lower bounds directly as betting rates, we can adopt a complete frequency interpretation of families, or we can put ourselves in the position of a partial observer of the special situation.

Isaac Levi's Theory of Credal Commitment. One possibility is to abandon frequency altogether and interpret the distributions in a family as distributions of belief. This has been advocated most clearly by Isaac Levi [10]. Like de Finetti, Levi interprets a probability distribution as a commitment to bet. But he goes beyond de Finetti by interpreting a family of probability distributions as a state of indecision among such commitments. He makes detailed proposals for using and modifying such families in the face of new evidence, changes of opinion, and opportunities to seek further information or take practical action.

In Levi's theory, as in de Finetti's theory, a person can choose whatever beliefs she wants. While this is undogmatic, it leaves us without guidance in using the theory to represent evidence. Levi's proposals for changing beliefs are more definite, but they are complex and arbitrary (Shafer [21]). This may be inevitable, because without the connection with frequency, the additivity of probability itself is arbitrary.

Betting Rates with Buy-Sell Spread. Another possibility is to interpret lower probability functions merely as systems of betting rates, under a scheme in which bettors are permitted a buy-sell spread. Suppose, indeed, that for each subset A of a frame Θ, Betty declares the highest price she will pay for a ticket on A. The prices she declares for different subsets should be related, because tickets on various events can sometimes to be compounded to form a ticket on another event. As it turns out, the prices should form a lower probability function for some family of distributions. We need not, however, attach any significance to the family. Thus this interpretation of lower probability functions does not compel us to consider to more general families of distributions.

The betting interpretation of lower probability functions has been studied thoroughly by Peter Walley [31]. It seems incomplete, however, as a canonical example for a theory of evidence. Why does the person in the canonical example have given betting rates? What kind of evidence lies behind these betting rates? We need to know more about that evidence in order to draw convincing analogies between it and our actual evidence. Moreover, the betting interpretation does not impose rules for combining

evidence. Walley recommends Bayesian conditioning for updating, but this seems arbitrary without an interpretation for the families of distributions bounded by the betting rates.

Families of Frequency Distributions with Bayesian Conditioning. Let me turn now to the most obvious and common interpretation of probability bounds: there is a true frequency distribution for some phenomenon, but we do not know it exactly. We know only a family of distributions that contains it. We use the lower probability function of this family as one expression of our beliefs. When the occasion arises, we change our beliefs by conditioning each distribution in the family.

This sounds deceptively straightforward. Having grown accustomed to thinking about frequentist probability in terms of a mathematical distribution rather than in terms of the special situation, we find it natural to generalize to a family of distributions. Having grown accustomed to the erroneous idea that arbitrary conditioning is meaningful, we take it for granted that conditioning a family is meaningful.

Actually, it is not so easy to imagine ways of obtaining the knowledge that a frequency distribution is in a certain family. The obvious way to get imperfect knowledge about a frequency distribution is to sample from it, but sample information does not lead to a family of distributions. It leads to estimates, confidence intervals, tests, and posterior distributions. Didier Dubois and Henri Prade argue that set-valued statistics do lead to families of distributions (with lower probability functions that happen to be belief functions), but statisticians will ask for clarification. Does the model include rules (albeit unknown) for how the range of uncertainty is determined? If so, set-valued statistics fall in the usual domain of mathematical statistics.

We can use families of distributions to represent vague guesses about frequencies, but the usefulness of this representation can be questioned. As critics often point out, it may be easier to make vague guesses into precise guesses than to make them into something equally precise and even more complex—a family of distributions. Often the problem with our guesses is not that they are too vague but that they may not be fully relevant. Pearl might be able to guess the rate of burglaries in Los Angeles area, but what does this tell him about the likelihood of a burglary of his home on this particular sunny afternoon?

Even if we do have knowledge that a frequency distribution is in a certain family, how can we justify Bayesian conditioning of the distributions in the family? It is important to remember that once our knowledge has been divorced from the step-by-step unfolding of events, we lack the strong justification of conditioning that is available in the special situation. Change of belief in accordance with the rule of conditioning is no longer an integral part of the model.

Bayesian theory gives up protocols for new information in favor of external judgments that the choice of information is not biased, but it retains protocols at least as an example—perhaps the best example—of such absence of bias. The fundamental point of Pearl's sandwich examples is that protocols are no longer a good example of absence of bias when we generalize to families of distributions.

Pearl presents his sandwich principle as a criticism of belief functions, but as Nic Wilson points out, the principle is also violated when we condition families of distributions. I agree with Wilson and Smets that the sandwich principle is unconvincing as a general principle of plausible reasoning, for the very statement of the principle involves the concept of conditioning, and this concept is meaningless outside the realm of probability. Even within probability, the appeal of the principle is limited to the situation in which the probability model says that our new knowledge will be either B or not B. When we have such a protocol, and we know that our belief in A will go up either way, it seems odd that it should not be up already. But when we do not have such a protocol—when what we condition on is chosen from a more diverse list of possibilities in an unbiased way—then what would happen if we were to condition on B or on not B does not loom so large.

When we generalize from a single distribution to a family of distribution, we generalize the idea of a protocol as well. We can no longer have a complete probability model for new information, but we can specify a partition of the space (such as B or not B) and specify that our new information will be knowledge of which element of the partition contains the truth. The examples that disturb Pearl suggest that such generalized protocols do not assure the absence of bias that can justify conditioning. This should not be surprising. Since the absence of bias is relative to our knowledge, making our knowledge more complicated may make it more difficult to be unbiased.

Partial Causal Models. If we retain the sequence of experiments that characterizes the special situation, but drop the assumption that we observe the outcomes of the experiments as they are performed, then we have what is often called a "causal model." If we also drop the assumption that we know fully the frequencies of outcomes for each experiment, assuming instead that we know only bounds for these frequencies, then we have a "partial causal model."

