9

Constructive Probability*

Glenn Shafer

In a series of papers published in the 1960's, A. P. Dempster developed a generalization of the Bayesian theory of statistical inference. In *A Mathematical Theory of Evidence*, published in 1976, I advocated extending Dempster's work to a general theory of probability judgement. The central idea of this new general theory is that we might decompose our evidence into intuitively independent components, make probability judgements based on each component, and then extend, adapt, and combine these judgements using formal rules. In this way we might be able to construct numerical degrees of belief based on total evidence that is too complicated or confusing to deal with holistically. The systems of numerical degrees of belief that the theory helps us construct are called *belief functions*. Belief functions have a certain structure, but they are not, in general, additive like Bayesian probability distributions: a belief function Bel may assign a proposition A and its negation $\overline{A}$ degrees of belief $Bel(A)$ and $Bel(\overline{A})$ that add to less than one.

The theory of belief functions should be sharply distinguished from the ideas on "upper and lower probabilities" that have been developed by I. J. Good [11], C. A. B. Smith [28], and, more recently, Peter Williams [30, 31]. It is true that the theory's degrees of belief $Bel(A)$ have some properties in common with these authors' lower probabilities $P_*(A)$. And it is also true that Dempster, in his writing, used the vocabulary of upper and lower probabilities. But the conceptual structure of the theory of belief functions is quite different from the structure underlying Good, Smith, and Williams' work.

Since its publication, *A Mathematical Theory of Evidence* has been reviewed or discussed by several authors, including Persi Diaconis [4], Terry

* From *Synthese* 48 (1981) 1–60, © 1981 by D. Reidel Publishing Co., Dordrecht, Holland, and Boston, U.S.A. The author has corrected some typographical errors and updated some references. The original research was partially supported by grant MCS 78-01887 from the National Science Foundation. The basic ideas of Sect. 1 were formulated in discussions with Amos Tversky. The author also profited from discussions with Zeno Swijtink and correspondence with Peter Williams.

Fine [5], Isaac Levi [16], Dennis Lindley [17], Teddy Seidenfeld [20], and Peter Williams [32]. Most of these critics, being themselves dissatisfied with the Bayesian theory, have welcomed the new theory. But they have been troubled by the absence of a behavioral interpretation for the theory. The Bayesian theory can appeal to its "betting interpretation" to explain what its degrees of belief mean and to justify its rules for these degrees of belief. No such interpretation has been supplied for the theory of belief functions. So what do its degrees of belief mean? And why should we accept the theory's rules for these degrees of belief? Why, in particular, should we prefer these rules to the rules suggested by Good, Smith, and Williams?

In this paper, I argue that a constructive theory of probability judgment need not rely for its meaning and justification on any behavioral interpretation. My argument is based on an understanding of constructive probability judgment developed in recent unpublished work by Amos Tversky and myself. According to this understanding, numerical probability judgment amounts to comparing one's evidence to a scale of canonical examples, and a constructive theory of probability judgment must supply both the scale of canonical examples and methods of breaking the task of comparison down into simpler judgments. As I explain in Sect. 1 below, the Bayesian theory, the theory of belief functions, and a theory of lower probability functions can all be developed in this framework. All three of these constructive theories use the idea of chance in their scale of canonical examples. The theory of belief functions uses examples where the meaning of a message depends on chance, while the other two theories use examples where the truth is generated by chance.

In the course of the paper I give particular attention to Peter Williams' review of *A Mathematical Theory of Evidence*. Williams' writing is exceptionally lucid, and he is exceptionally explicit in relating his criticisms of the theory of belief functions to the betting interpretation of probability.

Williams treats both lower probabilities and Bayesian (i.e., additive) probabilities as betting rates. And he hints that his intuitions about lower probabilities are inherent in the very idea of betting. One of the purposes of this paper is to show that this is not so. The theory of belief functions is as consistent with the use of probability judgments as betting rates as the theory of lower probabilities Williams favors. It is especially important to recognize that one cannot choose between the different rules of conditioning used by belief functions and by Williams' theory (see Sect. 3 below) on the basis of the idea of betting alone.

1 The Meaning of Probability

Williams begins his review of *A Mathematical Theory of Evidence* with two questions: "(i) What is meant by 'degree of belief' and how might an individual determine his degrees of belief in a particular case? (ii) For what reasons are degrees of belief required to satisfy the conditions imposed?"

On a practical level, making a probability judgment means assessing the strength and significance of one's evidence by fitting it into a scale of canonical examples. And the probability judgment or "degree of belief" itself means that we have made the comparison—perhaps with the aid of some theory—and found our evidence to match a certain example on the scale best. Thus the meaning of a degree of belief depends on the scale we use and, more generally, the theory we use in arriving at it.

To make numerical probability judgments we need, of course, a numerical scale, and the obvious approach to constructing such a scale is to use examples involving chance. There is, however, more than one way of using the idea of chance to construct a scale of examples, and different ways correspond to different theories of probability judgment. It will be helpful, before going into Williams' questions more fully, to compare three such theories—the Bayesian theory, the theory of belief functions, and a theory of lower probabilities.

1.1 The Bayesian Theory

In the classical picture of chance, we imagine a game that can be played repeatedly and for which we know the chances. These chances are long-run frequencies, they can be thought of as propensities, and they also define fair betting rates—rates at which a bettor would break even in the long run. Since they are known and there is no other evidence, these chances give a measure of how much reason we have to believe that one or another of the game's outcomes will occur on a particular occasion. So we can call them numerical degrees of belief. If we imagine a number of different games, with different chances, then we have a scale of numerical degrees of belief.

The Bayesian theory uses this scale in a straightforward way. The Bayesian's task is to compare his problem to a scale of examples in which the truth is generated according to known chances and to decide which of these examples is most like his problem. And so when he makes the probability judgment $P(A) = p$, say, he is saying that his evidence provides support for A comparable to what would be provided by knowledge that the truth is generated by a chance setup that produces a result in A exactly p of the time. He is not saying that his evidence is just like such knowledge in all respects, nor that the truth is in fact a result of chance. But he is measuring the strength of his evidence by comparing it to a scale of chance setups.

How can the Bayesian accomplish his task? How can he make his scale of chances and the affinity of his evidence to this scale vivid enough to his imagination that he can meaningfully locate the evidence on the scale? This question does not, I believe, have a simple general answer. In any particular case the Bayesian must struggle to find ways of understanding his evidence that facilitate its comparison to the scale of chances. Perhaps he can understand his evidence in terms of a causal model and assess numerically the propensity of the model to produce various outcomes. Perhaps he can discern relevant frequencies in his evidence. And perhaps he can make enough well-founded judgments of these sorts to enable him to construct an overall probability distribution that seems well-founded to him. Or perhaps he cannot. There is nothing in the Bayesian theory that can guarantee its success.

The probability distributions of the Bayesian theory have, of course, exactly the same structure as chance distributions: a function P defined for all subsets of a finite set Θ (*the frame of discernment*) is a *Bayesian* (or *additive*) probability distribution if there exist non-negative numbers $p(\theta)$ for the elements θ of Θ such that

$$P(A) = \sum_{\theta \in A} p(\theta) \tag{1}$$

for all $A \subset \Theta$. (It is also required that $\sum_{\theta \in \Theta} p(\theta) = 1$.) In words: the degree of belief $P(A)$ that the truth lies in A is the sum over the elements θ of A of the degrees of belief $p(\theta)$ that the truth is θ.

1.2 The Theory of Belief Functions

A function *Bel* defined for all subsets of a frame Θ is called *a belief function* if it is of the form

$$Bel(A) = \sum_{B \subset A} m(B), \tag{2}$$

where $m(B)$ are non-negative numbers satisfying $m(\phi) = 0$ and $\sum_{B \subset \Theta} m(B) = 1$. Every Bayesian probability distribution is a belief function. (The m-values for a Bayesian probability distribution P are obtained by setting $m(\{\theta\}) = p(\theta)$ and $m(B) = 0$ for all B that contain more than one element.) But not every belief function is a Bayesian probability distribution.

The theory of belief functions is based on a way of comparing our evidence to the scale of chances that is quite different from that of the Bayesian theory. Instead of comparing our evidence to a scale of examples where the truth is generated according to known chances, we compare it to a scale of examples where the reliability and meaning of a message depends on known chances.

Here is a way to develop the scale of examples needed for belief functions. Suppose someone chooses a code at random from a list of codes, uses the chosen code to encode a message, and then sends us the result. We know the list of codes and the chance of each code being chosen—say the list is $c_1, \ldots, c_n$, and the chance of c_i being chosen is p_i. We decode the encoded message using each of the codes and find that this always produces a message of the form "the truth is in A" for some non-empty subset A of Θ. Let A_i denote the subset we get when we decode using c_i, and set

$$m(A) = \sum \{p_i \mid 1 \le i \le n; A_i = A\}$$

for each $A \subset \Theta$. Then $m(A)$ is, in a certain sense, the total chance that the true message was A.[1] And $Bel(A)$, given by (2), is the total chance that the

[1] This is not to say that we are dealing with a random mechanism that produces the message A with chance $m(A)$. It is just that $m(A)$ is the sum of the chances for those codes that decode our encoded message to A.

true message implies A. If the true message is infallible and the coded message is our only evidence, then we will want to call $Bel(A)$ our degree of belief that the truth lies in A.

We can tell this story with whatever values of the $m(A)$ we please, and so it provides us a canonical example corresponding to every possible belief function Bel. Of course we will seldom or never encounter in practice a situation in which our evidence really does consist of a coded message and all the assumptions of the canonical example are satisfied. But it is also rare that our evidence amounts to knowledge of a chance distribution according to which the truth has been or will be generated. In both cases the canonical examples are meant not as realistic examples but as standards for comparison.

Our task, when we assess evidence using belief functions, is to choose values of $m(A)$ that make the canonical "coded-message" example most like that evidence. But how do we do this? In complicated problems it is absurd, surely, to suppose that we can simply look at our evidence holistically and write down the best values for the $m(A)$. So we need a theory—a set of tools for constructing belief functions from simpler, more elementary judgments. *A Mathematical Theory of Evidence* suggests a number of such tools: assessment using simple support functions, assessment using consonance, discounting, minimal extension, and Dempster's rule of combination. All these tools are readily intelligible in terms of the canonical examples.

Dempster's rule of combination is the most important single tool of the theory. This rule tells us how to combine a belief function Bel_1 (with m–values $m_1(A)$, say) representing one body of evidence with a belief function Bel_2 (with m-values $m_2(A)$) representing an unrelated body of evidence so as to obtain a belief function Bel (with m-values $m(A)$) representing the pooled evidence. The idea underlying the rule is that the unrelatedness of the two bodies of evidence makes pooling them like combining two stochastically independent randomly coded messages. We should, that is to say, combine the canonical examples corresponding to the two bodies of evidence by supposing that the two random choices of codes are stochastically independent. It is easy to see how this leads to a rule for obtaining the $m(C)$ from the $m_1(A)$

Let us denote by C the set of codes that decode our encoded message to A. If we had not yet seen the encoded message, it would certainly be natural to adopt $m(A)$ as our degree of belief that the code used is in C. The suggestion here is that it is still natural to do so in the situation where we have seen the encoded message and thus know that the code used being in C is equivalent to A being the true message.

A similar tack is often taken by non-Bayesian statisticians when they make probability judgments based on probability sampling or on randomization. Here, as in those cases, one might refuse to adopt the suggested degrees of belief and adopt instead a parametric model. In this case the model would have the true message as its parameter and the encoded message as its observable given each value of the parameter. In the absence of other evidence about the true message, this model does not seem very useful. (Cf. Kempthorne, [15].)

and the $m_2(B)$. Denote by $c_1, \ldots, c_n$ and by $p_1, \ldots, p_n$ the codes and their chances in the case of the first message, and by $c'_1, \ldots, c'_m$ and $p'_1, \ldots, p'_m$ the codes and their chances in the case of the second. Then independence means that there is a chance $p_i p'_j$ that the pair $\left(c_i, c'_j\right)$ of codes will be chosen. But notice that decoding may now tell us something. If the message A_i we get by decoding the first message with c_i contradicts the message B_j we get by decoding the second message with c'_j (i.e., if $A_i \cap B_j = \phi$), then we know that $\left(c_i, c'_j\right)$ could not be the pair of codes actually used. So we must condition the chance distribution, eliminating such pairs and multiplying the chances for the others by K, where

$$\begin{aligned} K^{-1} &= \sum \{p_i p'_j \mid 1 \leq i \leq n; 1 \leq j \leq m; A_i \cap B_j \neq \phi\} \\ &= \sum \{m_1(A) m_2(B) \mid A \subset \Theta; B \subset \Theta; A \cap B \neq \phi\}. \end{aligned}$$

Notice also that if the first message is A and the second message is B, then the overall message is $A \cap B$. Thus the total chance of the overall message being C is

$$\begin{aligned} m\,(C) &= K \sum \{p_i p'_j \mid 1 \leq i \leq n; 1 \leq j \leq m; A_i \cap B_j = C\} \qquad (3) \\ &= K \sum \{m_1\,(A)\, m_2\,(B) \mid A \subset \Theta; B \subset \Theta; A \cap B = C\}. \end{aligned}$$

Formula (3) is Dempster's rule.

The availability of Dempster's rule opens the possibility that we might construct a belief function based on complicated evidence by decomposing the evidence, breaking it down into small unrelated items whose message is relatively clear. The most convenient case, perhaps, is when each small item points clearly and unambiguously to a single subset of Θ. In this case the assessment of each item means the determination of a simple support function.

A *simple support function* focused on a subset A_0 of Θ and awarding it degree of support s is a belief function with m-values $m(A_0) = S$, $m(\Theta) = 1-s$ and $m(A) = 0$ for all other $A \subset \Theta$. This corresponds to a coded message which means A_0 with chance s and means Θ (i.e., means nothing at all) with chance $1 - s$. The values of the belief function are

$$Bel(A) = \begin{cases} 0 & \text{if } A_0 \not\subset A \\ s & \text{if } A_0 \subset A \neq \Theta \\ 1 & \text{if } A = \Theta. \end{cases}$$

In words: we have no positive beliefs beyond those implied by the degree of support s for A_0. Simple support functions are appropriate when the message of an argument or an item of evidence is clear and unambiguous, but its reliability must be assessed. The chance s corresponds, in such a case, to an assessment of that reliability. It is our assessment, so to speak, of the chance that the argument is sound.

