GLENN SHAFER*

THE CONSTRUCTION OF PROBABILITY ARGUMENTS†

I. INTRODUCTION

Though this book has been assembled by legal scholars interested in probability, I will address my contribution to a broader audience—students of the foundations of probability and statistics. This group includes statisticians, philosophers, and psychologists, as well as law professors. My primary purpose will be to suggest ways that the foundations of probability can benefit from the legal perspective on problems of evidence and argument.

Incorporating the legal perspective on probability does not necessarily mean seeking instruction from judges and law professors. It does mean taking seriously the situations of the various actors in the legal setting: the lawyer who must make a case, the opposing lawyer who must criticize it, and the judge or juror who must evaluate the conflicting arguments. Within the statistical tradition, at least, most thought about the foundations of probability has been inspired by the situations of rather different actors: the gambler in a pure game of chance or the scientist disentangling systematic from random variation. Putting the metaphor of the courtroom on a par with the metaphors of the gambler and the scientist can help us achieve a fuller understanding of probability judgment.

The courtroom metaphor forces us to pay more attention to the constructive nature of probability judgment. The lawyer in the courtroom must make a case. She must construct an argument; she cannot merely present facts. She must establish the relevance of the facts she presents. The judge and jurors must do more than understand the conclusion of the argument. They must decide how well the argument supports the conclusion.

Contrast this with the rhetoric of Bayesian statisticians. They have little to say about relevance, argument, or evaluation. Fundamentally, everything is relevant: we are supposed to condition on all the evidence. A probability analysis is supposed to be all-inclusive, and hence not subject to another level of evaluation. The final probabilities are the last word.

My purpose is not to caricature the practical statistician—Bayesian or non-Bayesian. Thoughtful statisticians are perfectly capable of judicious analyses. But they do so only by standing back from basic theory and rhetoric. I am suggesting that the courtroom metaphor can help integrate a more judicious attitude into this basic theory and rhetoric.

P. Tillers and E. D. Green (eds.), Probability and Inference in the Law of Evidence, 185–204.

Kluwer Academic Publishers.

To put this thought in perspective, recall that the legal tradition had a hand in creating numerical probability. Numerical probability was first invented by legal scholars.[1] These scholars did not develop the calculus of probability we know today because they did not connect their idea to games of chance. But Leibniz and Bernoulli, the seventeenth-century scholars who did connect probability to the mathematics of games of chance, also appreciated the legal tradition. In his famous *Ars Conjectandi*,[2] Bernoulli based probability squarely on the concept of argument. Many of his basic examples concerned legal problems and, as Garber and Zabell have shown,[3] some came from the Ciceronian legal tradition.

Unfortunately, Bernoulli's eighteenth-century successors did not preserve the tie between probability and the legal tradition. They were intent on developing the mathematical vistas that Bernoulli's innovative ideas had opened up, and this led to a renewed emphasis on the gambling metaphor. Legal ideas never subsequently regained a central role in probability theory: I argue that they should.

The next section of this article outlines the constructive philosophy of probability developed in more detail by Shafer and Tversky,[4] and relates this philosophy to the courtroom metaphor. The following sections consider some specific issues that are illuminated by the courtroom metaphor. Section III discusses how the significance of evidence depends on the ground rules for its acquisition. Finally, Section IV discusses the constructive nature of likelihood, using as an example the problem of assessing probabilities of paternity.

There are two disclaimers that need to be included in this introduction. First, I can claim no originality when I urge that students of probability pay more attention to the legal perspective; the thought has been formulated many times. I should call particular attention to the work of David Schum, who is commenting on this article. Professor Schum has done more than anyone else in recent years to bring insights from the law into other contexts in which probability is used.[5] My message here is that these kinds of insights are part of the foundations of probability. The issues involved in justifying Schum's cascaded inference schemes are at least as fundamental as the Dutch-book arguments that philosophers of probability are so fond of debating.