I am very sympathetic with the idea of a partial causal model, because it is a genuine generalization of the special situation, not merely a generalization of the idea of knowing a frequency distribution. I think that there is still work to be done, however, in understanding such models. We need to understand the extent to which the frequencies in the model can retain a belief interpretation, and the extent to which this can help justify the use of conditioning.

3.4. Conclusion

Bounds on probabilities and families of frequency distributions can be useful representations for evidence, but they are not as natural or ubiquitous a form of knowledge as Judea Pearl suggests.

The use of a frequency bound in a practical problem relies on the assumption that there is a correct frequency to bound—a correct reference class for our problem. This assumption is only sometimes reasonable. There may be no good reference class. There may be several reference classes—several different ways of describing a situation—that are relevant or partially relevant. We cannot necessarily say that one of them is the right one. They may simply be competing arguments, which must be weighed against each other without being combined into a single argument.

Belief functions are useful when different reference classes can be seen as relevant to different aspects of a problem, and hence can be combined. The idea is to treat reference classes for several different questions as independent arguments about a question for which we do not have a good reference class.

4. The Interpretation of Belief Functions

In this section, I compare my interpretation of belief functions with the interpretations proposed by Judea Pearl, by Philippe Smets, and by Enrique Ruspini, John Lowrance, and Thomas Strat.

I begin by discussing the transitory nature of the compatibility relations in the principle canonical examples for belief functions—the witness of uncertain reliability and the randomly coded message. A transitory compatibility relation applies only to the particular case, not to all the cases in the reference class that defines frequencies for the background probability space. My objections to the proposals by Pearl and by Ruspini, Lowrance, and Strat are based on the need for such transitory compatibility relations. Pearl's formulation—that belief functions express probabilities of provability—gives the impression that compatibility relations are always permanent. Ruspini, Lowrance, and Strat make the compatibility relation part of the probability model, thus reducing its transitoriness to random variation. This brings belief functions more firmly into the purview of formal probability theory, but it does so using a misleading and fictional superstructure.

I also discuss conditioning and combination. Didier Dubois and Henri Prade, as well as with Ruspini, Lowrance, and Strat, argue that belief functions use conditioning in the usual Bayesian sense, which is true. It is also true that the unbiasedness condition for Bayesian conditioning is a special case of the independence condition for Dempster's rule. The three prisoners puzzle involves essentially the same issues for belief functions as it does for Bayes.

I conclude this section by discussing Wilson's alternative rationale for Dempster's rule, Smets' rejection of any probabilistic interpretation of belief functions, and Larry Wasserman's call for an asymptotic theory for belief functions.

4.1. The Witness's Transitory Compatibility Relation

The witness of uncertain reliability provides a simple example of a transitory compatibility relation. The compatibility relation is established by what the witness says. The similar witnesses to whom we compare the witness in order to calibrate her credibility say different things and hence establish different compatibility relations. Even within the reference class in which we place the witness, the other witnesses are saying different things. The compatibility relation is transitory rather than fixed relative to repeated drawing from this reference class.

Recall, from my review article, the story of Betty's telling me that a tree fell on my car. Because I had found witnesses like Betty reliable 90% of the time in the past, her testimony gave me a 90% degree of belief that a tree had fallen on my car. This seems reasonable to me even though none of the earlier witnesses ever told me that a tree fell on my car.

Let me express the framework of the story of Betty formally, so that I can make my point about transitory compatibility relations in a more general way. My only evidence about Betty is that she was randomly chosen from a class of witnesses, p of whom are reliable, and 1-p of whom are not. I represent this evidence by a probability space (Ω,P), where

$$\Omega = \{\text{reliable, unreliable}\},$$

$P(\text{reliable})=p$, and $P(\text{unreliable})=1-p$. Betty's testimony, that a tree fell on my car, creates a compatibility relation between Ω and Θ, where

$$\Theta = \{\text{tree fell on my car, tree didn't fall}\}.$$

This compatibility relation can be expressed by the multivalued mapping Γ, where

$$\Gamma(\text{reliable}) = \{\text{tree fell on my car}\},$$

$$\Gamma(\text{unreliable}) = \{\text{tree fell on my car, tree didn't}\}.$$

The mapping Γ, together with the probability space (Ω,P), determines my belief function Bel on Θ. The probabilities given by P are full-fledged probabilities—frequencies *and* degrees of belief. Betty was chosen at random from a reference class of witnesses, so p is the frequency of reliable witnesses and my degree of belief that Betty is reliable. The frequency aspect of the interpretation is limited, however, to the reference class (Ω,P). It does not extend to the frame Θ, or to the compatibility relation between Ω and Θ. It is only on this particular occasion that the witness chosen at random testifies that a tree fell on my car and hence creates this particular compatibility relation.

The transitoriness of the compatibility relation is essential to the way this canonical example is used by the constructive theory of belief functions. When I make judgments about the reliability of a certain item of evidence (a handshake, a well-kept financial ledger, a vivid memory), I place this item of evidence in a reference class consisting of items of evidence in my past experience that had the same claim on my credence. The items in this reference class are similar but not identical to each other. In general, they bear on many different questions.

Belief functions can also be based on permanent compatibility relations, but transitory compatibility relations account for the most useful applications of belief functions. Thus the possible transitoriness of compatibility relations must be taken into account when we discuss the interpretation of belief functions. Such transitoriness limits the frequency interpretation of belief functions, and it therefore limits the sense in which belief functions can be said to express probability bounds or probabilities of provability.

4.2. Do Belief Functions Bound Probabilities?

Formally, any belief function Bel on a frame Θ is a lower probability function. If we let p be the family consisting of all probability distributions P on Θ such that $P(A)\geq Bel(A)$ for every subset A of Θ, then

$$Bel(A) = \inf \{P(A) \mid P\in \mathrm{p}\}.$$

But what interpretation do we put on the probabilities expressed by the distributions in p?