The idea of the chance that an argument is sound (as opposed to the Bayesian idea of the chance that an assertion is true) is illustrated by the following example, which is essentially due to J. H. Lambert (see Shafer [22]) and which could be used to provide an alternative scale of canonical examples for simple support functions. Suppose we know all α's are β's, and we are told, by a randomizing device that tells the truth with chance s and lies with chance $1 - s$, that γ is an α. If the device told the truth (chance s), then we have a syllogism:

$$\begin{array}{c}\text{All } \alpha\text{'s are } \beta\text{'s.}\\ \underline{\gamma \text{ is an } \alpha.}\\ \gamma \text{ is a } \beta.\end{array}$$

If the device lied (chance $1 - s$), then we have nothing, for when the minor premise in the syllogism Barbara is negated, there is no conclusion:

$$\begin{array}{c}\text{All } \alpha\text{'s are } \beta\text{'s.}\\ \underline{\gamma \text{ is not an } \alpha.}\\ \text{Maybe } \gamma \text{ is a } \beta\text{; maybe not.}\end{array}$$

So the argument for the proposition "γ is a β" is sound with chance s and unsound with chance $1 - s$. As evidence, it amounts to the same thing as a message that asserts this proposition with chance s and says nothing with chance $1 - s$.

There is no guarantee that a satisfactory analysis of one's evidence will be achieved using belief functions, just as there is no guarantee of success with the Bayesian theory. I do believe, however, that the greater flexibility of belief functions will often be valuable. In many cases our deliberation needs to be directed towards the structure and reliability of the evidence rather than towards the nature of the process by which the truth is generated, and this means that a random model for the evidence may fit our needs better than a random model for the truth.

1.3 Lower Probabilities

Suppose we know a certain process is governed by chance, but instead of knowing precisely the chance law P governing it, we know only that P is in a class $\mathcal{P}$ of chance laws. Denote by Θ the set of possible outcomes for the process. Then we might set our degree of belief that the outcome of a given trial will be in a subset A of Θ equal to

$$P_*(A) = \inf \{P(A) | P \in \mathcal{P}\} . \tag{4}$$

This seems natural because we know the chance of A is at least $P_*(A)$. And so, in particular, we can expect to at least break even in the long run if we offer to bet (with others who have no more knowledge than we) on A at the odds $P_*(A) : 1 - P_*(A)$.

By varying the class $\mathcal{P}$ in this story we obtain a scale of examples. Perhaps we can construct a theory of probability judgment—a "theory of lower probabilities"—using this scale as the standard to which to compare our evidence. It will rarely if ever happen, of course, that our evidence really consists of knowledge that the truth is generated by chance and the chance law is in a class $\mathcal{P}$. But we have said the same thing about the canonical examples underlying the Bayesian theory and the theory of belief functions.

But what are the elements of this theory of lower probabilities? What tools do we have for locating our evidence on its scale of canonical examples? How, that is to say, do we break the task of constructing the class $\mathcal{P}$ down into simple judgments?

Here is an idea. Suppose we assess our evidence by making judgments of the form "our evidence is like knowing that the truth is generated by chance and that the chances have such-and-such a property." Since there are many properties of chance distributions, this formulation permits a wide variety of judgments. We may say that our evidence is like knowing that the chance of A is greater than the chance of B, or like knowing that the conditional chance of A given C is greater than that of B given C, or like knowing that the mathematical expectation of some function of the truth is between certain bounds, etc. Our theory will ask us to make as many of these judgments as we think necessary to capture the message of the evidence, and $\mathcal{P}$ will consist of all the distributions that have all the properties we have specified.

Notice that this idea does not involve the decomposition of evidence. The task of constructing $\mathcal{P}$ is broken down into simple judgments by distinguishing different questions, not by distinguishing different items of evidence bearing on these questions. All the judgments are supposed to be based on the total evidence.

A class $\mathcal{P}$ of chance distributions determines, of course, more than the lower probabilities (4). It also determines *lower conditional probabilities*

$$P_*\left(A|B\right) = \inf\left\{P\left(A|B\right)|P \in \mathcal{P}; P\left(B\right) > 0\right\}, \tag{5}$$

which are defined whenever $P(B) > 0$ for some $P \in \mathcal{P}$,[2] and *lower expectations*

$$E_*\left(X\right) = \inf\left\{E_P\left(X\right)|P \in \mathcal{P}\right\},$$

which are defined (in the case where Θ is finite) for every real-valued function X on Θ. Since a lower unconditional probability is a special case of a lower conditional probability ($P_*\left(A\right) = P_*\left(A|\Theta\right)$) and a lower conditional probability can be determined from knowledge of lower expectations ($P_*\left(A|B\right) = p$ if

[2] De Finetti [8] assumes that $P(A|B)$ is defined for an additive probability distribution even if $P(B) = 0$, and Williams [30] accordingly supposes that $P_*(A|B)$ is always defined. But it is not necessary to explore these subtleties in the present discussion.

$E_*(X) = 0$, where $X(\theta) = 1 - p$ if $\theta \in A \cap B$, $-p$ if $\theta \in \overline{A} \cap B$, and 0 if $\theta \in \overline{B}$), we obtain more information about $\mathcal{P}$ as we pass from lower probabilities to lower conditional probabilities to lower expectations.

Example 1. Here are two classes $\mathcal{P}_1$ and $\mathcal{P}_2$ that have the same lower unconditional probabilities but can be distinguished by their lower conditional probabilities. Set $\Theta = \{a, b, c\}$, $\mathcal{P}_1 = \left\{P | P(\{a, b\}) \geq \frac{1}{2}\right\}$, and $\mathcal{P}_2 = \{P | P(\{b\} \mid \{b, c\}) \geq \frac{1}{2}\ \}$. Then $P_{*1}(A) = P_{*2}(A)$ for all $A \subset \Theta$. But $P_{*1}(\{b\} \mid \{b, c\}) = 0$, while $P_{*2}(\{b\} \mid \{b, c\}) = \frac{1}{2}$. (2) Here are two classes that have the same lower conditional probabilities but can be distinguished by other lower expectations. Set $\Theta = \{-2, -1, 1, 2\}$, set $\mathcal{P}_1 = \{P | E_P \geq 0\}$, where E_P denotes the mean of the distribution P, and set $\mathcal{P}_2 = \mathcal{P}_1 \cup \{P_2\}$, where P_2 is the distribution that puts mass $\frac{1}{2}$ on -2, $\frac{1}{3}$ on 1, and $\frac{1}{6}$ on 2. Then $P_{*1}(A|B) = P_{*2}(A|B)$ for all A and B, but the lower expectations of the identity function $X(\theta) = \theta$ are $E_{*1}(X) = 0$ and $E_{*2}(X) = -\frac{1}{3}$. (3) Here are two distinct classes that cannot be distinguished by their lower expectations. Set $\Theta = \{a, b\}$, $\mathcal{P}_1 = \{P | P(\{a\}) \geq .5\}$, and $\mathcal{P}_2 = \{P | .5 \leq P(\{a\}) \leq .6 \text{ or } P(\{a\}) \geq .9\}$.

Let us call a function P_*, defined for all $A \subset \Theta$, a *lower probability function* if it is given by (4) for some class $\mathcal{P}$. And let us call a function of two variables $P_*(A|B)$ a *lower conditional probability function* if it is given by (5) for some class $\mathcal{P}$; such a function is defined for $B = \Theta$ and for all other $B \subset \Theta$ such that $P_*(\overline{B}|\Theta) < 1$. In general, as we have seen, there are many classes that yield the same lower probability function or lower conditional probability function. But the largest class that yields a given lower probability function P_* is

$$\mathcal{P}(P_*) = \{P | P(A) \geq P_*(A) \ \text{ for all } A \subset \Theta\}, \tag{6}$$

and the largest class that yields a given lower conditional probability function $P_*(\cdot|\cdot)$ is

$$\mathcal{P}(P_*(\cdot|\cdot)) = \left\{P | \text{ if } P_*(\overline{B}|\Theta) < 1, \text{ then } P(B) > 0 \text{ and } P(A|B) > P_*(A|B)\right\}. \tag{7}$$

Lower probability functions have been characterized axiomatically by Williams [31], Huber [14], and Wolf [33]. I have not seen simple axioms for lower conditional probability functions, but see Williams [30].

Our "theory of lower probabilities," as I have described it so far, includes in its scale of canonical examples every possible class $\mathcal{P}$ of chance distributions over a frame Θ. For the theory allows us to specify an arbitrary property of a chance distribution and to say that our evidence is like knowing that the truth is generated according to chances having that property. Perhaps this is too rich a scale. In practice there will surely be a limit to the complexity and subtlety of properties that can sensibly be said to correspond to intuitive insights about our evidence. And it may be desirable, from a psychological point of view, for the theory to recognize this explicitly by specifying a somewhat sparser scale. It cannot help us in fitting our evidence to a scale of canonical examples to have that scale encumbered with confusing and superfluous possibilities.

Just what classes $\mathcal{P}$ should be included in the theory's scale? I see no definitive answer to this question, but it does seem that an adequate scale should include all $\mathcal{P}$ that can be defined by the sorts of constraints commonly placed on chance distributions—all that can be defined, say, by (1) bounds on chances, conditional chances, and expectations, (2) comparisons among chances and conditional chances, and (3) conditions of independence and conditional independence. This is a rich scale. It includes far more $\mathcal{P}$ than those of the form (6) or (7), and far more, even, that those that can be defined by bounds on expectations. (As we have already noted, bounds on chances and conditional chances can be reduced to bounds on expectations. Moreover, some comparisons can be reduced to bounds: the condition $P(A) > P(B)$, for example, is equivalent to $P\left(A \cap \overline{B} | A \triangle B\right) \geq \frac{1}{2}$, or simply to $P\left(A | A \cup B\right) \geq \frac{1}{2}$ if $A \cap B = \emptyset$. But conditions of independence and comparisons of the form $P(A|B) \geq P(A)$, say, go beyond bounds on expectations.)

Notice that if we were content with a scale consisting of $\mathcal{P}$ of the form (7), then the lower conditional probability function $P_*(\cdot|\cdot)$ would completely identify $\mathcal{P}$ and hence would be a complete report of our assessment of our evidence. If we agree, as I think we must, that a richer scale is necessary, then $P_*(\cdot|\cdot)$ cannot be regarded as a complete assessment. But it might be an adequate summary for some purposes.

1.4 The Literature on Lower Probabilities

The idea of constructing a class of distributions by comparing our evidence to knowledge that the truth is generated according to chances having certain properties is an adaptation of an idea developed by I. J. Good [11]. Good suggests that we pretend we have an additive probability distribution P in a black box. Initially we know nothing about P, except that it is defined for subsets of a frame Θ. But we make qualitative probability judgments about Θ, and we interpret these judgments as constraints on P. For example, we judge that A is more probable than B, and we interpret this as $P(A) > P(B)$. Or we judge that we would think A more probable than B if we knew C for certain, and we interpret this as $P(A|C) > P(B|C)$. If we manage to keep these constraints from conflicting, then they determine a non-empty set $\mathcal{P}$ of additive probability distributions.

Unfortunately, Good does not say that we are comparing our evidence with knowledge that the truth is generated by some chance law in $\mathcal{P}$. Instead he studiously avoids pinning down the nature of the unknown probability distribution P—he locates P in a "black box" precisely in order to avoid saying whether it is a chance law, a hidden subjective distribution, or something else. I believe this deliberate vagueness is untenable in a constructive theory. It leaves us uncertain about how to make the qualitative probability judgments and uneasy about whether we really want to interpret these judgments as constraints on P. We cannot make even qualitative probability judgments unless we have a definite language in which to work.

Most other recent literature on lower probabilities seems less relevant to our constructive view. Smith [28] and Williams [30, 31] study lower probabilities as betting rates, but as I argue in Sect. 2 below, it is difficult to relate talk about betting to constructive probability judgment. Huber's work on lower probabilities [13, 14] is mainly concerned with situations where the truth's being generated by chance is a serious hypothesis and not just a metaphor. For further references, see Shafer [22].

1.5 Belief Functions and Lower Probabilities

Mathematically, every belief function is a lower probability function. Every function of the form (2), that is to say, is also of the form (4). Here is one way to see this. Given a belief function Bel on a frame Θ, we can construct an additive probability distribution P such that $P(A) > Bel(A)$ for all $A \subset \Theta$ by choosing an element θ_B of every non-empty subset B of Θ and setting

$$p\left(\theta\right) = \sum \left\{ m\left(B\right) \left| \theta_B = \theta \right. \right\}.$$

Let $\mathcal{P}$ denote the class of distributions obtained by varying the choice of the θ_B. Then $P(A)$ is smallest for those P in $\mathcal{P}$ that choose θ_B to be outside A whenever possible—i.e., whenever $B \not\subset A$. So

$$\inf \left\{ P\left(A\right) \left| P \in \mathcal{P} \right. \right\} = \sum \left\{ m\left(B\right) \left| B \subset A \right. \right\} = Bel\left(A\right).$$

Not every lower probability function, on the other hand, is a belief function; Williams exhibits an example of one that is not on page 380 of his review.

Does the fact that every belief function is a lower probability function mean that our theory of lower probabilities is more general than the theory of belief functions? Certainly not. For the theory of belief functions uses a belief function in a different way than our theory of lower probabilities would use it. The meaning is quite different in the two cases. One theory is comparing our evidence to knowledge provided by a randomly coded message; the other is comparing our evidence to knowledge about chances governing the truth. I will discuss some of the implications of this difference in meaning in Sects. 3 and 5 below.

Since it does retain the Bayesian idea that our evidence is like knowing that the truth is generated by chance, our theory of lower probabilities is much closer in spirit to the Bayesian theory than the theory of belief functions is. And, as we shall see in Sect. 3 below, it does not escape as thoroughly as one might think from the Bayesian emphasis on prior probabilities.

I will not surprise the reader when I say that I find belief functions more interesting and promising than lower probabilities. In many cases, I believe, our evidence is so unlike knowledge that the truth is generated by chance that it is misleading to liken a conviction that the evidence supports A better than B to knowledge that the chance of A is greater than the chance of B.

I hope, on the other hand, that the theory of lower probabilities I have sketched here is more than a straw man. It is quite possible that judgments of the kind the theory suggests will sometimes provide the most useful and insightful way to analyze one's evidence. And, as I shall try to show in this paper, the theory provides explicit motivation for assumptions that Good, Smith, and Williams have taken for granted in their writings on lower probabilities.

1.6 What is a Degree of Belief?

What is meant by "degree of belief," and how might an individual determine his degrees of belief in a particular case?

The meaning of an "epistemic probability" or "degree of belief" is very rich. It depends, I have argued, on the whole theory by which the probability judgment is made or, as we might put it, on the whole language in which it is expressed. A degree of belief of .3, say, means one thing in the Bayesian theory and something different in the theory of belief functions. It also depends on the canons of judgment that have been established in the particular field of inquiry. A historian's valuation of certain kinds of evidence may differ from a judge's.

There is room for ambiguity in the question about how an individual might "determine his degrees of belief." Some Bayesians give the impression of thinking that we have numerical probabilities for everything hidden in our psyche; they would interpret "determine" as a synonym for "elicit." Others take a more constructive view; for them probability judgment is a matter of assessing evidence and constructing reasonable numerical beliefs. As I have tried to make clear, I subscribe to the constructive view. Probability judgment is a matter of construction. We may come to the task with some vague beliefs, but these will not be numerically precise and will usually not even have any very definite structure. (It would be silly, for example, to argue about whether our unreflective beliefs have a structure more like belief functions or more like Bayesian probability distributions. There simply is not that much structure there.) And the process of construction should ideally be sufficiently fruitful in new insights and understanding as to render obsolete much of any rudimentary structure that might be in these initial vague beliefs.