Second, the interaction between law and the world of probability and statistics has many strands, and this article is not concerned with all of them. The increasing use of conventional statistical evidence in legal settings constitutes one important strand.[6] I regard this development positively, but it is not directly related to the theme of this article. Another strand is the use of legal examples by scholars who have advanced alternatives to the usual calculus of probability. Two scholars come to mind in this connection—L. Jonathan Cohen and Per Olof Ekelöf. Cohen has argued that the usual calculus is less satisfactory in the legal setting than a simple numerical scoring of hypotheses.[7] Ekelöf has argued that the legal setting requires calculations similar to those made in the theory of belief functions.[8] I have made similar arguments.[9] but this is not my theme here. Here I am arguing not for any particular theory of probability, but for an incorporation into all these theories of the constructive attitude represented by the courtroom metaphor.

II. Constructive Probability

In *The Foundations of Statistics*,[10] L.J. Savage distinguished three categories of probability: the objectivistic, personalistic, and necessary interpretations. According to the objectivistic interpretation—also called frequentist—a probability is an objective fact about a repeatable event: it is the long-run frequency with which the event happens. According to the personalistic interpretation—also called subjective or Bayesian—a probability is a particular person's opinion; it can be deduced from the person's behavior when she chooses among bets or other acts with uncertain outcomes. According to the necessary interpretation—also called logical—a probability measures the extent to which one proposition, out of logical necessity and apart from human opinion, confirms the truth of another.

The personalistic interpretation, which Savage advocated, has gained wide support in recent decades, and it is now the most vigorous and self-confident of the three interpretations. Most probabilists who offer advice to the legal profession are Bayesians. Their message is that mathematical probability in general, and Bayes's theorem in particular, can help judges and jurors evaluate evidence.[11]

Unfortunately, none of the three interpretations emphasizes the constructive character of probability judgment. The objectivistic and necessary interpretations both treat probability as objective, making it independent of human action. The personalistic interpretation treats probability as a form of opinion, but it neither emphasizes nor requires that probability opinions be deliberately constructed. Indeed, much Bayesian writing gives the impression that these opinions are ready-made in our minds, waiting to be "elicited."[12]

A. *The Constructive Interpretation*

If we want an interpretation of probability that emphasizes its constructive character, we must explain what people are doing when they construct probabilities. Shafer and Tversky suggest that they are matching problems to canonical examples.[13] When we make Bayesian probability judgments, we are matching an actual problem to a scale of canonical examples from physics or games of chance. In these canonical examples, the possible answers to questions of interest have well-defined and known probabilities.

Of course, we do not match all the evidence we have about a problem to a complex canonical example in one fell swoop. Instead we match parts of the problem or parts of the evidence to more modest canonical examples, and then try to fit these partial matches together. In so doing we construct an argument, an argument that draws an analogy between our actual evidence and the knowledge of objective probabilities in a complex physical experiment or game of chance.

There are many choices in the design of a probability argument: we must decide how to break down our evidence and how to put it back together in a probability model: we must choose what to think of as fixed when making numerical probability judgments, and how much detail to include: and we must determine on what, if anything, to condition.

One advantage of the constructive interpretation is that it pulls us down from the fantasy that a numerical probability analysis can take all evidence into account and hence provide the final word on a question, to the reality that any probability analysis must be treated as just another argument. As an argument, it must be evaluated, and the result of the evaluation may be negative for a variety of reasons: our evidence may fail to fit the scale of canonical examples to which we are trying to fit it: our evidence may be inadequate to justify some of the numerical probabilities in our argument (traditionally, we worry about whether our evidence is adequate to justify prior probabilities for statistical hypotheses, but there is nothing special about these probabilities; we need evidence for every probability judgment); or our evidence may be inadequate to justify some of the judgments needed to put individual numerical judgments together.

B. *Subjective Aspects*

The constructive interpretation of probability preserves some of the features of the personalistic interpretation. It preserves the connection with betting behavior because the canonical examples to which we are matching the evidence are examples in which probabilities are reasonable betting rates. Since, however, we are only drawing an analogy to these examples, the interpretation is less dogmatic. There is no pretense that we would really offer to bet all comers.