If Bel is based on a transitory compatibility relation, then these probabilities are not frequencies. Of course, we can make up a new story that makes them frequencies. I can imagine witnesses like Betty telling me many times that a tree fell on my car; I can

imagine different frequencies with which it is true; and I can assume that these frequencies are all greater than p. But if I used the transitory compatibility relation to model my evidence, then this new story has nothing to do with that evidence.

In the absence of a frequency interpretation, what might it mean to interpret the probabilities expressed by the distributions in p as beliefs? We could follow Isaac Levi and say that p is a class of distributions of belief among which we are undecided. But in the absence of a connection with frequency, I see no reason for assuming additivity of belief, and in any case, this family of distributions is superfluous to my assessment of my evidence.

My conclusion is that in the case of a transitory compatibility relation, the family p is a purely mathematical construct, with no conceptual significance.

4.3. Do Belief Functions Express Probability of Provability?

Probabilities for ω produce degrees of belief for θ because ω, together with the compatibility relation, implies that θ is in $\Gamma(\omega)$. The degree of belief Bel(A) is the total probability of all the ω that together with the compatibility relation imply that θ is in A. Pearl, equating this implication with logical proof, calls Bel(A) a "probability of provability." It is "an ordinary probability of a bona fide event—the existence of a logical proof for A" (pp. 367-368).

This way of talking has its uses, but I am uncomfortable with it, because it gives the impression that compatibility relations are always permanent. Since the "proof" a transitory compatibility relation supports applies only to the particular case, and not to other cases in the reference class underlying the belief function, we are not dealing with the probability of an event, as that phrase is usually interpreted. The event does not happen with a certain frequency in the reference class. If Betty is a reliable witness, then her statement that a tree fell on my car proves that a tree fell on my car. Since witnesses like Betty are reliable 90% of the time, this gives me a 90% degree of belief that a tree fell on my car. But is this the probability of proving that a tree fell on my car? Will listening to such witnesses as I go through life prove 90% of the time that a tree fell on my car? No, because (I hope) they will not always tell me that a tree fell on my car.

My fear that Pearl's "probability of provability" makes the compatibility relation seem permanent is confirmed by his article's line of reasoning. After emphasizing examples of compatibility relations that are permanent and therefore determine bounds on frequencies, Pearl concludes (p. 386) that compatibility relations are representations of categorical domain knowledge—representations of "stable relationships that govern entities in a domain." But compatibility relations need not express stable relationships. They can use the circumstances of the particular case (what the witness said, what these accounting records say, the apparent state of this bird's wings) to relate frequency knowledge to the question of interest.

4.4. The Randomly Coded Message

The canonical example of the witness of uncertain reliability is sufficient to calibrate simple support functions, but sometimes we want to assess more complex belief functions directly from evidence. We want to say that a given item of evidence (a measurement, an apparent erasure, a certain kind of static) can mean any of several

things, with different probabilities. The canonical example of the randomly coded message generalizes the canonical example of the witness in order to deal with such evidence. Like the canonical example of the witness, it uses a transitory compatibility relation.

Here is what I mean by a randomly coded message. We are interested in a question whose possible answers constitute a frame Θ. Betty, who knows the true answer, decides to tell us a subset L of Θ that contains it. She selects L arbitrarily but honestly; the true answer really is in L. She gives L to Sally, who sends L to us in encoded form, using a code ω_0, so that we actually receive $M=\omega_0(L)$. Sally selects ω_0 randomly, independently of the message L, from a set Ω of codes, following a probability distribution P on Ω. Sally does not tell us ω_0, but we know Ω and P. For simplicity, suppose we can decode M using any code ω in Ω, and suppose the result, $\omega^{-1}(M)$, is always a nonempty subset of Θ. Then M determines a compatibility relation between Ω and Θ, which can be expressed by the multivalued mapping Γ that maps each code ω in Ω to the subset $\omega^{-1}(M)$ of Θ. This, together with P, determines a belief function Bel on Θ. Our degree of belief in A,

$$\text{Bel}(A) = P(\{\omega|\omega^{-1}(M)\ldots A\}),$$

is the total probability of the codes according to which the message would allow us to conclude that the true answer is in A.

The compatibility relation here depends on M, the encoded message, which depends in turn on L, the true message. Betty chooses L arbitrarily, and she varies L arbitrarily when the story is repeated. Thus the compatibility relation is again transitory. Here, as in the case of the witness, this transitoriness is appropriate for applications. The probabilities for the codes correspond to the frequencies with which similar evidence in the past should have been interpreted in various ways. Since the details of the past instances vary, the implications of the different interpretations vary.

4.5. Permanent Compatibility Relations

Some compatibility relations are permanent, in the sense that they do express stable relationships.

The simplest example of a permanent compatibility relation arises when known frequencies are adopted directly as a belief function. We can think of this as the case where Ω and Θ are equal, and Γ, rather than being multivalued, is simply the identity mapping from to Ω to Θ, so that Bel=P. In this case, the probabilities determine the truth about the question that interests us, not just the meaning of the evidence.

We should not, of course, automatically interpret in this way any belief function that happens to be additive. If the different codes in the story of the randomly coded message all produce incompatible messages when applied to the message received, then the belief function obtained will be additive, even though the compatibility relation is transitory. As Nic Wilson reminds us, additive belief functions can also result from infinite weights of evidence.

Permanent compatibility relations can also produce non-additive belief functions. This is the case where random drawing ω from Ω partly determines the true value of θ. The role of permanent compatibility relations in the theory of belief functions is clearly

limited, however. We cannot randomly determine the same θ twice. Very restrictive conditions are required to assure that two random partial determinations be consistent with each other. 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.

4.6. Conditioning and Combination

I usually treat conditioning as a special case of combination, because I consider the idea of combination more useful than the idea of conditioning for analyzing and sorting evidence. We can sort evidence more flexibly if we think of items of evidence as things to which we point—happenings or physical objects tagged with bits of knowledge or experience. It is more limiting to think of them as formal "propositions" or as subsets in frames already formulated. I agree with Philippe Smets, however, that conditioning can be put before combination in the formal theory of belief functions.