1.7 Why Belief Functions?

For what reasons are degrees of belief required to satisfy the conditions imposed? Why, that is to say, should "belief functions" be required to be of the form (2) instead of, say, the more general form (4)?

As I see it, the theory of belief functions is a language in which one can construct and express probability judgments. Asking why the theory uses degrees of belief with a given structure is like asking why some aspect of a language's grammar is as it is. Explanations can be given, but they are inevitably internal explanations—explanations of how that aspect fits in with other aspects of the language. Challenged to explain why belief functions are required to be of the form (2), I might point out that only functions of this form can be combined by Dempster's rule. Or I might point out that functions of this form result when evidence is assessed using the scale of canonical examples involving randomly coded messages. But these are only internal explanations. They do not rule out the usefulness or even superiority of a different theory using a different and possibly more general structure for degrees of belief.

As I have tried to make clear, I do not deny the possibility of a theory superior to the theory of belief functions. I believe, though, that the superiority of one theory of probability judgment to another can be demonstrated only by a preponderance of examples where the best analysis using the one theory is more insightful than the best analysis using the other. As Amos Tversky puts it, the unit of comparison for theories of probability judgment is the individual analysis.

The individual analyses we compare should be complete analyses—analyses beginning with an intuitive account of one's actual evidence and building up formal judgments step by step. (Examples of such analyses using belief functions are given in Shafer [24] and Shafer and Breipohl [27].) It may be unfair to ask a theory to deal with a problem which has already been translated from actual experience into the language of another theory.

It would be unfair, for example, to argue that the very existence of a class $\mathcal{P}$ of chance distributions such that (4) is not a belief function is proof of the inadequacy of the theory of belief functions. For it is not the case that we can ever really know, in a concrete problem, that the truth is generated by chance in accordance with some distribution in a class $\mathcal{P}$. Rather, the determination of the class $\mathcal{P}$ must itself be regarded as the first step in one particular approach to constructing probability judgments. And so it proves nothing that the theory of belief functions may be unable to carry on from this first step. The important questions are: (1) Can a theory of lower probability functions show us how to carry out this first step insightfully? (2) In real examples where such a theory succeeds, can the theory of belief functions do as good or better using some other first step?

2 Betting

Since they use the picture of chance, our three constructive theories inevitably lead us to think about betting. But what exactly is the significance of betting for these theories?

Certainly we should not, in a constructive theory, interpret a probability judgment as an actual commitment to bet. Nor should we interpret it as a declaration that the person making the judgment has exactly the same attitude towards a bet in accordance with that judgment as he has towards a fair bet in a game of chance. Our relative equanimity about fair bets in games of chance is based on the assurance that the chances are objective facts and on the assurance that no possible opponent can gain an advantage over us through deeper understanding or knowledge of the game, and these elements are missing when we construct probability judgments on the basis of ordinary evidence. A probability judgment using the Bayesian theory, for example, is merely a judgment that our evidence is more similar in strength and significance to the evidence provided by knowledge of given chances than to the evidence provided by knowledge of different chances. We will not be happy unless we feel that the similarity is substantial and instructive and that our judgment is sound, but we will not pretend that the similarity is complete, nor that we are certain no one else could make a better judgment.

2.1 Long-Run Policies

So what are we saying about betting when we announce a probability judgment in one of our constructive theories? We are only saying, I think, that we judge our evidence to be similar to knowledge of a chance model where certain bets conform to a prudent long-run policy.

It is instructive to spell this out for each of our three theories.

- When we construct a Bayesian probability distribution P, we are judging our evidence to be like knowledge of a chance model where betting on A at the rate $P(A)$ conforms to a policy that breaks even in the long run. (If, for $i = 1, 2, \ldots$, a chance distribution P_i over Θ_i is used to generate an independent outcome $\theta_i \in \Theta_i$, and if on each occasion we choose a subset A_i of Θ and bet on it at rate $P_i(A_i)$, then we break even in the long run.)
- When we construct a belief function Bel, we are judging our evidence to be like knowledge of a chance model where betting on A at the rate $Bel(A)$ would conform to a policy that at least breaks even in the long run. (Consider a sequence of randomly and independently coded messages. Suppose the ith message bears on Θ_i. If we choose a subset A_i of each Θ_i, and if $Bel_i(A_i)$ turns out to be the total chance that the ith true message implies A_i, then we at least break even in the long run by betting on A_i at the rate $Bel_i(A_i)$.)
- When we construct a lower probability function P_*, we are judging our evidence to be like knowledge of a chance model where betting on A at the rate $P_*(A)$ would conform to a policy that at least breaks even in the long run. (If, for $i = 1, 2, \ldots$, a chance distribution P_i over Θ_i is used to generate an independent outcome $\theta_i \in \Theta_i$, and if on each occasion we choose a subset A_i of Θ_i and bet on it at a rate $P_{*i}(A_i) \leq P_i(A_i)$, then we at least break even in the long run.)

Notice that we can make statements for belief functions and lower probability functions that are identical on the surface. But in making these statements we have chance models and long-run policies in mind that are quite different in the two cases. A belief function and a lower probability function that are mathematically equivalent evoke the same bets in our actual problem, but they refer these bets to different chance models and embed them in different long-run policies.

Notice also that our statements about the long-run policies breaking even in the chance models are not quite theorems. They can be turned into theorems only by giving some mathematical form to the implicit assumption that our choice of the A_i is independent of the truth and of the random action of the model.

In formulating the statements about the models, I have been careful to embed each probability judgment in a sequence of judgments with different chance models and even different frames. For the chance model and the frame are constructed to represent the evidence in the problem at hand, and the next problem, and its evidence, will be different. If we were to allow ourselves to envision repeated trials using the same model (P, Θ), then we could make much stronger and more mathematically precise statements for the Bayesian and lower probability models. We could, for example, say the following:

- If a chance distribution P over Θ is used to generate a sequence $\theta_1, \theta_2, \ldots,$ of independent outcomes, and on each occasion we bet on $A \subset \Theta$ at the rate $P(A)$, then we will break even in the long run. In fact, we will break even even if we offer such bets for all $A \subset \Theta$ and let our opponents choose, on each occasion, which bets to accept.

But since P is a product of our particular problem, these strong statements are utterly irrelevant.

In the case of the chance model for belief functions there is no such temptation to talk about repetitions. For the belief function *Bel* is determined, in the model, by the random choice of a code and would vary even if the chance distribution for the code were kept fixed.

To summarize: Constructive probability judgments can be related to betting, but the relation is tenuous on two counts. It is tenuous because we are only comparing our evidence to a chance model. And it is tenuous because even in the model the bets can be justified only when embedded in a particular long-run policy involving other models.

2.2 The Dutch-Book Arguments

Williams must have a more intimate relation between probability and betting in mind when he writes about the "betting interpretation" of Bayesian degrees of belief and of lower probabilities and pleads for a similar "operational interpretation" for belief functions. But what more intimate relation can there be if we insist on a constructive understanding of probability judgment?

Williams' answer, apparently, is that our primary purpose in constructing probability judgments should be the setting of rates at which we will offer bets in accordance with some betting scheme.

There is, Williams reminds us, a betting scheme that seems to force a Bayesian structure on betting rates and another, looser one that seems to force the less restrictive structure of lower probability functions on them.

- Suppose we must choose, for each subset A of Θ, a betting rate $P(A)$ and then offer to take either side of a bet on A at odds $P(A) : 1 - P(A)$. Then an opponent can compile a book of bets from our offers that assures a net gain from us (a "Dutch book") if and only if the function P fails to be an additive probability distribution.
- Suppose we must choose a betting rate p_* for each A and then offer to bet on A at the odds $p_* : 1 - p_*$, but we are not required to offer to take the other side of the bet. Let $P_*(A)$ denote the greatest rate at which we have offered to bet on A—either explicitly or because such a bet can be compounded from our other offers. Then a Dutch book can be made against us if and only if P_* fails to be a lower probability function. (See Smith [28] or Williams [31]. Williams' proof of this result is especially elegant.)

But there does not seem to be a betting scheme in which the avoidance of Dutch book yields precisely the class of belief functions.

The Dutch-book arguments are interesting, but it is hard to accept the claim that the setting of betting rates in some particular betting scheme is the primary purpose of probability judgment.

It is often argued in this connection that every choice or action is like a bet and that probability judgments ultimately have no purpose other than to guide future choice and action.

But how well do human choices and actions fit the picture of a bet? How well, that is to say, do they fit the apparatus of "decision theory," where alternatives are weighed by the combination of probabilities and utilities? I believe that they do not fit very well. One way to understand why they do not fit is to recognize that utilities, like probabilities, do not simply exist. They are constructed. And in the case of utilities the construction is accomplished not so much by reflective thought as by our choices and actions themselves. It is only after a human being or a society of human beings has established a self-conception through crucial choices in a given domain that we can speak in any detail about his or its preferences in that domain. (For a review of some recent thinking about the inadequacy of decision theory, see March [18].)

Probability judgments should help guide our future choice and action, but it is also important to remember that the proximate purpose of probability judgment is always understanding. Human beings often seem to prize understanding for its own sake, and it is not easy to argue that this is always mere appearance. For it is only after we have gained understanding that we can formulate other goals.

Sometimes we are told that the Bayesian theory is a theory about the betting behavior of ideal rational agents, and that as such it is "normative"—it provides us with a definition of rationality that is so inherently attractive that we should try to conform to it, even if we cannot fully succeed. But surely this line of thought begs all the important questions. It is vacuous to call a mode of thinking or behavior an ideal unless it is appropriate to our needs and capabilities. And though the Bayesian theory is clearly a norm for behavior within a particular betting scheme, this does not make it a useful norm in ordinary thought and action.

I conclude that it is misleading to speak of a "betting interpretation" of probability. All three of our theories of probability judgment produce degrees of belief that can be used to set betting rates without fear of Dutch book. But this is only a minor aspect of their meaning.

2.3 Betting as a Tool in Probability Judgment

Another possible way of relating betting to probability might be to use introspection about betting as a tool in constructing probability judgments.

In the context of our three constructive theories, this would mean using such introspection to help us compare our evidence to canonical examples involving chance. We might try to locate the strength of our evidence on the scale of chances by asking ourselves at what odds our attitude towards a given bet would be comparable to our attitude towards a fair bet (Bayesian theory), or perhaps at what odds our attitude would be comparable to our attitude towards a bet we know to be at least fair (theory of lower probability functions). This might be more effective psychologically than trying to think about our evidence in terms of frequencies or propensities. The prospect of monetary loss or gain might concentrate our minds and thus permit a more honest and acute assessment of the strength of our evidence than we could obtain by thinking about it directly.

Here we have a reasonably sharp empirical question. Does it help people assess their evidence to think about betting? Or is it more helpful to think about frequencies or propensities? This question has not, perhaps, been investigated as directly as it might be. But the many empirical studies that have been made in this area do not seem to indicate that the betting metaphor is any more useful than the frequency metaphor, say, as a psychological aid in constructing degrees of belief.

I do not personally find that talk about betting concentrates my mind on my evidence; instead it tends to divert my mind to extraneous questions: my attitude towards the monetary and social consequences of winning or losing the bet, my assessment of the knowledge and astuteness of my opponent, etc. I find it inherently implausible, moreover, that I could better understand the strength of my evidence by asking myself about my willingness to bet. In a situation where I had somehow made a thorough and unimprovable but not fully conscious analysis of my evidence, it might be sensible for me to forget about the evidence and concentrate on my own hidden attitudes. But so far as I know, I do not make such unconscious analyses of evidence.

2.4 Lower Expectations

A function X which assigns a real number $X(\theta)$ to every $\theta \in \Theta$ can be thought of as a gamble: if $X(\theta) > 0$, then $X(\theta)$ is the amount we win; if $X(\theta) < 0$, then $-X(\theta)$ is the amount we lose. The idea of buying a gamble generalizes the idea of betting, for betting the amount p on A at the odds $p : 1-p$ means paying p to buy the gamble

$$X(\theta) = \begin{cases} 1 & \text{if } \theta \in A \\ 0 & \text{if } \theta \notin A. \end{cases}$$

Let us consider how each of our three theories would price a gamble.

- *Bayes.* If the truth is generated by chance in accordance with the chance distribution P, then the fair price for the gamble X is, of course, its expectation with respect to P, $E_P(X)$. Paying $E_P(X)$ for X is a policy that at least breaks even in the long run.
- *Belief Functions.* If we receive an infallible message that the truth is in $A \subset \Theta$, then we know the gamble $X(\theta)$ is worth at least $\inf\{X(\theta) | \theta \in A\}$ to us. So if we receive a randomly coded message and the chance of the message meaning A turns out to be $m(A)$ for each $A \subset \Theta$, then it is natural to price the gamble at the average value

$$\widehat{Bel}(X) = \sum_{A \subset \Theta} m(A) \left[\inf_{\theta \in A} X(\theta) \right]. \tag{8}$$

  Let us call $\widehat{Bel}(X)$ the *lower expectation* of X. It is a fair price to pay for X in the sense that we will at least break even if we pay such prices for gambles in a long run of independent randomly coded messages.
- *Lower probabilities.* Suppose we know the truth is generated by chance in accordance with some distribution in a class $\mathcal{P}$. Then we know the expectation of X is at least

$$E_*(X) = \inf_{P \in \mathcal{P}} E_P(X). \tag{9}$$

  And we will at least break even in the long run if we follow the policy of paying this price for X.

In Sect. 1 above I called (9) the lower expectation of X. Is it consistent to call both (8) and (9) by the same name? As it turns out, it is; if

$$\mathcal{P} = \{P | P(A) \geq Bel(A) \ \text{ for all } \ A \subset \Theta\},$$

then (8) and (9) will be equal. (See Huber [13] and Shafer [23].)

3 Conditioning

The idea of conditioning has its origin in the theory of chance.

Conditioning occurs most naturally, perhaps, in the case of a game of chance that unfolds step by step. When such a game has been only partly played out (when only the first die has been thrown, say), chance still has a role to play. And this role can be described by the conditional chance distribution. Suppose, indeed, that X denotes the set of complete outcomes for the game, and that the chance for each outcome x is denoted $p(x)$, so that the chance law P governing the game is given by

$$P(A) = \sum\{p(x)|x \in A\}$$

for all $A \subset X$. Say the partial playing out of the game determines only that the eventual outcome will be in the subset X_0 of X. Then the conditional chances $p'(x)$ governing the remainder of play are obtained by reducing the $p(x)$ for $x \notin X_0$ to zero and multiplying the $p(x)$ for $x \in X_0$ by the factor $P(X_0)^{-1}$. And the conditional chance distribution $P(\cdot|X_0)^{-1}$ is given by

$$P(A|X_0) = \sum\{p'(x)|x \in A\} = \frac{P\left(A \cap X_0\right)}{P\left(X_0\right)} \tag{10}$$

for all $A \subset X$. We can see that this is the right way to define the conditional chances by thinking about long-run frequencies: $P(A|X_0)$ is simply the proportion of the games that reduce to X_0 during the first stage of play that will go on to have their eventual outcome in A.