Subjectivity also enters the constructive interpretation in a more fundamental way. We often must make subjective judgments about which canonical example on the scale best matches the evidence in a problem, and about whether this best match is good enough to constitute a sound argument.

One important way in which the constructive interpretation differs from the personalistic interpretation is in its ability to acknowledge that the parts of a probability argument may differ in quality. The personalistic interpretation is based on the demand that a person should have betting rates, or preferences from which such rates can be derived, for all questions. This demand is made categorically; a person cannot be excused from the demand because of the poor quality of her evidence. The constructive interpretation, on the other hand, directs attention to the evidence. It asks about the adequacy of the evidence for each of the probability judgments we make.

Since one part of a probability argument may be better than the rest, we are led to ask whether conviction should ever be carried by partial arguments. I believe it should. An argument that can be seen as only part of a Bayesian argument, but is based on high quality evidence, may be more cogent than a completed argument that draws on weaker evidence. Section IV of this article argues for the cogency of partial likelihood arguments. Other partial arguments that are often cogent include frequentist arguments (interpreted subjectively) and belief-function arguments.[14]

C. *Objective Aspects*

The constructive interpretation also preserves some aspects of the frequentist or objectivistic interpretation. The probabilities in the canonical examples to which we compare our evidence are objective frequencies as well as subjective betting rates, and the analogy a probability argument draws to one of these canonical examples is therefore stronger when the probabilities we use in the argument are clearly relevant frequencies.

Since clearly relevant frequencies are seldom available, dogmatic statements of the frequentist interpretation make it seem that frequency ideas have very limited applicability. The constructive interpretation permits us to recognize a broader field of applicability for frequency ideas because it allows consideration of cases in which there is only an analogy to the random sampling that the frequentist interpretation requires. The constructive interpretation allows us to make arguments that turn on the implausibility of certain results were a situation the product of random sampling, without pretending that it actually is.

One way to explain the analogy to random sampling is to say that the probability model we have constructed represents a thought experiment. In this thought experiment, we imagine that facts were determined by a random drawing from a certain reference class. The results of such a thought experiment may constitute a persuasive argument, even after we acknowlege that the selection of the reference class was somewhat arbitrary. Section IV will apply this idea of a thought experiment to the problem of assessing the evidence provided by blood tests in paternity cases.

D. *The Courtroom Metaphor*

The constructive interpretation of probability was not developed with the courtroom metaphor in mind, but the metaphor can reinforce and add to the interpretation. One contribution the courtroom metaphor can make is to shift our attention from the idea of fixed evidence to the idea of argument. Most of what is written about the foundations of probability, including the preceding paragraphs, seems to take for granted that when a probability judgment is at issue, the evidence on which it is based is fixed and well-defined. But the courtroom metaphor encourages us to think of evidence as something that develops as an argument is made. Evidence is introduced into court. When a witness testifies, she gives evidence that was not there before, and when a lawyer cross-examines a witness, they together create evidence.

The creative nature of argument is relevant to many issues in the foundations of probability. Consider, for example, the suggestion sometimes made that we should base certain probabilities on our background knowledge. This suggestion assumes the background knowledge is well-defined. Certain things are stored in our heads—or on our library shelves—and others are not. What is stored, however, may not be so well-defined, and what we will find when we look certainly is not well-defined in advance. Rather than speak of the probability defined by our background knowledge, we should speak of the probability or the arguments that we happen to discover when we search our background knowledge.

A related contribution of the courtroom metaphor is to draw attention to how evidence is acquired. We will study this in the next section. Another important contribution the courtroom metaphor can make stems from its separation of the roles of creator and evaluator of argument. This separation forces us to think about both processes. We will return to this point in Section V.

III. The Acquisition of Evidence

Modern textbooks on mathematical probability teach that the conditional probability of an event A given an event B, denoted by Pr[A|B], is given by the following formula:

$$\Pr[A|B] = \Pr[A\&B]/\Pr[B]$$

Figure 1.