One way to do this is to permit the compatibility relation to fall in such a way as to rule out certain elements of the probability space Ω. If our witness says something we know to be false, for example, then this eliminates the possibility that she is reliable, and we will condition on the subset of Ω consisting of the single element "unreliable," which will then have probability one. Since this element of Ω is compatible with either element of Θ, our belief function on her evidence will be vacuous.

In the example of the randomly coded message, we can similarly relax the assumption that $\omega^{-1}(M)$ is always a nonempty subset of Θ. It may sometimes be empty, indicating that ω could not be the code Sally used. In this case, we must assume that Betty chooses the message L in a way that is not systematically misleading. Then, by the general principles of Bayesian conditioning, we can condition P on $\{\omega|\omega^{-1}(M)\neq\varnothing\}$. This leads to the belief function Bel given by

$$\mathrm{Bel}(A) = \frac{P(\{\omega|\omega^{-1}(M)\ldots A\})}{P(\{\omega|\omega^{-1}(M)\neq\varnothing\})} \ . \qquad (4.6.1)$$

This is the total probability of the codes that put the true answer in A, calculated using the conditional probabilities.

How can we elaborate the condition that Betty chooses the message L in a way that is not systematically misleading? The simplest way, I think, is to assume that Betty chooses L without any knowledge of the set of codes that Sally uses, and certainly without any knowledge of the particular code ω_0. If Betty selected L knowing ω_0, she might be able to select it so that $\omega_0(L)$ would decode to false statements using most of the codes in Ω. If Betty merely knew the probability space of codes (Ω,P), she might be able to choose L so as to make the normalization in (4.6.1) misleading. Of course, we can only exclude systematic bias. Betty might accidentally choose L in a way that misleads us on this particular occasion, but such is life. She will not make exactly the same mistake again. Since the compatibility relation is transitory, she will be sending a message about a different topic next time.

In the case of randomly coded messages, Dempster's rule is easy to justify. We consider two probability spaces of codes, (Ω_1,P_1) and (Ω_2,P_2), that Sally samples from independently. This is the same as sampling from the product space $(\Omega_1 x \Omega_2, P_1 x P_2)$. On this particular occasion, Sally uses the two codes to send true messages about Θ

from Betty and Jane. Sally uses the code ω_{01} that she draws from Ω_1 to encode a subset L_1 supplied by Betty and she uses the code ω_{02} that she draws from Ω_2 to encode a subset L_2 supplied by Jane In effect, Sally sends a joint message $L_1 \cap L_2$, encoded with the joint code, $(\omega_{01},\omega_{02})$. Since Betty's and Jane's choice of the messages is independent of Sally's choice of the codes, the absence of systematic bias in choosing L_1 and L_2 extends to absence of bias in choosing the way they interact. For example, ω_{01} does not influence Jane's choice of L_2. So we can say that $L_1 \cap L_2$ is chosen without systematic bias relative to $(\Omega_1 x \Omega_2, P_1 x P_2)$, just as L_1 was chosen without systematic bias to (Ω_1, P_1) and L_2 was chosen without systematic bias relative to (Ω_2, P_2). So we can apply (4.6.1) to $L_1 \cap L_2$ and $(\Omega_1 x \Omega_2, P_1 x P_2)$, and this gives the product belief function $Bel_1 \oplus Bel_2$.

The essential condition for Dempster's rule is independence between the two belief functions. This means independence of the background probability spaces and absence of any systematic bias in the way the compatibility relations interact. This condition is met easily in the case of the randomly coded messages. It may be more difficult to meet when one of the compatibility relations is permanent, for then the weight of the condition of absence of systematic bias must be borne fully by the transitory compatibility relation. We will see this in the three prisoners puzzle.

4.7. The Three Prisoners Puzzle for Belief Functions

The three-prisoners puzzle and other similar puzzles for Bayesian conditioning are puzzles for belief functions for two reasons: because Bayesian conditioning is a special case of Dempster's rule, and because it is sometimes thought that belief functions, as a generalization that overcomes limitations of the Bayesian theory, should be able to overcome the limitations of Bayesian conditioning.

Is Bayesian conditioning a special case of Dempster's rule in only a formal mathematical sense, or is it a special case in the sense that the conditions under which it is legitimate (absence of bias in choice of information) can be subsumed under the conditions under which Dempster's rule is legitimate (independence, including absence of bias in the interaction of the compatibility relations)?

The formal description of Bayesian conditioning as a special case of Dempster's rule is straightforward. We have two belief functions on a frame Θ. The first, Bel_1, is additive. The second, Bel_2, puts mass one on the subset B of Θ. The product $Bel_1 \oplus Bel_2$ gives the same degrees of belief as the conditional probability function $Bel_1(\cdot|B)$. In the case of the three prisoners puzzle,

$$\Theta = \{Art,Bob,Carl\}.$$

The additive belief function Bel_1 gives degree of belief one-third to each element of Θ, and

$$B = \{Art,Bob\}.$$

The product $Bel_1 \oplus Bel_2$ gives degree of belief one-half to Art and degree of belief one-half to Bob.

As we learned in Section 2, there are two ways to legitimize conditioning on B. We can assume that B is learned in accordance with a protocol. Or, more generally, we can assume merely that B is chosen in some way that is not systematically biased. The guard in the three prisoners problem violates both assumptions by always choosing B to

contain Art. In order to compare this with the conditions for Dempster's rule, we need to talk about the background probability spaces, (Ω_1,P_1) and (Ω_2,P_2), and their compatibility relations with Θ, expressed by multivalued mappings Γ_1 and Γ_2. And we have to interpret the multivalued mappings as transitory or permanent compatibility relations.