Conditioning can, of course, be applied in the case of any subset X_0 of X, even if X_0 does not correspond to a partial completion of the game. There are several ways of explaining what meaning conditioning might have in this more general case. One way is to turn our attention from the chances to the degrees of belief they justify. If we know the chance distribution P and have therefore adopted its values as our degrees of belief concerning how the game will turn out, then news that the outcome has fallen in X_0 will naturally lead us to revise our beliefs by (10). Of all the games in which this news is true, we will tell ourselves, $P(A|X_0)$ is the proportion in which the outcome is in A. And so adopting $P(A|X_0)$ as our new degree of belief seems reasonable, provided there is no trickery involved in our having received the news that the outcome is in X_0—provided, in other words, that our receipt of this news is not the result of some fiendish scheme to mislead us.

Now suppose we represent ordinary evidence by constructing degrees of belief over a frame Θ and then obtain new evidence whose direct effect on Θ is to establish with certainty that the truth is in a subset Θ_0. How should we change our degrees of belief to take this new evidence into account? Each of our constructive theories of probability has its own way of translating the rule of conditioning for chance distributions into an answer to this questions.

- *Bayes.* In the Bayesian case we have constructed an additive probability distribution P over Θ, with the understanding that our evidence is comparable to knowledge that the truth is generated by P. So we will simply adopt the conditional distribution $P\left(\cdot|\Theta_0\right)$ as our new additive probability distribution.
- *Belief functions.* In the case of belief functions, the chance distribution in our model is a distribution for the random choice of a code, and when we take the news that the truth is in Θ_0 into account, we have to condition this distribution on a subset of codes.

  Say we have represented our old evidence by a belief function Bel, corresponding to a randomly coded message with possible codes $c_1, \ldots, c_n$, where code c_i was used with chance p_i and decoding by code c_i produces the message $A_i \subset \Theta$. We can simply incorporate the news that the truth is in Θ_0 into the messages, thus changing A_i to $A_i \cap \Theta_0$. But we must also notice that the news may tell us something about which code was used: if $A_i \cap \Theta_0 = \emptyset$, then code c_i cannot be the code that was used. So in addition to changing A_i to $A_i \cap \Theta_0$ we must also condition the chance distribution for the codes on the subset $\{c_i | A_i \cap \Theta_0 \neq \emptyset\}$ of codes. This means we replace the p_i by p_i', where

$$p_i' = \begin{cases} 0 & \text{if } A_i \cap \Theta_0 = \emptyset \\ \frac{p_i}{\sum\{p_j | A_j \cap \Theta_0 \neq \emptyset\}} & \text{if } A_i \cap \Theta_0 \neq \emptyset. \end{cases}$$

  These two changes (replacing p_i with p_i' and A_i with $A_i \cap \Theta_0$) give us a new randomly coded message representing the total evidence. The belief function $Bel\left(\cdot|\Theta_0\right)$ corresponding to this randomly coded message has m-values

$$m\left(A|\Theta_0\right) = \sum\{p_i' | A_i \cap \Theta_0 = A\} = \frac{\sum\{p_i | A_i \cap \Theta_0 = A\}}{\sum\{p_i | A_i \cap \Theta_0 \neq \emptyset\}}$$

  for all $A \neq \emptyset$, and so

$$\begin{aligned} Bel\left(A|\Theta_0\right) &= \sum\left\{m\left(B|\Theta_0\right) | B \subset A\right\} \\ &= \frac{\sum\{p_i | A_i \cap \Theta_0 \subset A\} - \sum\{p_i | A_i \cap \Theta_0 = \emptyset\}}{1 - \sum\{p_i | A_i \cap \Theta_0 = \emptyset\}} \\ &= \frac{Bel\left(A \cup \overline{\Theta_0}\right) - Bel\left(\overline{\Theta_0}\right)}{1 - Bel\left(\overline{\Theta_0}\right)} \end{aligned}$$

  for all $A \subset \Theta$. This is the rule of conditioning for belief functions.
- *Lower probabilities.* Suppose we think the evidence bearing on a frame Θ is similar in strength to knowledge that the truth is generated by chance in accordance with some distribution in a class $\mathcal{P}$. Then we can take new evidence that the truth is in Θ_0 into account by saying that our total evidence is similar in strength to knowledge that the truth is generated

by chance in accordance with some distribution in the class $\mathcal{P}'$ obtained by conditioning on Θ_0 each element of $\mathcal{P}$ that can be so conditioned. In particular, we replace our lower probability P_* by P'_*, where

$$P'_*(A) = \inf \{P(A|\Theta_0)|P \in \mathcal{P}; P(\Theta_0) > 0\},$$

and we replace our lower conditional probability function $P_*(\cdot|\cdot)$ by

$$P'_*(A|B) = \inf \{P(A|B \cap \Theta_0)|P \in \mathcal{P}; P(B \cap \Theta_0) > 0\} = P_*(A|B \cap \Theta_0).$$

Notice that $P'_*(A|B)$ is undefined if $P_*(\overline{B \cap \Theta_0}) = 1$, in which case $P'_*(\overline{B}|\Theta_0) = 1$.

3.1 The Role of Conditioning

It should be emphasized that the decision to use the rule of conditioning in one of our constructive theories is itself a constructive judgment. We condition on B, as I have said, when the direct effect of new evidence on our frame Θ is to establish that the truth is in B. But whether this is the direct effect of the new evidence is a matter of judgment, not of fact. "The direct effect of the new evidence" is an idea that has reality only within our language of probability judgment. We learn the meaning of this idea by example, just as we learn the meaning of other elements of a language, and our application of the idea to particular evidence is, like other probability judgments, a comparison of that evidence with other examples.

The decision to condition is just one place where the idea of "the direct effect of given evidence" comes into play in the theory of belief functions. It also comes into play when we represent an item of evidence by a simple support function; in this case we must judge that the item's only direct effect on Θ is to support a given subset. And, as we shall see in Sect. 4 below, this is merely a special case of the judgment that the direct effect of given evidence on Θ is discerned by a given subalgebra.

The theory of belief functions is so concerned to identify the direct effect of given evidence because it often works with limited items of evidence. As I pointed out in Sect. 1 above, the fundamental strategy of the theory is to make judgments based on different items of evidence and then to combine these judgments. Conditioning is merely one example of such decomposition and recombination, and it is unusual only in that the message of one of the items of evidence is conclusive.

Theories which compare evidence to knowledge that the truth is generated by chance do not depend so extensively on the decomposition of evidence. Our theory of lower probabilities, for example, breaks the overall task of judgment down by distinguishing different questions, not by distinguishing different items of evidence bearing on those questions. We construct a lower probability function from many judgments of the form "our evidence is like knowing the chance of A to be greater than p," but it is "A" and "p" that vary from judgment to judgment, not the evidence; all the judgments are supposed to be based on the total evidence. In this theory, as in the Bayesian theory, it is only in the case of conditioning that we decompose our evidence, and so it is only in the case of conditioning that we are concerned with identifying the direct effect of a limited item of evidence.

How important is conditioning? Some Bayesians have given it a central role in their theory, perhaps because it is the only way their theory decomposes evidence and is hence the only way they can formally combine "new" evidence with old. (See, for example, de Finetti [8], p. 141.) But I am inclined to think of conditioning as a tool we will not use very often in a constructive theory. It will happen fairly often, no doubt, that we can formulate a frame and distinguish evidence whose direct effect is to establish that the truth is in a certain subset. But how often will this frame be the same as the one we have used or want to use in assessing the balance of our evidence? New evidence that we actually obtain after constructing numerical probability judgments over a frame Θ will seldom affect Θ so simply. And I also find it doubtful whether the assessment of a body of evidence already obtained will very often be best accomplished by singling out a part that establishes a subset B of a frame Θ, using the rest to construct degrees of belief over all of Θ, and then conditioning on B. It will usually, I think, be more sensible and efficient to treat knowledge of B as just another element of our background knowledge and to concentrate our probability judgments on matters that we really find uncertain. For a discussion of this point in the context of a detailed example, see Shafer [24].

One aspect of a decision to use conditioning in our constructive theories is the implicit judgment that the news that the truth is in B has not been selected from the many things we might be told just because it will interact with other evidence in such a way as to mislead us. This judgment can be translated into statements about the chance models used by the theories. In the Bayesian theory and the other theories that think of the truth as being generated by chance, the judgment comes down to saying that our new evidence is like learning the truth is in B by means of some mechanism that selects this message to send us without regard to the chances by which the truth was generated. In the theory of belief functions, the judgment comes down to saying that the selection of the message was without regard either to how the random coding of previous messages was set up or to how that random coding turned out. Notice that these statements assure, within the chance models, that betting in accord with the new degrees of belief remains a policy that at least breaks even in the long run.

3.2 A Comparison of Two Rules

The theory of belief functions and our theory of lower probabilities have very different rules of conditioning—rules that can give very different results even when applied to the same degrees of belief. We can gain insight into the difference between the two theories by studying a simple example of this divergence.

Let us first consider how the theory of belief functions conditions a simple support function. Suppose A_1 is a proper non-empty subset of Θ and we represent strong but inconclusive evidence that the truth is in A_1 by the simple support function

$$Bel\,(A) = \begin{cases} 0 & \text{if } A_1 \not\subset A \\ .95 & \text{if } A_1 \subset A \neq \Theta \\ 1 & \text{if } A = \Theta. \end{cases} \tag{11}$$

This belief function has m-values $m(A_1) = .95$, $m(\Theta) = .05$, and $m(A) = 0$ for all other A. In adopting it we are likening our evidence to a message that probably means A (chance .95) but might possibly (chance .05) mean nothing. Now suppose we obtain new evidence whose direct effect on Θ is to establish that the truth is in A_2, where A_2 is some other subset of Θ such that $A_1 \cap A_2 \neq \emptyset$. Then we condition Bel on A_2, obtaining

$$Bel\,(A|A_2) = \begin{cases} 0 & \text{if } A_1 \cap A_2 \not\subset A \\ .95 & \text{if } A_1 \cap A_2 \subset A \not\supset A_2 \\ 1 & \text{if } A_2 \subset A; \end{cases} \tag{12}$$

the news that the truth is in A_2 changes the message that it is probably in A_1 into the more specific message that it is probably in $A_1 \cap A_2$.

Let us make the story more concrete. Suppose a burglar is traced to a rooming house, in such a way as to make it highly probable that he is actually one of the roomers, though it is believed that he keeps his tools and loot elsewhere. A police detective searches the rooming house and interviews the five roomers, but on this first examination finds nothing that either exonerates or further incriminates any of them. At this point the detective might formulate a frame Θ which includes, for each roomer i, a subset B_i corresponding to the possibility that roomer i is the burglar. (See Fig. 1.) And he might adopt (11) as a representation of his evidence, where A_1 is the union of the B_i's.

Suppose now that roomers 4 and 5 produce airtight alibis, conclusively establishing that neither is the burglar. Such alibis, in order to be convincing,

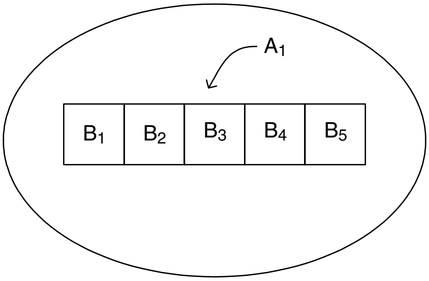

Fig. 1. Roomer i is the burglar

would have to involve great detail, and this detail would inevitably provide less conclusive evidence about other questions. But we may suppose that these other questions are not germane to the investigation and therefore need not be introduced into the frame Θ. Thus the detective may judge that the only direct effect of this new evidence on Θ is to eliminate B_4 and B_5 from consideration. In this case he will want to condition (11) on the set $A_2 = \overline{B_4 \cup B_5}$, which corresponds to the burglar being someone other than roomer 4 or roomer 5. The set $A_1 \cap A_2 = B_1 \cup B_2 \cup B_3$ corresponds to the burglar being one of the first three roomers. And according to the new belief function (12), the suspicion against the rooming house now points to these three.

Here is another way the story might go. Suppose the new evidence, instead of consisting of alibis, is evidence from the scene of the crime establishing that the burglar has blood type O. In this case the detective might introduce the question of the burglar's blood type into our frame Θ, so that there is a subset A_2 of Θ corresponding to its being type O. (This set A_2 is pictured in Fig. 2; since we do not yet know the roomers' blood types, A_2 intersects with each B_i.) And he will then condition (11) on A_2. The resulting belief function (12) awards degree of belief .95 to $A_1 \cap A_2$, which corresponds to the proposition that the burglar is one of the roomers and has blood type O. Under these circumstances the detective's next step will no doubt be to find out the blood type of each of the roomers and to condition (12) on this further information. I will refrain from illustrating this further conditioning graphically, because a very complicated picture arises when we introduce distinctions about each roomer's blood type into Θ. But the final result is obvious: if none of the roomers have type O blood then the suspicion against them is dispelled; otherwise it is focused on those that do.

One might challenge the adequacy of (11) and (12) as an analysis of this detective story on the grounds that there is probably other evidence that it does not take into account. Surely the detective acquired some hints and hunches in the course of interviewing the roomers. And might he not have some prior inclination to expect type O blood, given its high frequency in the

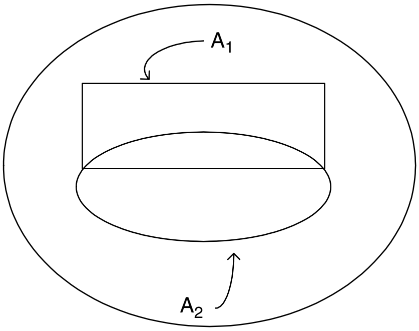

Fig. 2. The intersection of A_1 and A_2

population? The answer to this challenge is that the theory of belief functions can always accommodate further evidence, provided its relevance is identified and its value is assessed. The detective can decide he has further evidence worth introducing into the analysis, or he can decide he does not.

Let us now consider how to analyze the detective story using our theory of lower probabilities.

The most obvious approach is to liken the initial evidence in favor of A_1 to knowledge that the truth is generated by chance and that the chance of A_1 is at least .95. This means representing the evidence by the class $\mathcal{P} = \{P | P(A_1) \geq .95\}$ or by the lower probability function

$$P_*(A) = \begin{cases} 0 & \text{if } A_1 \not\subset A \\ .95 & \text{if } A_1 \subset A \neq \Theta \\ 1 & \text{if } A = \Theta, \end{cases} \tag{13}$$

which is mathematically identical to the belief function (11). But if we condition $\mathcal{P}$ on a subset A_2 that intersects both A_1 and $\overline{A_1}$, then we will obtain the new lower probability function

$$P_*'(A) = \begin{cases} 0 & \text{if } A_2 \not\subset A \\ 1 & \text{if } A_2 \subset A, \end{cases} \tag{14}$$

which indicates no particular support at all for $A_1 \cap A_2$. In fact, (14) seems to ignore the initial evidence. It is presumably the lower probability function we would adopt if we had only the new evidence establishing A_2.