They say that this formula is a mathematical definition. Mathematicians have simply decreed that the Pr[A&B]/Pr[A|B] should be called the conditional probability of A given B.

When we are taught the personalistic or subjective Bayesian interpretation of probability, we are further told that when you learn that B is true, you should "condition on B"—i.e., you should change your probability for A from Pr[A] to Pr[A|B]. This is a fundamental personalistic doctrine. One should always take new evidence into account through conditioning. But why? Surely there is something missing here. An arbitrary mathematical definition cannot determine how you should change your beliefs.

A. *Freund's Puzzle*

The suspicion that something is missing is confirmed by a number of puzzles and paradoxes. My favorite is Freund's puzzle of the two aces.[15]

I show you a deck containing only four cards: the ace and deuce of spades, and the ace and deuce of hearts. I shuffle them, deal myself two of the cards, and look at them, taking care not to let you see them. You realize that there are six equally likely possibilities:

Ace of spades and ace of hearts,
Ace of spades and deuce of spades,
Ace of spades and deuce of hearts,
Ace of hearts and deuce of spades,
Ace of hearts and deuce of hearts,
Deuce of spades and deuce of hearts.

If A denotes the event that I have two aces, B_1 denotes the event that I have at least one ace, and B_2 denotes the event that I have the ace of spades, then your initial probabilities are $\Pr[B_1] = 5/6$, $\Pr[B_2] = 1/2$, and $\Pr[A] = \Pr[A\&B_1] = \Pr[A\&B_2] = 1/6$.

Now I smile and say, "I have an ace." You are supposed to react to this new information by conditioning on the event B_1. Thus you change your probability for my having two aces from 1/6 to

$$\Pr[A|B_1] = \Pr[A\&B_1]/\Pr[B_1] = (1/6)/(5/6) = 1/5$$

The information that I have at least one ace has increased your probability that I have two.

Now I smile again and announce, "As a matter of fact, I have the ace of spades." You are supposed to condition again, this time on the event B_2, obtaining the new probability

$$\Pr[A|B_1 \& B_2] = \Pr[A|B_2] = \Pr[A\&B_2]/\Pr[B_2] = (1/6)/(1/2) = 1/3.$$

The more specific information that I have the ace of spades has increased even further your probability that I have two aces. Is this second change reasonable? Should my decision to identify a suit make any difference?

Most people will agree, on reflection, that what is reasonable depends on what ground rules are established in advance regarding what I was supposed to tell you. If it had been agreed that I would first tell you whether I had an ace and then *whether* I had the ace of spades, then the change from 1/5 to 1/3 is reasonable (your probability that I had both aces would go down from 1/5 to zero if I told you that I did not have the ace of spades, so it is reasonable that it should go up from 1/5 to 1/3 when I tell you that I do have the ace of spades). But if it had been agreed that I would first tell you whether I had an ace and then tell you the suit of an ace that I had, if I did have one, then the change from 1/5 to 1/3 is not reasonable. Once I told you I had an ace, I had to tell you either spades or hearts, so it makes no sense for you to raise your probability from 1/5 to 1/3.

B. *A Personalistic Treatment of the Puzzle*

The personalistic interpretation helps us to understand the puzzle. The key is to observe that you should condition not only on the event that I have the ace of spades, but also on the event that I have told you so. You should calculate $\Pr[A|B_1\&B_2\&B_3] = \Pr[A|B_3]$, where B_3 is the event that I tell you I have the ace of spades. This event is part of your evidence, and you should condition on all your evidence.

Suppose, for example, that it is settled that I am going to tell you whether I have the ace of spades. Then B_3 is equivalent to B_2. The event that I tell you I have the ace of spades is equivalent to the event that I have it. So the change from 1/5 to 1/3 is correct.