Implicit in most comments on the puzzle (including the comments by Pearl and most of the discussants) is the assumption that Γ_1 represents a permanent compatibility relation. The probabilities one-third each for three prisoners represent the idea that someone chose at random which prisoner to execute, and repetition means repeating this random choice. So we simply make (Ω_1,P_1) a copy of (Θ,Bel_1), with Γ_1 the identity mapping from Ω_1 to Θ. We can construct (Ω_2,P_2) by making Ω_2 consist of a single point, with probability one; Γ_2 maps this single point to the subset B of Θ. Thus the choice of B defines the compatibility relation for the second item of evidence. Since B is chosen in a biased way, and since the first compatibility relation is fixed, the interaction of the compatibility relations is biased. Thus the bias that makes Bayesian conditioning illegitimate also makes Dempster's rule illegitimate.

Most of the discussants, following Diaconis [2] and Pearl, present the problem in a different way. They enlarge the frame Θ so that it tells both who will be executed and what the guard will say, and they vacuously extend the probabilities of one-third each for who will be executed to a belief function on this larger frame. Then they ask whether we can condition this belief function on what the guard says. In this setting, we have a generalized protocol, in the sense that I discussed in Section 3. Does the fact that we are working with such a generalized protocol mean that conditioning is justified? Evidently not, if we retain the assumption that the compatibility relation is permanent.

I do not agree with Nic Wilson that the problem lies in the formal additivity of the belief function. It lies in the interpretation of that additivity. If we interpret the additivity as a sign of a permanent compatibility relation (the victim is chosen at random), then we are essentially in the domain of probability bounds, and the discussion of Section 3 applies. But if the additivity were the accidental result of a transitory compatibility relation, then Dempster's rule might be appropriate.

4.8. Ruspini's Probabilistic Model for Belief Functions

Enrique H. Ruspini [14] has advanced a probabilistic model for belief functions that reduces the conditions for Dempster's rule to conventional formal independence conditions. As Ruspini, Lowrance, and Strat explain in their contribution to this discussion, belief-function degrees of belief are both probabilities of provability and bounds on probabilities in this model.

As the most thorough elaboration of the “probability of provability” idea (though it dates from 1987, before the idea was taken up by other authors),.Ruspini's model merits attention. It casts valuable light on the theory of belief functions. But it sets up a superstructure of fictional and uninterpretable probabilities. A sorting out of which probabilities in the model are meaningful, in the sense of Matheron [11], will lead back, I believe, to the less elaborate approach I favor.

To explain, I will put the canonical example of the witness of uncertain reliability into Ruspini's framework. I will first consider the single witness Betty telling me a tree fell on

my car, and then I will consider Betty and Sally both telling me something about whether a tree fell on my car.

In order to put Betty's into Ruspini's framework, we need, at the least, a probability distribution P for $(\underline{\omega},\underline{\theta},\underline{B})$, where $\underline{\omega}$ is a variable that tells whether Betty is reliable, with values in

$$\Omega = \{\text{reliable},\text{unreliable}\},$$

$\underline{\theta}$ is a variable that tells whether a tree fell on my car, with values in

$$\Theta = \{\text{did},\text{didn't}\},$$

and $\underline{B}$ is a variable that tells what Betty says on the subject, with values in

$$\mathfrak{b} = \{\{\text{did}\}, \{\text{didn't}\}, \Theta\}.$$

Here I am resorting to abbreviation. By

$$\underline{\theta}=\text{did}$$

I mean that a tree did fall on my car. By

$$\underline{B}=\{\text{did}\},$$

I mean that Betty says a tree fell on my car.

In this setting, we have a supercompatibility relation—a compatibility relation between Ω and Θ for each possible value of the testimony $\underline{B}$. This supercompatibility relation can be expressed by the mapping Γ^* from $\Omega x\mathfrak{b}$ to $\mathfrak{b}$ given by

$$\Gamma^*(\text{reliable},\{\text{did}\}) = \{\text{did}\},$$

$$\Gamma^*(\text{reliable},\{\text{didn't}\}) = \{\text{didn't}\},$$

$$\Gamma^*(\text{reliable},\Theta) = \Theta,$$

$$\Gamma^*(\text{unreliable},\{\text{did}\}) = \Theta,$$

$$\Gamma^*(\text{unreliable},\{\text{didn't}\}) = \Theta,$$

$$\Gamma^*(\text{reliable},\Theta) = \Theta.$$

In general, $\Gamma^*(\omega,B)$ is the subset of Θ consisting of the values we would consider possible for $\underline{\theta}$ if we knew that ω was Betty's reliability and B was her testimony. The distribution P must give probability one to a reliable Betty telling the truth: $P(\underline{\theta}\in\Gamma^*(\underline{\omega},\underline{B}))=1$.

Suppose Betty says that the tree fell on my car. In other words, we observe $\underline{B}=\{\text{did}\}$. We condition P on this observation. This gives the conditional distribution $P(\cdot|\ \underline{B}=\{\text{did}\})$ for the still unknown variables $\underline{\omega}$ and $\underline{\theta}$. Using this conditional distribution, we define a belief function Bel on Θ. For each subset A of Θ, we write

$$\begin{aligned}\text{Bel}(A) &= P(\Gamma^*(\underline{\omega},\underline{B})\ldots A \mid \underline{B}=\{\text{did}\}),\\ &= P(\Gamma^*(\underline{\omega},\{\text{did}\})\ldots A \mid \underline{B}=\{\text{did}\}).\end{aligned} \qquad (4.8.1)$$

This is the probability given by $P(\cdot|\ \underline{B}=\{\text{did}\})$ to the event that A is proven by {did}, Betty's known testimony, together with $\underline{\omega}$, her still unknown reliability.