It will be agreed, I think, that (14) is unsatisfactory. How is it to be avoided?

The natural move is to challenge the adequacy of the class $\mathcal{P} = \{P | P(A_1) \geq .95\}$ as a representation of our initial evidence. There is, one might argue, more to be said on the basis of the initial evidence than that the chance of A_1 is at least .95. In order to prepare for conditioning on the alibis of the two roomers for example, we might decide that the five roomers have equal chances of being the burglar, thus narrowing the class $\mathcal{P}$ down to the class

$$\mathcal{P}_1 = \{P | P(A_1) \geq .95; P(B_1) = P(B_2) = P(B_3) = P(B_4) = P(B_5)\}.$$

This already awards a lower probability of .57 to $B_1 \cup B_2 \cup B_3$. And when we condition $\mathcal{P}_1$ on $A_2 = \overline{B_4 \cup B_5}$, we obtain

$$\mathcal{P}_1' = \left\{P | P(A_2) = 1, P(A_1 \cap A_2) \geq \frac{.57}{.62} \approx .92; P(B_1) = P(B_2) = P(B_3)\right\},$$

which awards a lower probability of .92 to $A_1 \cap A_2 = B_1 \cup B_2 \cup B_3$. This is nearly as great as the degree of belief .95 awarded by the belief function (12). Notice, though, that this analysis is sensitive to the number of roomers and the proportion with alibis in a way that the analysis using belief functions is not. If four out of the five roomers have alibis, then the final lower probability for the remaining one would be only $\frac{.19}{.24} \approx .79$; if there were 20 and 19 were similarly exonerated, then the final lower probability for the remaining one would be $\frac{.0475}{.0975} \approx .49$. And these figures could easily be altered if we claimed that our initial evidence justified unequal prior chances for the roomers.

The initial class $\mathcal{P}$ can also be adapted to give sensible results when conditioned on the burglar's blood type. In this case the natural move is to narrow $\mathcal{P}$ down to

$$\mathcal{P}_2 = \{P | P(A_1) \geq .95; P(A_1 \cap A_2) = P(A_1) P(A_2)\} .$$

We require, that is to say, that A_1 and A_2 be independent. This is reasonable; once we have decided to think of the truth as random, it is natural to think of the random determination of the burglar's blood type as stochastically independent of the random determination of whether he is one of the roomers. Conditioning $\mathcal{P}_2$ on A_2 yields

$$\mathcal{P}_2' = \{P | P(A_2) = 1; P(A_1 \cap A_2) \geq .95\} ,$$

which gives a lower probability function mathematically identical to the belief function (12).

This last analysis can be extended to an analysis incorporating further conditioning on the roomers' blood types that will continue to agree with the analysis using belief functions. Here is the set-up. Let T denote the burglar's blood type, let T_i denote the ith roomer's blood type, and set

$$X = \begin{cases} 0 & \text{if the burglar is not one of the roomers} \\ i & \text{if the burglar is the } i\text{th roomer.} \end{cases}$$

(Notice that $T = T_i$ when $X = i$. And "$X \neq 0$" is equivalent to A_1.) Replace the initial class $\mathcal{P}$ by the class $\mathcal{P}_3$ consisting of all P such that $P(A_1) \geq .95$, $(X, T_1, .., T_5)$ are jointly independent with respect to P, all the T_i have the same marginal distribution, and T has this same distribution conditional on $X = 0$. We may take the burglar's and the roomers' blood types into account by conditioning $\mathcal{P}_3$ on the values of T and the T_i, and if there is a subset of roomers whose blood type agrees with the burglar's they will inherit the full .95 suspicion against the rooming house.

To summarize: A basic idea of the theory of belief functions is the idea of evidence whose only direct effect on the frame Θ is to support a subset A_1, and an implicit aspect of this idea is that when this evidence is combined with further evidence whose only direct effect on Θ is to establish a compatible subset A_2, the support for A_1 is inherited by $A_1 \cap A_2$. The theory of lower probabilities does not have a fully equivalent idea. New evidence establishing A_2 *may* cause prior support for a subset A_1 to be inherited by $A_1 \cap A_2$ in the theory of lower probabilities, but whether this happens will depend, as in the Bayesian theory, on various "prior probabilities."

Indeed, the similarity between our theory of lower probabilities and the Bayesian theory in their dependence on prior probabilities is striking. Our theory of lower probabilities does not, apparently, always get us away from the Bayesian bemusement over how to assess prior probabilities when the evidence is weak. In the case of our five roomers there was a natural symmetry on which to pin “equal prior probabilities,” but one could easily construct similar examples where there are no obvious symmetries or else competing ones, so that the prior probabilities needed in order to get sensible answers from conditioning seem much more arbitrary. This makes us wonder just how much is gained in the generalization from the Bayesian theory to the theory of lower probabilities.

However we answer this question, the drastically different results we get by conditioning (11) and (13) should bring home to us that a belief function can have quite a different meaning from a mathematically identical lower probability function. Saying our evidence is like a message that probably means A_1 but might mean nothing is quite different from saying it is like knowing that the truth is generated by chance and that the chance of A_1 is great. So we must decide when we make a probability judgment, just which formulation fits the significance of our evidence. We cannot simply make a vague judgment that the evidence supports A_1, express it numerically by (11), and then interpret (11) indifferently either as a belief function or as a lower probability function.

3.3 Conditional Bets

Consider again two proper subsets A_1 and A_2 of Θ such that $A_1 \neq A_2$ and $A_1 \cap A_2 \neq \emptyset$. Following de Finetti, let us call a gamble of the form

$$X(\theta) = \begin{cases} 1-p & \text{if } \theta \in A_1 \cap A_2 \\ -p & \text{if } \theta \in \overline{A_1} \cap A_2 \\ 0 & \text{if } \theta \notin A_2, \end{cases} \tag{15}$$

where $0 < p \leq 1$, a “bet on A_1 conditional on A_2.” The idea behind this name is that if we agree to this gamble (i.e., buy it for the price zero), then we will be betting on A_1 at odds $p : 1-p$ and total stakes $p + (1-p) = 1$, with the understanding that the bet will be called off if the truth turns out, when it is revealed, not to be in A_2.

In our constructive theories of probability judgment, our attitude towards a gamble depends, in the tenuous way discussed in Sect. 2 above, on the gamble’s expectation or lower expectation. This is true in particular of a conditional bet. If the expectation or lower expectation of the conditional bet is nonnegative, then the bet conforms, in the chance model we have used to represent our evidence, to a policy that at least breaks even in the long run.

Our attitude towards any gamble will, in general, change as we acquire new evidence. And in the theory of belief functions, our attitude towards a conditional bet can change dramatically when we obtain new evidence establishing the condition of the bet. Suppose, for example, that we have represented our evidence about Θ by the belief function (11). Then our lower expectation for the conditional bet (15) is

$$\widehat{Bel}\,(X) = .95\left[\inf_{\theta\in A_1} X\,(\theta)\right] + .05\left[\inf_{\theta\in\Theta} X\,(\theta)\right] = -.05p.$$

Since this is negative, the theory gives no sanction to the bet. But if we obtain new evidence establishing A_2 and change our belief function to (12), then the lower expectation changes to

$$\widehat{Bel}\,(X|A_2) = .95\left[\inf_{\theta\in A_1\cap A_2} X\,(\theta)\right] + .05\left[\inf_{\theta\in A_2} X\,(\theta)\right] = .95\,(1-p) + .05\,(-p)\,.$$

If $p < .95$, then this will be positive and so the theory will sanction the bet as reasonable policy. It is easy to see intuitively why our attitude towards the bet changes in this way. The bet is essentially a bet on $A_1 \cap A_2$, and the original evidence, while supporting A_1, does not provide any particular support for $A_1 \cap A_2$ until it is conjoined with the evidence establishing A_2.

Neither the Bayesian theory nor the theory of lower probabilities, in contrast, ever changes its willingness to sanction a conditional bet because of new evidence whose direct effect is to establish the bet's condition. Indeed, when we condition a Bayesian probability distribution P on A_2, the expectation of (15) changes only from $E_P(X)$ to

$$E_P(X|A_2) = \frac{E_P(X)}{P(A_2)};$$

it cannot change in sign. And when we condition a class $\mathcal{P}$ of distribution on A_2, the lower expectation of (15) changes only from

$$E_*\,(X) = \inf\left\{E_P(X)|P\in\mathcal{P}\right\}$$

to

$$E_*\,(X|A_2) = \inf\left\{\frac{E_P(X)}{P(A_2)}|P\in\mathcal{P}; P(A_2) > 0\right\},$$

and while this may be a change from zero to a positive quantity it cannot be a change from a negative to a non-negative quantity or vice-versa.

This contrast can also be expressed in terms of maximum rates for conditional bets. *The maximum rate for betting on A_1 conditional on A_2* is defined as follows:

- In the case of a Bayesian probability distribution P such that $P(A_2) > 0$, it is

  $$\sup\left\{p|E_P\,(X) \geq 0\right\},$$

  where X, which depends on p, is the conditional bet (15).

- In the case of a belief function Bel such that $Bel(\overline{A_2}) < 1$, it is

$$\sup \left\{ p | \widehat{Bel}\,(X) \geq 0 \right\}.$$

- In the case of a class $\mathcal{P}$ of distributions such that $p_*(\overline{A_2}) < 1$ (i.e., $P(A_2) > 0$ for some $P \in \mathcal{P}$), it is

$$\sup \{p | E_*\,(X) \geq 0\} = \sup \{p | E_P\,(X) \geq 0 \text{ for all } P \in \mathcal{P}\}.$$

These definitions all say the same thing: the maximum rate is defined except when we are certain the truth is not in A_2 (in which case the conditional bet is of no interest), and it is defined to be the greatest value p for which the bet is sanctioned. In general, a bet on A_1 conditional on A_2 is sanctioned in one of the constructive theories only if the bet's value for p is less than or equal to this maximum rate. Thus the contrast between belief functions and the other two theories can be expressed by saying that the maximum rate for betting on A_1 conditional on A_2 may change when one conditions on A_2 in the theory of belief functions, but not in the other theories.

The picture becomes clearer, perhaps, then we notice that in the Bayesian theory the maximum rate for betting on A_1 conditional on A_2 happens to be equal to the conditional probability $P(A_1|A_2)$. This is because $E_P\,(X) \geq 0$ if and only if

$$P\,(A_1 \cap A_2)\,(1-p) + P\left(\overline{A_1} \cap A_2\right)(-p) \geq 0$$

or

$$p \leq \frac{P\,(A_1 \cap A_2)}{P\,(A_2)} = P(A_1|A_2).$$

Bear in mind that though this maximum rate might be called a "conditional betting rate," it is the bet that is conditional; the rate itself is "unconditional" in the sense that it is our rate prior to obtaining new evidence and "conditioning" on A_2. But when we obtain this new evidence the conditional bet becomes, for practical purposes, unconditional—for we know its condition is satisfied. Thus our new maximum rate for the conditional bet will be the same as our new maximum rate for an unconditional bet on A_1—i.e., our new degree of belief in A_1. But this new degree of belief is $P(A_1|A_2)$. This is how it happens that our maximum rate for this particular conditional bet is unchanged.

The same thing happens in our theory of lower probabilities: the maximum rate for betting on A_1 conditional on A_2 happens to be equal to $P_*(A_1|A_2)$, and hence remains unchanged when we condition on A_2. But in the theory of belief functions this does not happen: our "prior" maximum rate for betting on A_1 conditional on A_2 is usually not equal to $Bel(A_1|A_2)$, our "posterior" maximum rate for betting on A_1.

3.4 The Dynamic Assumption of the Betting Theories

In this essay I have insisted on understanding both the Bayesian theory and the theory of lower probabilities as constructive theories. I have assumed that the degrees of belief given by both theories are the result of comparing one's evidence to knowledge about chances governing the truth. And I have used this assumption to derive the theories' methods for pricing gambles and their rules of conditioning.

In the literature that treats probability theory as a theory about the gambling behavior of "idealized rational agents," on the other hand, there is no possibility of appealing to chance models to derive rules of conditioning. And thus these rules for changing degrees of belief or betting rates become, to use Ian Hacking's eloquent phrase, *dynamic assumptions*.[3] And one faces the problem of making these assumptions plausible.

Here is how de Finetti tries to make the Bayesian rule of conditioning plausible. He begins by *defining* a Bayesian's "conditional probability of A given B," denoted $P(A|B)$, as his rate for betting "on A conditional on B"—his rate for betting, that is to say, on A with the understanding that the bet will be called off unless B is true. (See de Finetti [6], p. 109, [7], p. 82, [8], p. 135.) He then proceeds to interpret $P(A|B)$ as the probability of A conditional on B in the usual sense—i.e., as the Bayesian's degree of belief or betting rate for A after he has obtained new evidence establishing B. (See de Finetti [6], p. 119, [7], p. 210, [8], p. 141).

What are we to make of this procedure? It obviously takes for granted that *one's betting rate for a conditional bet should be unchanged when new evidence is obtained whose direct effect is to establish the truth of the bet's condition.* Let us call this *de Finetti's principle.* I have been unable to find a critical discussion of this principle in de Finetti's writing. He seems to consider the principle too self-evident to require such a discussion.

As one who finds the theory of belief functions, which does not obey de Finetti's principle, self-consistent and appealing; I find the idea that de Finetti's principle is self-evident baffling. I see the correctness of the principle when betting rates are based on knowledge of chances governing the truth. I am willing to accept the principle as part of a theory that compares our evidence to knowledge of chances. But I do not see that it is inherent to the idea of betting *per se.* It is clear enough that a bettor should change his betting rates when he learns that B is true, and that his new rate for an unconditional bet on A should be the same as his new rate for a bet on A conditional on B.

[3] See Hacking [12]. In this paper Hacking complains about the lack of any justification for the rule of conditioning in the Bayesian literature. The literature on lower probabilities is equally lacking. Since Hacking wrote, Teller [29] has given a Dutch-book argument for the Bayesian rule of conditioning, but this argument depends on the Bayesian rule of additivity and also on the assumption that we know before obtaining the new evidence that the subset established by it will be an element of a certain partition. See also Freedman and Purves [10].

Moreover, these new rates should be the same as the new rate for a bet on A conditional on any B' such that $B \subset B' \subset \Theta$. All these bets are equivalent for someone who knows that the truth is in B. But why should the new rates for all these bets be the same as the old rate for the bet conditional on B? Why should this particular rate remain unchanged while the others change?

De Finetti's principle can similarly serve as the dynamic assumption of a betting theory of lower probabilities. Smith [28] seems to use it in this way, for he gives the name "lower conditional probability" to a bettor's maximum rate for a bet on A which is to be called off unless B is true (p. 6) and then takes it for granted that this should become his betting rate for an unconditional bet on A when he obtains new evidence establishing B. Williams [30] similarly identifies lower conditional probabilities as betting rates for conditional bets but does not discuss changes in betting rates resulting from new evidence.