On the other hand, suppose it is settled that if I have at least one ace, then I am going to tell you a suit. If I have only one, I will tell you its suit. If I have both, I will secretly flip a fair coin to decide which suit to tell you—spades if heads, hearts if tails. In this case, the event that I tell you I have the ace of spades is equivalent to the event that either it is my only ace or else the coin came up heads. When we work out the details, we find that your probability of 1/5 for my having both aces should remain unchanged.[16]

But suppose there are no ground rules; I just deal the cards and volunteer the information. How can you use Bayesian conditioning in this situation? The personalistic answer is that even though you do not know what ground rules I am following, you should have probabilities for the different possible sets of ground rules I might be following. Your probability model should model what I am going to tell you, not just what cards I have.

C. *A Contribution from the Constructive Interpretation*

This personalistic treatment of the puzzle gives us important insights, but it stops short of explaining why it is so valuable to you to know the ground rules.

If we stand back from the personalistic interpretation, it is clear that probability calculations are worth a lot more if ground rules are established in advance. If no such rules were established, then you would not know what was going on, and the assertion that you should have probabilities for the rules I might have been following would not help much. Even the assertion that I was following some unknown set of rules might not have much content. I did what I did and said what I said, and, in the absence of any prior agreement, the question of what I might have said lies more in the realm of imagination than of fact.

The personalistic treatment of the puzzle seems to deny this. Since it insists that we be able to supply a subjective probability for any event, regardless of the quality of our evidence, the personalistic interpretation is always unable to acknowledge that the strength of a probability argument depends on the quality of evidence. In this case, it is unable to acknowledge that the strength of the argument depends on the quality of the evidence for ground rules for the acquisition of evidence on which to condition.

Moreover, the personalistic treatment of the puzzle fails to deal with the question with which we began: Why is it imperative that your beliefs should change in accordance with the formula $Pr[A|B] = Pr[A\&B]/Pr[B]$ if this formula is merely an arbitrary mathematical definition?

The constructive interpretation of probability can do better. If probabilities are interpreted as betting rates, and if there are ground rules that single out B as one of the possibilities for what you might have learned at some time, then the formula $Pr[A|B] = Pr[A\&B]/Pr[B]$ is more than just a definition—it can be derived as a theorem.[17] Moreover, there are such ground rules in games of chance. Card games, for example, have ground rules for when the values of various cards are to be revealed to other players. Hence the existence of ground rules governing our acquisition of evidence in a practical problem is one aspect of the quality of any analogy that we can draw to a game of chance, and hence one aspect of the quality of any Bayesian probability argument that we can construct for the problem.

D. *A Contribution from the Courtroom Metaphor*

Teachers of probability and statistics have not dealt with these issues well.[18] This is due in part to the dominance of traditional metaphors: physical experiments and games of chance. Ground rules for the acquisition of evidence are so integral to both these metaphors that their presence is not even noticed. In a game of chance, the very rules of the game constitute such ground rules. In an experiment, such ground rules are implicit in the set-up of the experiment, in the specification of what is to be measured or observed. These metaphors do not, therefore, prepare students to deal with problems for which ground rules for the acquisition of evidence are problematic.

The courtroom metaphor might help us do better. Courtroom ground rules are not complete, and they are not there automatically; the contending parties must fight for them. The courtroom metaphor therefore forces us to pay attention to the presence or absence of ground rules.

There are many aspects of the courtroom metaphor that a teacher might use to reinforce this point. For example, jurors must often struggle with the significance of the absence of evidence. If a plaintiff fails to present evidence that should have been available had her claims been true, and if she fails to account plausibly for such absence, then a juror may conclude that the claims are probably false.

We can also point out the significance of give and take in the courtroom presentation of evidence. In order to evaluate a witness's testimony, a juror does not limit herself to what the witness has said. She also considers what questions the witness was answering and how well the cross-examining lawyer has tested the witness's credibility. She considers what the witness did not say and how the witness said what she did say. The importance of this give and take is recognized in law, for example, by the prohibition against hearsay testimony. It should also be recognized by the philosophy of probability.