To see that Bel is a belief function, note that it depends only on the marginal of $P(\cdot|\ \underline{B}=\{\text{did}\})$ for $\underline{\omega}$. Write Q for this marginal. Then define a multivalued mapping Γ from Ω to subsets of Θ by

$$\Gamma(\omega) = \Gamma^*(\omega,\{\text{did}\}).$$

This is the same multivalued mapping Γ we studied in Section 4.1 above, and (4.8.1) can be written.

$$\text{Bel}(A) = Q(\{\omega \mid \Gamma(\omega) \ldots A\}).$$

So Bel is the belief function determined by the probability space (Ω,Q) and the multivalued mapping Γ. If we write p for Q(reliable), then we get Bel(did)=p and Bel(didn't)=0. So Bel differs from the belief function we obtained in Section 4.1 only in that it is based on conditional probabilities for Betty's reliability given what she said rather than on frequencies for the reliability of similar witnesses.

Now consider the problem of combining the testimony of Betty and Sally. Here we need a joint distribution P for $(\omega_1,\omega_2,\theta,\underline{B}_1,\underline{B}_2)$, where ω_1 and $\underline{B}_1$ represent Betty's reliability and testimony, and ω_2 and $\underline{B}_2$ represent Sally's reliability and testimony. We can formulate a mapping Γ^*_1 to express the condition that a reliable Betty tells the truth, and an analogous mapping Γ^*_2 to express the condition that a reliable Sally tells the truth. The mapping Γ^* given by

$$\Gamma^*(\omega_1,\omega_2,\theta,B_1,B_2) = \Gamma^*_1(\omega_1,B_1)\cap\Gamma^*_2(\omega_2,B_2)$$

then expresses both conditions. The marginal of P for $(\omega_1,\underline{B}_1)$ conditioned on the observed value of $\underline{B}_1$, together with Γ^*_1, determines a belief function Bel_1. Similarly, the marginal for $(\omega_2,\underline{B}_2)$ conditioned on the observed value of $\underline{B}_2$, together with Γ^*_2, determines a belief function Bel_2. And the marginal for $(\omega_1,\omega_2,\underline{B}_1,\underline{B}_2)$, conditioned on both observed values, together with Γ^*, determines a belief function Bel.

Under what conditions on P will Bel always be the result of combining Bel_1 and Bel_2 by Dempster's rule, no matter what the observations are? The answer, as Ruspini, Lowrance, and Strat explain, is that the events

$$\Gamma^*_1(\omega_1,\underline{B}_1)=A_1 \text{ and } \Gamma^*_2(\omega_2,\underline{B}_2)=A_2 \qquad (4.8.2)$$

must be independent with respect to P whenever their conjunction has positive probability. This comes down to saying that whether Betty is reliable and what she says is independent of whether Sally is reliable and what she says, within the constraint that they cannot both be reliable and say contradictory things.

The simplest instance of (4.8.2) is

$$\Gamma^*_1(\omega_1,\underline{B}_1)=\{\text{did}\} \text{ is independent of } \Gamma^*_2(\omega_2,\underline{B}_2)=\{\text{did}\},$$

which translates into the condition that

Betty is reliable and says a tree fell on my car

is independent of

Sally is reliable and says a tree fell on my car.

Notice that we do not say that

$$\Gamma^*_1(\omega_1,\underline{B}_1)=\{\text{did}\} \text{ is independent of } \Gamma^*_2(\omega_2,\underline{B}_2)=\{\text{didn't}\},$$

because the events

Betty is reliable and says a tree fell on my car

and

Sally is reliable and says a tree didn't fall on my car

cannot both happen; their conjunction has zero probability.

Perhaps this is a reasonable probability model, provided we have evidence on which to base all the probabilities. But if we do have this evidence, and we do construct all the probabilities in the model, then Pearl's objection to the belief functions that we construct

within the model is very cogent. Why should we be interested in the probability that A is provable rather than the probability that A is true?

Ruspini, Lowrance, and Strat respond to this question with a scenario in which a statistician observes many instances of given "evidential conditions" and records only the frequency with which certain things were proven, not the frequency with which they were true. In the example of Betty, this would mean that the statistician observes many instances where a witness like Betty tells me a tree fell on my car, and records each time whether the witness was generally reliable, though not whether a tree fell on my car. Why should the statistician act this way? I think we must take the story as an extension of the canonical example. It is merely a story to which we compare our evidence. We compare our evidence to knowledge of frequencies that such a quirky statistician might have kept.

Unfortunately, there is more to the story than the frequencies that the quirky statistician might have kept. There are also the frequencies he did not keep—the probabilities for θ and $\underline{B}$. On what evidence are these probabilities based? We have no evidence for them. Ruspini, Lowrance, and Strat express this by saying that we cannot measure them. I think it more appropriate to say that they do not exist and we cannot make them up. There is no reference class, no evidence, no repeatable experiment to which we can appeal to define them. They are purely fictional. This is why I do not want to say that Bel(A) is a lower bound for P(A). There is no P(A).

When we bring these fictional probabilities for θ and $\underline{B}$ into our model, they weaken the probabilities that do have an evidential basis. In the context of a model that includes probabilities for a tree falling on my car, how can I ignore them in assessing the probability of Betty's reliability given that she said a tree fell on my car? Should I suppose that

Betty is reliable and says a tree fell on my car

is independent of

Sally is reliable and says a tree fell on my car

when both are correlated with the event, also in the probability model, that a tree falls on my car?

To me, the constructive theory of belief functions is a way of using probability moderately—a way of making some probability judgments without creating so many fictional probabilities that speculation about them swamps what little evidence we have. This is why we should avoid Ruspini's model as a foundation for the constructive theory of belief functions. Yet the model deserves further theoretical attention. Sorting out within it which probabilities are meaningful might be one way for us to gain more experience in the use of belief functions.

4.9. Nic Wilson's Justification of Dempster's Rule

Nic Wilson, in his reply to my article, reviews very concisely the rationale for Dempster's rule based on the canonical examples of the witness or the randomly coded message. He is dissatisfied with this rationale, however, and he advances an alternative justification for Dempster's rule for simple support functions, based on judgements about conditional probabilities after hearing the witnesses.