3.5 Williams' Argument on Conditional Bets

On p. 381 of his review, Williams discusses the pricing of conditional bets in the theory of belief functions. He casts his argument in terms of a numerical example, but we can easily recast it in general terms. It begins, essentially, with the following fact: *offers to bet on A_2 at rate p and on A_1 conditional on A_2 at rate q entail an offer to bet on $A_1 \cap A_2$ at rate pq.* (Proof: If the bet on A_2 has total stakes q, then it is the gamble

$$X_1(\theta) = \begin{cases} (1-p)\,q & \text{if } \theta \in A_2 \\ (-p)\,q & \text{if } \theta \notin A_2. \end{cases}$$

If the conditional bet has unit stakes, then it is the gamble

$$X_2(\theta) = \begin{cases} 1-q & \text{if } \theta \in A_1 \cap A_2 \\ -q & \text{if } \theta \in \overline{A_1} \cap A_2 \\ 0 & \text{if } \theta \notin A_2. \end{cases} \tag{16}$$

Taking both these gambles means taking the gamble

$$X_1(\theta) + X_2(\theta) = \begin{cases} 1-pq & \text{if } \theta \in A_1 \cap A_2 \\ -pq & \text{if } \theta \in A_1 \cap A_2, \end{cases}$$

which is merely a bet on $A_1 \cap A_2$ at the rate pq.)

Suppose we price gambles using a belief function Bel, so that $Bel(A_2)$ and $Bel(A_1 \cap A_2)$ are the greatest rates at which we will bet on A_2 and $A_1 \cap A_2$, respectively. If q is a rate at which we bet on A_1 conditional on A_2, then our willingness to bet on A_2 at the rate $Bel(A_2)$ implies, by the italicized sentence, a willingness to bet on $A_1 \cap A_2$ at the rate $Bel(A_2)q$. So the assertion that $Bel(A_1 \cap A_2)$ is the greatest rate at which we will bet on $A_1 \cap A_2$ will be valid only if

$$Bel(A_1 \cap A_2) \geq Bel(A_2)q. \tag{17}$$

Williams asks, in effect, whether the pricing of conditional gambles in the theory of belief functions guarantees that (17) will be true.

In fact, the theory of belief function does guarantee (17). For it sanctions the conditional bet (16) only if (16) has a non-negative lower expectation—i.e., only if

$$(1-q)\sum\{m(A)\,|A\subset A_1\cap A_2\}\geq q\sum\left\{m(A)\,|A\cap\overline{A_1}\cap A_2\neq\emptyset\right\},$$

which implies

$$(1-q)\sum\{m(A)\,|A\subset A_1\cap A_2\}\geq q\sum\{m(A)\,|A\subset A_2;A\not\subset A_1\cap A_2\},$$

or

$$(1-q)\,Bel(A_1\cap A_2)\geq q\,(Bel\,(A_2)-Bel\,(A_1\cap A_2)),$$

which is equivalent to (17).

There is, of course, a more general issue here. The question is whether interpreting $Bel(A)$, for each $A\subset\Theta$, as the greatest rate at which a bet on A is sanctioned is consistent with sanctioning every gamble with non-negative lower expectation. We easily see that a bet on A at rate p has non-negative lower expectation if and only if $p\leq Bel(A)$. But perhaps it is possible, in some cases, to build up a bet on A at a rate higher than $Bel(A)$ by compounding other sanctioned gambles. In fact, it is not possible. One way to verify this is to check directly that the lower expectation $\widehat{Bel}$ obeys $\widehat{Bel}(X_1+X_2)\geq\widehat{Bel}(X_1)+\widehat{Bel}(X_2)$ and $\widehat{Bel}(aX)=a\widehat{Bel}(X)$ for $a\geq 0$. Another way is to apply the general theory developed by Smith and Williams.

The relation (17) *would* be a problem for belief functions if we interpreted the conditional degree of belief $Bel(A_1|A_2)$ as a sanctioned rate for a bet on A_1 conditional on A_2. For then (17) would imply

$$Bel\,(A_1\cap A_2)\geq Bel\,(A_2)\,Bel(A_1|A_2), \tag{18}$$

and, as Williams shows using a numerical example, this relation can easily be violated by belief functions.

Unfortunately, Williams finds the identification of conditional degrees of belief with betting rates for conditional bets so compelling that he takes the failure of (18) to be a shortcoming of the theory of belief functions. He concedes (p. 381) that one might say that "$Bel(A|B)$ as defined by Shafer should be interpreted as the largest rate at which the subject would be prepared to bet on A if B were discovered to be true (whatever this means), whereas the interpretation given is in terms of the subject's prior readiness to accept conditional bets." But he evidently finds this too bizarre to take seriously, for he concludes (p. 387) that the theory's rule of conditioning "excludes the possibility of interpreting degrees of belief in terms of acceptable betting rates."

I have, I hope, adequately explained why the theory of belief functions does not identify conditional degrees of belief with betting rates for conditional bets. And I think we may conclude from the example provided by the theory of belief functions that such an identification is not inherent in the idea of betting itself. So if we apply to Williams' ideas on lower probabilities the same standards of justification that he has applied to the rules for belief functions, we must ask him to justify this identification. Perhaps the best justification is the one I have developed in this essay: the identification holds if our model for evidence is partial knowledge of chances governing the truth.

4 Minimal Extension

A lower probability function defined only on a restricted class of subsets of a frame Θ can always be extended in a minimal way to a lower probability function defined on all subsets of Θ. Belief functions can be extended in a similar way provided that the restricted class is closed under intersections but not, in general, otherwise. And this, Williams argues, makes it "difficult, in certain cases, to find a belief function which might adequately express a subject's opinions."

Here, as elsewhere in his review, it is not clear whether Williams is taking a constructive point of view. His talk about "expressing a subject's opinion" could be construed to mean that we are concerned not so much with constructive probability judgment as with the task of eliciting opinions already determined. I shall, however, respond to Williams' criticism within the constructive framework of this essay.

4.1 Minimal Extension for Belief Functions

Consider a detective who is trying to find out who stabbed a man to death. Many questions will engage his interest: the circumstances of the killing, the circumstances of the victim, etc. But few of his sources of evidence will bear directly on more than a few of these questions. A medical specialist might, for example, give evidence that bears directly only on the time of death and the nature of the struggle. Evidence that bears on the time of death may, of course, ultimately point to the killer, but only indirectly, through its interaction with other evidence.

It may be the case, as I suggested in Sect. 3 above, that the idea of "direct effect of evidence" cannot be reduced to simpler ideas and so must be learned by example. Be this as it may, it is a clear and commonplace idea, and one that is fundamental in the theory of belief functions. The use of the idea is quite simple. When we judge that given evidence bears directly only on certain questions, we formulate a frame that deals only with these questions and then construct a belief function *Bel* over this frame to represent the evidence. We then think of this frame as a coarsening of a finer frame Θ that takes into account the other questions with which we are concerned. (See Chap. 6 of *A Mathematical Theory of Evidence.*) Or, to use a more familiar vocabulary, we think of the subsets of the first frame as forming a subalgebra $\mathcal{B}$ of the algebra of all subsets of the finer frame Θ. And we adopt the belief function $\overline{Bel}$ over Θ, where

$$\overline{Bel}\,(A) = \sup\left\{Bel\,(B)\,|B \in \mathcal{B}, B \subset A\right\} \tag{19}$$

for each $A \subset \Theta$. The belief function $\overline{Bel}$ is called the *minimal* (or *vacuous*, or *canonical*) *extension* of *Bel*; it gives each element of $\mathcal{B}$ the same degree of belief as *Bel* does, and it gives the other subsets the smallest degrees of belief consistent with these. (See Sect. 7.3 of *A Mathematical Theory of Evidence.*)

The subalgebra $\mathcal{B}$ may be more or less detailed. The detective and medical specialist, for example, may judge that the direct significance of certain medical evidence is exhausted by saying that it is highly probable that death took place between 5 and 10 hours ago. Or they may think this evidence also provides some support for a more exact time of death. Or they may think it provides both this and also some indication of the nature of the struggle. In the first case they might set $\mathcal{B} = \{\emptyset, B_0, \overline{B_0}, \Theta\}$, where B_0 corresponds to the death taking place between 5 and 10 hours ago, set $Bel(\emptyset) = Bel(\overline{B_0}) = 0$, $Bel(B_0) = .95$, and $Bel(\Theta) = 1$, and thus obtain for $\overline{Bel}$ a simple support function focused on B_0. But in the other cases $\mathcal{B}$ will be more detailed and $\overline{Bel}$ will be more complicated.

The idea of minimal extension can be generalized to the case where the initial belief function *Bel* is defined not on a subalgebra but merely on a collection $\mathcal{E}$ of subsets of Θ that is closed under intersections. (A function on such a collection is called a belief function if there is at least one way to extend it to a belief function over Θ.) As it turns out, there always exists in this general case a belief function $\overline{Bel}$ over Θ that extends such a belief function *Bel* (i.e., agrees with it on $\mathcal{E}$) and gives all subsets of Θ the smallest degrees of belief given to them by any belief function that extends *Bel*. To put it another way, the function $\overline{Bel}$ defined by

$$Bel(A) = \inf\{Bel'(A)|Bel' \text{ is an extension of } Bel\} \tag{20}$$

for all $A \subset \Theta$ is a belief function. If $\mathcal{E}$ is not an algebra, then the formula (19) for $\overline{Bel}$ may not be valid, but a more complicated formula can be given. (See Shafer [26].)

The notion of minimal extension breaks down for belief functions if the collection $\mathcal{E}$ on which *Bel* is initially defined is not even closed under intersections. For in this case there may not be a single extension of *Bel* which assigns smallest degree of belief to all subsets of Θ. To put it another way, the function $\overline{Bel}$ given by (20) may fail to be a belief function. The practical implication of this is that probability judgments based on a single item of evidence should include direct judgments about $A \cap B$ whenever they include direct judgments about A and about B. If, for example, our medical specialist judges given evidence to indicate both that the death occurred within the last ten hours and that the victim resisted, then his numerical judgments should include not only judgments about the support for each of these propositions but also a judgment about the support for their conjunction. If the specialist judges that the support for the two propositions comes from intuitively independent items of evidence or aspects of the evidence, then he can use Dempster's rule to determine the degree of support for the conjunction, but otherwise he must make a direct judgment.

In practice, the theory of belief functions applies minimal extension mainly to the case where initial judgments determine a belief function on a subalgebra. For the intuitive judgment that given evidence bears directly only on certain questions seems to translate naturally into the idea that it bears directly only on a subalgebra. And most of the theory's relevant tools (assessment relative to a single dichotomy, consonant assessment, discounting of frequencies) are readily understood as tools for constructing belief functions on subalgebras. The generalization to the case of a collection of subsets closed only under intersection seems to be of interest only as a technical tool in a theoretical context. (See Shafer [23].)

4.2 Minimal Extension for Lower Probabilities

As Williams points out, minimal extension can be applied to lower probabilities defined on an arbitrary collection $\mathcal{E}$. Suppose, indeed, that we make direct judgments that give us lower probabilities $P_*(A)$ for A in such a collection $\mathcal{E}$ and then make the judgment that those lower probabilities exhaust the impact of the evidence. If we have arranged the judgments $P_*(A)$ for $A \subset \mathcal{E}$ so that there is at least one extension to a lower probability function over Θ (i.e., so that there is at least one lower probability function P'_* defined for all subsets of Θ such that $P'_*(A) = P_*(A)$ for all $A \subset \mathcal{E}$; this may be a difficult condition to check), then there exists a *minimal extension*—a lower probability function $\overline{P'_*}$ defined for all $A \subset \mathcal{E}$ and awarding all subsets the least values awarded by any P'_* that extends P_*. In other words,

$$\overline{P'_*}(A) = \inf\left\{P'_*(A) \,|P'_* \text{ is an extension of } P_*\right\}$$

defines a lower probability function. This is obviously the same concept of minimal extension as the one used by the theory of belief functions. The only difference is that it works for all $\mathcal{E}$, not just for $\mathcal{E}$ that are closed under intersections.

The matter can be put most concisely by saying that there always exists a minimum element in the class of those lower probability functions assigning given values to given subsets. Notice, however, that there are many other properties such that there does not exist a minimum element in the class of lower probability functions having the property. If, for example, $\Theta = \{-1, 0, 1\}$, then there is no minimum element in the class of lower probability functions having lower expectation zero. Thus even lower probability functions are limited in this respect. One cannot specify arbitrary properties for a lower probability function, decide that these specifications exhaust the impact of the evidence, and then adopt the minimum lower probability function having the properties.

Williams' notion of minimal extension finds a place in the general constructive theory of lower probabilities that I developed in Sect. 1 above, but only as a rather special case. For in that theory we make judgments that impose a rather wide variety of constraints on a supposed chance distribution P before judging that we have exhausted the impact of the evidence and proceeding to derive a lower probability function P_* from the class $\mathcal{P}$ of distributions satisfying the constraints. And only if the constraints are all of the particular form "$P(A) > c$" can we think of each judgment as establishing a particular value $P_*(A)$.

4.3 Williams' Example

The tool of minimal extension is more widely available for lower probabilities than for belief functions. But what significance does this have? It seems to me that it has little immediate significance, and that its ultimate significance can only emerge from comparing the two theories as a whole in the context of actual examples. Discussing the question in isolation is rather like comparing two tool boxes on the basis of the weight of their hammers without regard for the different roles the two hammers play.

Williams does give an example to support his belief that minimal extension for arbitrary $\mathcal{E}$ is needed. He writes as follows:

> ... suppose there is evidence relating to the unknown outcomes of two tosses of a coin giving rise, for each toss, to a belief function
>
> $$Bel(\{H\}) = \frac{1}{2}, \qquad Bel(\{T\}) = 0.$$
>
> The upper and lower probabilities of heads, on either toss are therefore $\frac{1}{2}$ and 1, respectively. Now consider which belief function might be chosen to express the impact of the evidence on the set of possible joint outcomes $\{HH, HT, TH, TT\}$. We must have
>
> $$Bel(\{HH, HT\}) = Bel(\{HH, TH\}) = \frac{1}{2}, \tag{5}$$
>
> $$Bel(\{TH, TT\}) = Bel(\{HT, TT\}) = 0 \tag{6}$$
>
> since the arguments in (5) are respectively the events 'heads on the first toss' and 'heads on the second toss', whilst the arguments in (6) refer correspondingly to tails. Furthermore, one can imagine situations in which it would seem reasonable to say that no more support accrues to the remaining sets of possibilities than is required by (5) and (6). That is to say, we should look for a minimum element in the set of belief functions satisfying these conditions. ...

But, as Mr. Williams points out, there is no minimum in the class of belief functions over the frame $\Theta = \{HH, HT, TH, TT\}$ satisfying (5) and (6). (Here we have, in effect, $\mathcal{E} = \{\emptyset, \{HH, HT\}, \{HH, TH\}, \{TH, TT\}, \{HT, TT\}, \emptyset\}$, and this is not closed under intersection. We have made judgments about the degree of support for $\{HH, HT\}$ and about the degree of support for $\{HH, TH\}$, but not about the degree of support for $\{HH\} = \{HH, HT\} \cap \{HH, TH\}$.)

What are we to make of this example? Does it demonstrate that the wider availability of minimal extension can enable a theory of lower probabilities to do better than the theory of belief functions? No. The deficiency of the example in this respect is its abstract starting point. To compare theories fairly we need to compare complete analyses—analyses beginning with a full intuitive account of one's evidence and then building up the formal judgments step by step. Williams begins with the assumption that his evidence is best represented by the judgments (5) and (6) and the further judgment that $\mathcal{E}$ exhausts the impact of the evidence, and this assumption begs the real questions. If we do begin with an intuitive account of the evidence, then it may emerge that these judgments provide one sensible analysis, but it is unlikely that they will provide the only one. It is quite possible that there will be sensible analyses using belief functions that take quite different tacks. We might even choose to make a direct judgment about $\{HH\}$.

The only gesture Williams makes towards giving an intuitive basis to his example is the following:

> ...Suppose the evidence to consist of the outcome of a single toss of the coin. It is hard to see how this could provide evidence for or against any particular correlation...

And this, to my mind, says nothing about the real evidence. It seems to indicate that we have dreamed up a statistical model as one approach to analyzing the evidence. Apparently we are regarding two possible events (here called coin tosses) as repeatable experiments, with some joint chance distribution governing the pair of outcomes (X_1, X_2), say. And apparently our statistical model consists of those chance laws with identical marginals for X_1 and X_2. We are to observe another toss independent of (X_1, X_2) but governed by the same marginal and to infer what we can about the joint distribution and hence about how (X_1, X_2) will turn out. This is a parametric statistical problem. But where does it come from? What is the evidence for the model? A sensible analysis using belief functions would require answers to these questions.

5 The Independence of Evidence

Both the Bayesian theory and the theory of belief functions have a concept of independence for evidence. Both recognize different items of evidence as intuitively independent and model this intuitive independence in terms of stochastic independence. But since the two theories use the picture of chance in different ways, their concepts of independence are different.

In the theory of belief functions we liken evidence to a message whose meaning is random, or to a randomly valid argument—one whose validity depends on chance. We call different items of evidence intuitively independent when they can be likened to stochastically independent randomly valid arguments.

In the Bayesian theory, on the other hand, we liken our evidence to knowledge that the truth is generated by certain chances. Thus we do not, in general, think of the evidence itself as random. If, however, we single out a few items of our evidence, imagine that we have not yet obtained these items of evidence, and include the question of whether we will obtain them among the questions about which we are making probability judgments, then whether or not these items will occur becomes part of the truth which we are modeling as random, and so it becomes possible to think of these items of evidence as stochastically independent.

The two theories' concepts of independence have much in common. In many cases, the two theories can agree on calling certain items of evidence independent. And in both theories independence is relative to a given frame of discernment. In the theory of belief functions, this is expressed by saying that different arguments should be treated as independent only relative to a frame that discerns the interactions of their conclusions, while in the Bayesian theory it is expressed by saying that different items of evidence may be independent only conditionally given certain hypotheses.

We should not be misled, however, into thinking that the two concepts of independence are practically identical—that the two theories will always agree on whether given items of evidence are independent.[4] The fact is that they will often disagree. As we shall see in this section, the theory of belief functions may allow us to discern independent items of evidence in situations where the Bayesian theory suggests dependent items of evidence or even suggests that we need not distinguish separate items of evidence at all.

Confusion between the two theories' concepts of independence can be held responsible for the suggestion, made by Williams in his review, that the theory of belief functions cannot do as well as the Bayesian theory in taking dependencies in evidence into account. One goal of this section is to understand the thinking behind this claim and to explain why it is wrong.

5.1 Independence in the Theory of Belief Functions

The concurrence of many independent arguments can justify a high degree of belief. And it is natural to account for this by reasoning about chances. There may be a substantial chance, we tell ourselves, for any single one of the arguments to be invalid, but there is a much smaller chance that they should

[4] In Shafer [24] I suggested, wrongly, that there was such a practical identity.

all be invalid. If p_i is the chance that the ith argument is invalid, and the arguments are independent, then the chance that they are all invalid is the product of the p_i.[5]

This is a sensible account, but it must be rightly understood. When we say that the chance of an argument's validity is p_i we do not mean that the argument is literally a repeatable experiment, sometimes valid, sometimes not, and that we know the chance p_i in the way we might know the chance of heads when tossing a well-studied coin. We mean rather that we judge the force of the argument to be comparable to the force of such a randomly valid argument. And when we say that the arguments are independent, we do not mean that their validities are literally stochastically independent random events. We mean rather that we judge the arguments to be independent in an intuitive sense that is well-represented by stochastic independence[6]—i.e., that we judge the uncertainties in the arguments to be sufficiently unrelated that the combination of the arguments should have the force of the concurrence of two stochastically independent randomly valid arguments.

Dempster's rule of combination is merely an extension of this simple idea of combining the force of independent arguments by multiplication. As I explained in Sect. 1 above, the rule pools two bodies of evidence by treating the two randomly coded messages representing them as stochastically independent. When one uses the rule, one is making a judgment that the two bodies of evidence are sufficiently unrelated that pooling them is like pooling stochastically independent randomly coded messages.

Consider a simple example from *A Mathematical Theory of Evidence.* A detective investigating a burglary turns up one argument indicating that the burglar was lefthanded and another argument indicating that the burglary was an inside job. Suppose these two arguments are intuitively independent, in the sense that they involve different uncertainties and that the evaluation of each depends on a different small world of experience. Say the argument for the burglar being left-handed is based on smudges on the door of the safe, and thus depends for its evaluation on the detective's experience and insight into the question of how safes are forced open, whereas the argument for the burglary being an inside job is based on the detective's understanding of the possibilities

[5] This rule was discussed by James Bernoulli in his *Ars Conjectandi*, published posthumously in 1713. Bernoulli also gave several other rules for combining probabilities based on independent arguments. Since most of these rules are special cases of Dempster's rule of combination, Bernoulli can be regarded as the founder of the theory of belief functions. Though Bernoulli's account of the combination of arguments was popular during the 17th century, it was eventually displaced by the Bayesian account developed by Condorcet and Laplace. See pp. 345–349 of Shafer [22] and pp. 465–469 of Pearson [19].

[6] We should bear in mind that chance is never an objective fact but is always an abstract picture that we impose on nature to aid our understanding. Stochastic independence, in particular, is an abstract concept that we use to model situations where we have first perceived a causal or intuitive independence.

for entering the building. It might, in such a case, be quite reasonable for the detective to treat the two arguments as if they were stochastically independent randomly coded messages. It is not that his train of thought in forming each argument is an independent chance process and that he knows the chance that each process has to produce a valid result; it is just that he can evaluate his confidence in each argument by comparing it with the scale of randomly coded messages and he can judge that there is no important common element in the uncertainties in the two arguments.

We might, of course, challenge the detective's judgment. We might discover a soft spot which is common to both arguments and which the detective failed to notice—perhaps he is too readily ruling out some hypothesis that could explain both the smudge on the safe door and an unnoticed entry into the building. But the possibility of challenge is not peculiar to judgments of independence. Every probability judgment is open to challenge.

One point that emerges from this example is that the idea of independence applies not to isolated facts or propositions but to whole small worlds of experience and human interaction with experience. When we explain what arguments we are combining, it is natural to identify each by a proposition: Argument 1 = "there were smudges on the door of the safe;" Argument 2 = "the building was being watched." But these propositions are only tags. We are really combining whole "bodies of evidence"—whole bodies of concrete experience and interactive human evaluation of that experience.

It is inherent in the idea of analyzing our evidence into independent arguments that the force of each argument is evaluated in abstraction from the other arguments. Each argument is evaluated, that is to say, in abstraction from the other evidence bearing on its conclusions. But when we combine arguments we must take the interaction of the conclusions into account—we must take into account whether the arguments concur, what they support when they are combined, and whether they conflict, either in pairs or in more complicated interactions. Since conflict modifies our evaluation of the weight of the arguments (through the renormalizing constant K in (3)) even when the conflict is not on a point of substantive interest to us, we must take all conflict in conclusions into account. So we should apply Dempster's rule to belief functions representing different arguments only if the frame Θ over which these belief functions are defined is fine enough to take all conflict and other relevant interaction into account.

So we have two requirements for the use of Dempster's rule of combination: (i) The bodies of evidence must be entirely distinct. The uncertainties in the arguments being combined, that is to say, must be independent when the arguments are viewed abstractly—i.e., before the interactions of their conclusions are taken into account. (ii) The frame Θ must be fine enough to discern all relevant interaction of the conclusions.

5.2 Is There an Objective Criterion for Independence?

Peter Williams is not satisfied with the preceding explanation of the conditions for the legitimate use of Dempster's rule of combination. It is not clear, he tells us,

> that this formulation is sufficient to distinguish unambiguously between permissible and impermissible applications of the rule. To begin with, the identity criteria for bodies of evidence are unclear if these cannot be expressed as propositions. Indeed, even if they can be, do two propositions which are not logically equivalent, but are nonetheless equivalent by virtue of natural laws, express 'entirely distinct bodies of evidence'? Or again, suppose that two bodies of evidence are distinct, taken as wholes, but nonetheless partly overlap. ... [H]ow is one to extract the common part, given that bodies of evidence are not necessarily expressible as propositions?

In this passage Williams seems to be demanding some objective criterion for deciding when two bodies of evidence are independent and, more generally, some mechanical way of analyzing evidence into distinct or independent items. Do these demands make sense?

It seems to me that the idea of an objective criterion for the independence of evidence—the idea of a criterion exterior to the judgment—is a chimera. The judgment that two bodies of evidence are independent is a probability judgment, and the appropriateness of probability judgments can never be justified on the basis of criteria that do not themselves demand the application of judgment.

The analysis of evidence into distinct and independent arguments is, moreover, always a constructive act of judgment. Williams is quite right to suggest that there is no unambiguous formula telling us how to do it. It is usually the most creative and the most difficult part of our effort to understand a problem.

There is, in short, no royal road. The analysis of evidence is difficult, and foolish mistakes are always possible. As James Bernoulli put it, "many things will happen which can cause one to err frequently and shamefully unless one proceeds cautiously in discerning arguments. For sometimes arguments can seem distinct which are in fact one and the same argument. Or, vice versa, those which are distinct can seem identical..." (See p. 337 of Shafer [22]).

As Williams' comments indicate, one concomitant of the desire for a mechanical approach to the analysis of evidence is a desire to express evidence as sentences or as propositions. If we could translate all our evidence into statements of fact, then we could, it would seem, give rules for mechanically analyzing this evidence using symbolic logic together with background knowledge encoded as prior probabilities. But we cannot usually translate our evidence into statements of fact.

We can always describe our evidence, the reader may protest. This is true. But the description will usually have to include not only statements of fact but also statements of probability judgment. How might the detective describe the evidence that convinces him that a person cannot enter the building without being seen by the watchman? The evidence consists, in a very real sense, of mental experiments that the detective carried out on the scene. He tried everything he could think of, and nothing seemed plausible. Perhaps he can describe some of this mental experimentation—at least if you allow him to draw pictures. But how can he reduce his conviction that a certain trick will not work to statements of fact? How can he formulate statements of fact to express his degree of conviction that he has tried everything? In the end he will simply have to supplement his statements of fact with probability judgments.[7]

5.3 Independence in the Bayesian Theory

The Bayesian theory can combine intuitively independent items of evidence, but it does not do so, as the theory of belief functions does, by regarding each as an independent argument. Instead it asks us to think of the occurrence of each item of evidence as a random event and to assess the probabilities of these events under various hypotheses. And it asks us to model the intuitive independence of the different items of evidence by stochastic independence, conditional on the various hypotheses, of the events that these items of evidence will occur.

The idea is that we should single out certain items of evidence and then imagine ourselves assessing, before these items of evidence occur, both the probabilities of the hypotheses on which we want to bring these items of evidence to bear and also the probability that these items of evidence will occur, given each of the hypotheses. Suppose, for example, that we are considering an exhaustive list of mutually exclusive hypotheses $H_1, \ldots, H_k$ and we single out two items of evidence E_1 and E_2. Then our task is to use "old evidence" (evidence other than the occurrence of E_1 and E_2) to construct Bayesian probabilities $P(H_i)$ and $P\left(E_1 \text{ and } E_2 | H_i\right)$. And if we judge E_1 and E_2 to be like independent random events given H_i—if, that is to say, our old evidence together with knowledge of H_i can be compared to knowledge that E_1 and E_2 are stochastically independent—then we can construct $P\left(E_1 \text{ and } E_2 | H_i\right)$ by making separate probability judgments $P\left(E_1 | H_i\right)$ and $P\left(E_2 | H_i\right)$ and then setting

$$P\left(E_1 \text{ and } E_2 | H_i\right) = P\left(E_1 | H_i\right) P\left(E_2 | H_i\right). \tag{21}$$

[7] In another passage, Williams coments on my insistence on the "hazy and non-propositional nature of evidence." While standing by the claim that evidence cannot usually be reduced to statements of fact, I would like to withdraw any suggestion (see, for example, p. 120 of *A Mathematical Theory of Evidence*) that evidence is "vague" or "hazy." These epithets are themselves vague, and no useful idea is conveyed when they are applied to evidence. (Cf. Austin [1], pp. 125–127.)

Notice that making all these probability judgments amounts to constructing a Bayesian probability distribution P over a certain frame of discernment Θ. We can suppose, indeed that the H_i and E_i are subsets of this frame, and that the $4k$ subsets $H_i \cap E_1 \cap E_2$, $H_i \cap E_1 \cap \overline{E_2}$, $H_i \cap \overline{E_1} \cap E_2$ and $H_i \cap \overline{E_1} \cap \overline{E_2}$ are disjoint and each contain exactly one element.

The point of constructing this probability distribution P is that we may then take the "new evidence"

$$E_1 \text{ and } E_2 = E_1 \cap E_2$$

into account by conditioning. We can calculate, in particular, the probability

$$P\left(H_i|E_1 \cap E_2\right) = \frac{P\left(H_i\right) P\left(E_1|H_i\right) P\left(E_2|H_i\right)}{\sum\limits_{j=1}^{k} P\left(H_j\right) P\left(E_1|H_j\right) P\left(E_2|H_j\right)}, \tag{22}$$

our probability for H_i based on the total evidence. Formula (22) is known as *Bayes' Theorem.*

Consider, for example, the detective who has evidence that the burglar was lefthanded and evidence that the burglary was an inside job. Give names to these two items of evidence—say E_1 and E_2. The propositions of substantive interest are

$$I = \text{ an insider was involved in the burglary,}$$

and

$$L = \text{ the safe was opened by a left-hander,}$$

and so the hypotheses are $H_1 = I \cap L$, $H_2 = I \cap \overline{L}$, $H_3 = \overline{I} \cap L$, and $H_4 = \overline{I} \cap \overline{L}$. And formula (22) provides a way of constructing probability judgments concerning the H_i using the total evidence.