Wilson's rationale for Dempster's rule is very close to Ruspini's. Wilson does not posit probabilities directly for whether the tree fell on my car, but he makes the same a posteriori independence assumptions about the reliability of the witnesses as Ruspini makes. These a posteriori independence assumptions are, of course, satisfied by the joint probability distribution for the witnesses' reliability that underlies Dempster's rule, so the only difference between Wilson and me is whether it is more convincing to advance them directly or to justify them in terms of conditioning a single initial judgment of independence. I prefer conditioning a single initial judgment of independence because this allows a frequency interpretation farther into the story and because it minimizes the use of the slippery idea of conditional probability. Because conditional independence is so slippery (in the absence of a protocol), we can find plausible both Wilson's a posteriori independence and opposite judgments advanced by others (e.g., Freeling and Sahlin [6]).

4.10. Smets's Escape From Probability

Philippe Smets advocates the complete avoidance of probability in the interpretation of belief functions. He prefers that belief functions stand on their own.

The theory of belief functions can certainly stand on its own as a mathematical theory. Here are certain axioms and rules; here are their consequences. In order to use the theory, however, we need something more. At the very least, we need canonical examples with which to compare and calibrate actual evidence. Ideally, these canonical examples should motivate the axioms and rules.

My 1976 book suggested such canonical examples for weights of evidence, which are related logarithmically to the degrees of belief for simple support functions, and which therefore combine additively. We combine different items of evidence by adding weights that score their strength. We need a scale of measurement for these weights, and I suggested using statistical likelihoods to provide this scale. (Pearl starts to rediscover this idea in Section 4 of his article.)

The canonical examples that I have emphasized in my articles since 1976 have tied probability more directly to degrees of belief. My motivation for doing this was to allow belief functions more direct access to frequency evidence, and to allow belief functions to use people's willingness to compare non-frequency evidence to frequency evidence.

I am uncertain about Smets's attitude towards canonical examples. In practice, he sometimes uses frequencies in this role. He appeals to “Hacking's principle” to justify equating his degrees of belief with frequencies when frequencies are known, and this creates a standard that can be used to calibrate non-frequency evidence. But he stops short of using such examples to justify the axioms and rules of the theory. He does not, for example, provide any justification of the normalization involved in Dempster's rules of conditioning and combination. As I explained earlier, this normalization is explained by an appeal to Bayesian conditioning when we use the witness of uncertain reliability or the randomly coded message as our canonical examples. It also comes out in the wash when we use weights of evidence; here it is part of a convenient translation of the weights of evidence to an alternative numerical scale.

While commenting on Smets's views of normalization, I want to deny that I make a closed world assumption. The choice between an open world and a closed world does

not enter into the general theory of belief functions; it is a choice we make when we set up the frame of discernment Θ for a particular problem. By including an element in Θ that is not impugned by any of our items of evidence, we forestall any normalization.

I would also like to register my disagreement with the idea that belief functions should additive in order to reach a "pignistic level" where decisions can be made. I agree with Ruspini, Lowrance, and Strat that we should not pretend to knowledge or evidence we do not have simply to make our decision making look tidier.

4.11. Asymptotics for Belief Functions

I share Larry Wasserman's interest in developing an asymptotic theory for belief functions, not because this would bring belief functions closer to Bayesian theory, but because it would root belief functions more firmly in a tradition much older and deeper than Bayesian theory.

The belief-function methods that have been proposed for statistical inference tend to share the asymptotic properties of Bayesian and sampling-theory methods for the same problems. If we simply use a consonant belief function with plausibilities of singletons proportional to the joint likelihood of the observations, then the expected concentration of the likelihood on the correct parameter value will imply degrees of belief approaching one for neighborhoods of that value. Alternatively, if we translate individual observations into belief functions using some other method that still makes the plausibilities of singletons proportional to the likelihoods, and then combine these belief functions by Dempster's rule, then the result will agree asymptotically with a Bayesian analysis.

As I have already argued, however, this familiar setting is not the most appropriate for belief functions. I would be more interested in asymptotic results for the situation where the accumulation of evidence involves the investigation of more and more related questions, with a steady enlargement of the frame. There should be conditions of unbiasedness and haphazardness on the distribution of compatibility relations (we need some chaos here, not necessarily a probability distribution) and on the selection of propositions whose degree of belief we examine, which would enable us to relate the degrees of belief for these propositions to the frequency with which they are true.

5. Monte Carlo Implementation of Dempster's Rule

I do not fully share Nic Wilson's enthusiasm for Monte Carlo implementation of Dempster's rule. My conversations with Augustine Kong, who has explored extensively the use of Monte Carlo in Bayesian and belief-function networks, suggest that the method is useful but that it is not a panacea.

As Wilson points out, the computational cost of the Monte Carlo implementation tends to be proportional to the frame size. But it is also proportional to κ, where $\log(\kappa)$ is the weight of conflict. Wilson assumes that κ is bounded, and hence he treats it as a constant. In the examples that have interested me, however, the weight of conflict grows with the size of the problem and becomes the limiting factor in the usefulness of the Monte Carlo approach. When the weight of conflict is large, the Monte Carlo approach is infeasible, because only a tiny fraction of the trials give usable information.

Wilson argues that an extremely high weight of conflict will not occur if the analysis is valid. If evidence is too conflicting, we should not be trying to combine it. I agree that too high a weight of conflict signals problems, but what is too high is relative. As we combine more and more items of evidence, we would expect a steady multiplicative increase in κ. Typically, we get large frames because we are increasing the number of items of evidence. As I explain in Section 8 of my review article, new items of evidence generally involve collateral questions that must be brought into the frame, especially if we are trying to sort our evidence into independent items. Consequently, I associate an increase in frame size with an increase of the number of items of evidence and hence an increase in the weight of conflict.

As Wilson points out, the Monte Carlo approach can be combined with propagation in a Markov tree. Locally in the tree, both the frame sizes and the weight of conflict will be lower, and hence the Monte Carlo approach will be more feasible. If the tree has large cliques or a large branching factor, however, even the local weight of conflict may make the Monte Carlo approach infeasible.

Another way to broaden the usefulness of the Monte Carlo approach might be to combine it with Bayesian approximations when the weight of conflict is large. Often, though not always, large weights of conflict signal a concentration of belief on disjoint subsets or even singletons. If we can recognize when this happens, then we may be able to reduce some aspects of the computation by working with the partitions formed by the disjoint subsets.

6. Apologies

When I undertook to write a review article on belief functions, I feared that I would overlook important work and make errors of attribution and interpretation. I am grateful, therefore, that the article was accompanied by the articles of Pearl, Strat, Dubois and Prade, and Provan, which added a number of significant references to those that I had given. I am also grateful to the discussants for pointing out some of my errors.

In response to the comments by Didier Dubois and Henri Prade, I would like to apologize for failing to absorb the probabilistic interpretation of their disjunctive analogue of Dempster's rule (Ref. 3), and for overlooking Smets contribution in this area (Ref. 30) as well as his contribution with Kennes (Ref. 8) on the fast Möbius transform.

There are numerous recent contributions that none of us have mentioned. Among these, I would like to call attention to Mary McLeish's work on belief functions and non-monotonic logic (Ref. 12) and to Jürg Kohlas's work on the mathematical foundations for belief functions on infinite spaces (Ref. 9).

Acknowledgements

Research for this response was partially supported by the National Science Foundation through grant IRI8902444 to the University of Kansas.

References

1. Dawes, R.M., Faust, D., and Meehl, P.E., Clinical versus actuarial judgment, *Science* 243, 1668-1674, 1989.

2. Diaconis, P., Review of "A Mathematical Theory of Evidence," by Glenn Shafer, *Journal of the American Statistical Association* 73, 677-678, 1978.
3. Dubois, D., and Prade, H., A set-theoretic view of belief functions. *International Journal of General Systems* 12, 193-226, 1986.
4. Finetti, B. de , *Theory of Probability*, 2 vols., Wiley, New York, 1974-1975.
5. Freedman, D., and Lane, D., A nonstochastic interpretation of reported significance levels, *Journal of Business and Economic Statistics* 1, 292-298, 1983.
6. Freeling, A.N.S., and Sahlin, N.-E., Combining evidence, in *Evidentiary Value: Philosophical, Judicial and Psychological Aspects of a Theory* (P. Gärdenfors, B. Hansson, and N.-E. Sahlin, Eds.) Lund, 58-74, 1983.
7. Jeffrey, R., *The Logic of Decision*, McGraw-Hill, 1965.
8. Kennes, R., and Smets, P., Computational aspects of the Möbius transform, *Uncertainty in Artificial Intelligence*, Cambridge, MA, 1990.
9. Kohlas, J., A mathematical theory of hints, Technical report, University of Fribourg, 1991.
10. Levi, I., *The Enterprise of Knowledge*, MIT Press, Cambridge, 1980.
11. Matheron, G., *Estimating and Choosing*, Springer-Verlag, Berlin, 1989.
12. McLeish, M., A model for non-monotonic reasoning using Dempster's rule. To appear in Uncertainty in Artificial Intelligence, Vol. 6, 1992.
13. Mosteller, F., *Fifty Challenging Problems in Probability*, Addison-Wesley, 1967.
14. Ruspini, E.H., The logical foundations of evidential reasoning, Technical note No. 408, Artificial Intelligence Center, SRI International, Menlo Park, California, 1987.
15. Savage, L. J., *The Foundations of Statistics*, Wiley, New York, 1954.
16. Shafer, G., *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, N.J., 1976.
17. Shafer, G., Constructive probability, *Synthese* 48, 1-60, 1981.
18. Shafer, G., Lindley's paradox (with discussion), *Journal of the American Statistical Association* 77, 325-351, 1982.
19. Shafer, G., Bayes's two arguments for the rule of conditioning, *Annals of Statistics* 10, 1075-1089, 1982.
20. Shafer, G., Belief functions and parametric models (with discussion), *Journal of the Royal Statistical Society*, Series B 44, 322-352, 1982.
21. Shafer, G., Review of "The Enterprise of Knowledge," by Isaac Levi, *Technometrics* 24, 164-165, 1982.
22. Shafer, G., Conditional probability (with discussion), *International Statistical Review* 53, 261-277, 1985.
23. Shafer, G., Probability judgment in artificial intelligence and expert systems (with discussion), *Statistical Science* 2, 3-44, 1987.
24. Shafer, G., The unity of probability, in *Acting Under Uncertainty: Multidisciplinary Conceptions*, (G. von Furstenberg, Ed.), Kluwer, 95-126, 1990.
25. Shafer, G., The unity and diversity of probability (with discussion), *Statistical Science* 5, 435-462, 1990.
26. Shafer, G., Can the various meanings of probability be reconciled? In *Methodological and Quantitative Issues in the Analysis of Psychological Data*, Second Edition, (G. Keren and C. Lewis, Eds.), Lawrence Erlbaum, Hillsdale, New Jersey. To appear.
27. Shafer, G., The early development of mathematical probability, in *Encyclopedia of the History and Philosophy of the Mathematical Sciences*, (I. Grattan-Guinness, Ed.), Routledge, London. To appear.

28. Shafer, G., What is probability? In *Perspectives on Contemporary Statistics*, (D. C. Hoaglin and D. S. Moore, Eds.), Mathematical Association of America. To appear.
29. Shafer, G., and Tversky, A., Languages and designs for probability judgment, *Cognitive Science* 9, 309-339,1985.
30. Smets, P., Un modèle mathético-statistique simulant le procesus du diagnostic médical, Doctoral Dissertation, Free University of Brussels, 1978.
31. Walley, P. (1991), *Statistical Reasoning with Imprecise Probabilities*, Chapman and Hall, London, 1991.