We must always ask, of course, whether the independent judgment (21) is reasonable. Is it reasonable to think of the evidence E_1 involving the smudge on the safe and the evidence E_2 involving access to the building as random events that are stochastically independent given the H_i?

A more fundamental question is whether it is reasonable or helpful to think of E_1 and E_2 as random events at all. In our belief-function analysis we regarded E_1 and E_2 as arguments involving independent uncertainties. Here the perspective is different. Here we think of E_1 and E_2 not as arguments but as facts. And we transfer all the uncertainties to the hypothetical question of whether these facts would have occurred, given each of the hypotheses. But does this make sense? Can we, for example, intelligibly translate the question of how strongly E_2, the detective's study of access to the building supports I into the question of how likely his study would have been to turn out as it did, given that I is true and given that it is false?

In my opinion, we often cannot intelligibly translate our understanding of the significance of given evidence into answers to the question of how likely that evidence would be to occur. And this, I believe, is the fundamental objection to the version of the Bayesian theory that would have us assess all new evidence using Bayes' theorem. For a detailed discussion, see Shafer [24].

It should be noted, in any case, that the Bayesian theory, like the theory of belief functions, has no objective criterion for independence. In both theories the judgment that two items of evidence should be treated as independent is itself a probability judgment.[8]

5.4 Dependent Evidence?

Bayesian assessment of two items of new evidence does not necessarily require a judgment that the items are conditionally independent. Even if E_1 and E_2 are judged dependent, we can still construct the probability judgment $P\left(E_1 \cap E_2 | H_i\right)$ through the formula

$$P\left(E_1 \cap E_2 | H_i\right) = P\left(E_1 | H_i\right) P\left(E_2 | E_1 \cup H_i\right),$$

where $P\left(E_2 | E_1 \cap H_i\right)$ is a judgment as to how likely E_2 would be to occur based on the old evidence together with knowledge that E_1 has occurred and that H_i is true. And thus we can still use Bayes' theorem, in the form

$$P\left(H_i | E_1 \cap E_2\right) = \frac{P\left(H_i\right) P\left(E_1 | H_i\right) P\left(E_2 | E_1 \cap H_i\right)}{\sum_{j=1}^{k} P\left(H_j\right) P\left(E_1 | H_j\right) P\left(E_2 | E_1 \cap H_j\right)}.$$

So if we do use the Bayesian idea of assessing new evidence in terms of its likelihood to occur, it is not very important whether two items of evidence are independent or not.

The independence of different items of evidence is much more important in the theory of belief functions. For Dempster's rule of combination can be used to combine arguments only if those arguments are judged independent.

There seems to be a paradox here. The Bayesian theory can be understood as a special case of the theory of belief functions, and then Bayesian conditioning is seen as a special case of Dempster's rule of combination. (See p. 20 of *A Mathematical Theory of Evidence.*) But how can Bayesian conditioning be a special case of Dempster's rule if it can be used with dependent evidence and Dempster's rule cannot be?

The paradox is quickly resolved when we remind ourselves that "independence" does not have the same meaning in the two theories. The fact is that two items of evidence that are taken into account by conditioning are necessarily independent in the sense of the theory of belief functions, even though they may be either independent or dependent in the sense of the Bayesian theory.

[8] Seidenfeld [20] seems to think otherwise. The Bayesian theory, he writes, "provides the machinery for deciding whether the data are mutually independent." What machinery?

Let us recall the relation between conditioning and Dempster's rule. We explained conditioning in Sect. 3 above by saying that we condition a belief function Bel on a subset E_1 of its frame Θ in order to take into account new evidence whose direct effect on the frame Θ is to establish for certain that the truth is in E_1. But we can also treat such new evidence as an argument for E_1 whose validity is certain and represent it by a belief function Bel_1 with m-values $m_1(E_1) = 1$ and $m_1(A) = 0$ for all other $A \subset \Theta$. And it is because combining Bel with Bel_1 by Dempster's rule gives the same result as conditioning Bel on E_1 that we say that conditioning is a special case of Dempster's rule.

Now consider a second item of new evidence whose direct effect on Θ is to establish for certain that the truth is in $E_2 \subset \Theta$. This evidence can be represented by a belief function Bel_2 with m-values $m_2(E_2) = 1$ and $m_2(A) = 0$ for all other $A \subset \Theta$. Are the uncertainties in the two new items of evidence independent? Yes, for there are no uncertainties; we are modeling each item of evidence as a randomly valid argument in which the chance of validity is one, and so stochastic independence is automatic and it is legitimate to combine Bel_1 and Bel_2 by Dempster's rule. When we do combine Bel_1 and Bel_2, we obtain a belief function $Bel_1 \oplus Bel_2$ that gives $E_1 \cap E_2$ the m-value one, and combining Bel with $Bel_1 \oplus Bel_2$ by Dempster's rule amounts to conditioning Bel on $E_1 \cap E_2$.

One way of putting the matter is to say that the only decompositions of evidence recognized by the Bayesian theory are decompositions into items of evidence that are, from the point of view of the theory of belief functions, independent. The Bayesian theory permits the combination of evidence only through conditioning, and this means that only one of the bodies of evidence being combined, the "old evidence," can involve uncertainties. The other items of evidence must amount to certainties relative to the frame Θ and hence will be trivially independent of each other and of the old evidence.

When we assign names ("E_1" and "E") to new items of evidence and incorporate them into our frame of discernment, we are, in effect, reducing them from uncertain arguments to facts. We are stripping them of their uncertainties and putting all these uncertainties into what we call the "old evidence," the evidence on which the probability distribution P over the frame Θ must be based.

From the point of view of the theory of belief functions, the concentration of all our uncertainties in the "old evidence" does not, of course, solve the problem of probability judgment. Nor does it necessarily exhaust our interest in the combination of evidence. For we face a new problem of assessment of evidence, the problem of constructing a Bayesian probability distribution P (or, more generally, a belief function Bel) over the frame Θ based on this old evidence. And one way of doing this may be to decompose the old evidence into independent items that can be recombined by Dempster's rule.

It may deepen our understanding of the differences between the Bayesian and belief-function concepts of independence to recognize that Bayesian dependence of E_1 and E_2 may be compatible with belief-function independence not only of the items of evidence provided by the occurrence of E_1 and E_2 but also of the components of the old evidence that bear on E_1 and E_2. It is possible, that is to say, for the combination of belief functions representing intuitively independent components of the old evidence to produce a belief function over Θ which happens to be Bayesian and in terms of which E_1 and E_2 are dependent in the Bayesian sense. In fact, any Bayesian probability distribution P over Θ can, in theory, be produced by such a combination of belief functions.

5.5 Sorting out the Uncertainties

The preceding comments should not be construed as a denial of the practical problems that dependent arguments cause in the theory of belief functions. In many problems it will be easy to analyze the evidence into dependent arguments and more difficult to analyze it into independent arguments.

How do we go about analyzing our evidence into independent arguments? How, to put it another way, do we sort our evidence into arguments that involve distinct uncertainties? Perhaps there is no general answer to this question. But we can gain some insight by thinking about examples.

Suppose we are charged with deciding whether an aerial sprayer has allowed insecticide to drift onto the property of a neighboring landowner. Two arguments are presented by the prosecution: (1) The homeowner testifies that spray billowed across the road from the field being sprayed and settled onto her house and that this drift was significant enough to cause her and her family to suffer from headaches and burning eyes and lips. (2) A government bee inspector testifies that he found dead honey bees lying around the homeowner's beehive, that in his judgment they were killed by insecticide, and that the availability of flowering plants indicates that the bees must have been on the homeowner's property rather than on the field being sprayed when they were exposed.

Both items of evidence seem to directly support the charge of negligence. But one can argue that they involve overlapping uncertainties. The main uncertainties are distinct. The main uncertainty in the first item of evidence is how precise and reliable the homeowner is—how well she remembers and how much she exaggerates. The main uncertainty in the second item of evidence is the reliability of the bee inspector's judgment. But suppose the homeowner, out of pure malice, made up the story about drift and poisoned the bees herself. This possibility constitutes, it would seem, an uncertainty common to both items of evidence. And so if we take the possibility seriously we must count the two items as dependent.

There is, however, an obvious way of getting this common uncertainty out of the two items of evidence: incorporate it into the frame of discernment. We might, for example, consider a frame of discernment Θ consisting of three possibilities:

θ_1: The sprayer was not negligent; the homeowner was inaccurate, and the bee inspector was mistaken.

θ_2: The sprayer was not negligent; the homeowner is lying, and she poisoned the bees herself.

θ_3: The sprayer was negligent.

Relative to this frame of discernment we might describe our two items of evidence a little differently. The first item is our evidence for the reliability and probity of the homeowner (we have listened to her testify, etc.), and it supports θ_3 to some extent, and $\{\theta_1, \theta_3\}$ to a stronger extent. The second item is our evidence from the bee inspector, and it supports $\{\theta_2, \theta_3\}$. Notice that though the two items no longer both directly support negligence (θ_3), they still interact to support it. And they can now be regarded as independent arguments.

This example illustrates a reasonably general idea: often two arguments which seem dependent because of common uncertainties can be understood as independent once the common uncertainties are incorporated into the frame of discernment as explicit possibilities. This idea is the basis for saying that Dempster's rule should be used only when the frame "discerns the relevant interaction" of the different arguments.

The task of sorting our uncertainties into distinct arguments is not always so easy, of course. But I would argue that a theory that directs us to this task is grappling with the real problems in the assessment of evidence.

References

[1] J. L. Austin, 1962. *Sense and Sensibilia.*Oxford.

[2] A. P. Dempster, 1966. New methods for reasoning towards posterior distributions based on sample data. *Annals of Mathematical Statistics*, 37:355–374.

[3] A. P. Dempster, 1968. A generalization of Bayesian inference (with discussion). *Journal of the Royal Statistical Society Series B*, 30:205–247.

[4] Persi Diaconis, 1978. Review of *A Mathematical Theory of Evidence. Journal of the American Statistical Association*, 73:677–678.

[5] Terrence L. Fine, 1977. Review of *A Mathematical Theory of Evidence. Bulletin of the American Mathematical Society*, 83:667–672.

[6] Bruno de Finetti, 1964. Foresight: Its logical laws, its subjective sources. In Kyburg and Smokler, editors, *Studies in Subjective Probability*, pp. 93–158. Wiley.

[7] Bruno de Finetti, 1972. *Probability, Induction, and Statistics.* Wiley.

[8] Bruno de Finetti, 1974. *Theory of Probability.* Vol. 1, Wiley.

[9] Bruno de Finetti, 1975. *Theory of Probability.* Vol. 2, Wiley.

[10] D. A. Freedman and R. A. Purves, 1969. Bayes' methods for bookies. *Annals of Mathematical Statistics*, 40:1177–1186.

[11] I. J. Good, 1962. The measure of a non-measurable set. In E. Nagel, P. Suppes, and A. Tarski, editors, *Logic, Methodology and Philosophy of Science*, pp. 319–329. Stanford University Press, Stanford.

[12] Ian Hacking, 1967. Slightly more realistic personal probability. *Philosophy of Science*, 34:311–325.

[13] Peter Huber, 1973. The use of Choquet capacities in statistics. *Bulletin of the International Statistical Institute*, 45, Book 4:181–188.

[14] Peter Huber, 1976. Kapazitäten statt Wahrscheinlichkeiten? Gedanken zur Grundlegung der Statistik. *Jber. Deutsch. Math. Verein*, 78(2):81–92.

[15] Oscar Kempthorne, 1975. Inference from experiments and randomization. In J. N. Srivastava, editor, *A Survey of Statistical Design and Linear Models.* North-Holland.

[16] Isaac Levi, 1981. Dissonance and consistency according to Shackle and Shafer. In *Proceedings of the Biennial Meeting of Philosophy of Science Association (PSA-78)*, Vol. 2. Philosophy of Science Association, East Lansing, Michigan. Asquith and Hacking, editors.

[17] Dennis V. Lindley, 1977. Review of *A Mathematical Theory of Evidence*. *Bulletin of the London Mathematical Society*, 9:237–238.

[18] James G. March. Bounded rationality, ambiguity, and the engineering of choice. *The Bell Journal of Economics*, 9:586–608, 1978.

[19] Karl Pearson, 1978. *The History of Statistics in the 17th and 18th Centuries.* Griffin.

[20] Teddy Seidenfeld, 1981. Statistical evidence and belief functions. In *Proceedings of the Biennial Meeting of Philosophy of Science Association (PSA-78)*, Vol. 2. Philosophy of Science Association, East Lansing, Michigan. Asquith and Hacking, editors.

[21] Glenn Shafer, 1976. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, NJ.

[22] Glenn Shafer, 1978. Non-additive probabilities in the work of Bernoulli and Lambert. *Archive for History of Exact Sciences*, 19:309–370.

[23] Glenn Shafer, 1978. Dempster's rule of combination. Unpublished paper.

[24] Glenn Shafer, 1981. Two theories of probability. In *Proceedings of the Biennial Meeting of Philosophy of Science Association (PSA-78)*, Vol. 2. Philosophy of Science Association, East Lansing, Michigan. Asquith and Hacking, editors.

[25] Glenn Shafer, 1979. Lindley's paradox. Technical Report No. 125, Department of Statistics, Stanford University. (Later appeared in *Journal of the American Statistical Association*, 77:325–351, 1982.)

[26] Glenn Shafer, 1979. Allocations of probability. *Annals of Probability*, 7: 827–839.

[27] Glenn Shafer and A. M. Breipohl, 1979. Reliability described by belief functions. In *Proceedings of the 1979 Reliability and Maintainability Symposium*, pp. 23–27.

[28] C. A. B. Smith, 1961. Consistency in statistical inference and decision (with discussion). *Journal of the Royal Statistical Society Series B*, 23:1–25.

[29] Paul Teller, 1973. Conditionalization and observation. *Synthese*, 26:218–258.

[30] Peter M. Williams, 1975. Coherence, strict coherence and zero probabilities. *Contributed Papers*, Fifth International Congress of Logic, Methodology and Philosophy of Science, VI, 29,30.

[31] Peter M. Williams, 1976. Indeterminate probabilities. Pp. 229–246 of M. Przelecki, K. Szaniawski, and R. Wójciki, editors, *Formal Methods in the Methodology of Empirical Sciences.* Ossolineum and D. Reidel.

[32] Peter M. Williams, 1978. On a new theory of epistemic probability (Review of *A Mathematical Theory of Evidence*). *The British Journal for the Philosophy of Science*, 29:375–387.

[33] Guus Wolf, 1977. *Obere und Untere Wahrscheinlichkeiten.* PhD thesis, Eidgenössische Technische Hochschule, Zürich. (Diss. ETH 5884).