E. *The Completeness of Evidence*

Professor L. Jonathan Cohen, in his contribution to this book, contends that a person is justified in changing the probability for A from Pr[A] to Pr[A|B] (where B is the evidence) only if the evidence B is complete. This contention will seem obtuse to most Bayesians because there are many very different scenarios in which conditioning is equally justified, and the evidence will be much less complete in some of these than in others. Professor Cohen's contention does have a kernel of truth, however, because the legitimacy of conditioning depends on the existence of ground rules for the acquisition of evidence. It is completeness relative to these ground rules that is needed. We cannot justify conditioning on B if B, instead of representing one possible body of evidence permitted by the ground rules, represents just part of such a body of evidence.

IV. The Quality of Likelihoods

Ancillary to the personalistic doctrine that new evidence should always be taken into account by conditioning is the doctrine that this conditioning should be carried out using Bayes's Theorem. It is often asserted that Bayes's Theorem provides a recipe for dealing with new evidence: determine the "likelihoods" associated with the evidence, and multiply prior probabilities by these likelihoods. This means that the likelihoods fully capture the import of the new evidence.

The idea that evidence should be assessed in terms of its likelihood has influenced even those who are hesitant to rely on full Bayesian thought experiments. We are often urged to think about the import of given evidence in terms of likelihood even if we are unable to assess prior probabilities.

The constructive interpretation of probability questions these dogmas. Bayes's Theorem is one tool for constructing probability arguments, but it is not always the best tool. There is no reason to think that probability judgments corresponding to likelihoods will always be the most meaningful or effective probability judgments we can make. In some cases, the analogy that a given likelihood judgment represents may be too poor to make an effective argument. We must seek other arguments in these cases. The arguments we find may involve only partial likelihoods, likelihood judgments based only on part of the evidence initially considered, or they may not involve likelihood ideas at all.

This section will illustrate the limitations of likelihood using as an example the problem of assessing the evidence for paternity provided by blood tests. It turns out that in this problem the most effective arguments usually do have a likelihood form, but that in some cases it is best to rely on partial likelihoods.

A. *Bayes's Theorem*

It will suffice for our example to consider Bayes's Theorem in its simplest form, where prior odds for a hypothesis are multiplied by a likelihood ratio for that hypothesis. A likelihood ratio in favor of A on evidence B may be denoted by L[B|A]. By definition,

$$L[B|A] = Pr[B|A]/Pr[B|not\text{-}A]$$

Figure 2

Bayes's Theorem says that,

$$O[A|B] = L[B|A]O[A]$$

Figure 3

where O[A] denotes prior odds for A, a ratio Pr[A]/Pr[not-A], and O[A|B] denotes posterior odds, Pr[A|B]/Pr[not-A|B].

B. *The Probability of Exclusion*

In paternity disputes, blood tests are often used to check whether a particular man could be the child's father. They are also used, when the possibility of the particular man's being the father is not excluded, to calculate relevant probabilities. These calculations are the subject of a large literature, dating back to Essen-Möller.[19]

The blood tests determine an individual's types for a set of antigens—the individual's phenotype. The law of genetics rules out certain combinations of phenotypes for mother, child, and father. It is possible, therefore, that by testing a mother, child, and alleged father we can exclude the possibility that he is the real father. If this does not happen, then it would seem that the evidence against him has been strengthened. How can this be measured?

One simple approach is the following.[20] Suppose the phenotypes of the mother and child have been determined but that the alleged father has not yet been tested. We calculate a probability p that he will be excluded, given that he is not the real father. If p is nearly one and yet, after testing, the man is not excluded, then p can be thought of as a measure of the doubt cast on his not being the father.

This is equivalent to a likelihood argument because the ratio $p/(1-p)$ is approximately equal to a likelihood ratio. Indeed, if B denotes the event that the alleged father is not excluded, and A denotes the event that he is the father, then Pr[B|not-A], calculated with the phenotypes of the mother and child fixed, is equal to $1-p$. And since Pr[B|A] = 1, the likelihood ratio

L[B|A] = Pr[B|A]/Pr[B|not-A] is equal to 1/(1 − p). If p is near 1, this is approximately equal to p/(1 − p).

How do we calculate p, or equivalently, Pr[B|not-A]? Here we must rely on the fact that large numbers of individuals from different populations have been tested, and tables of the frequencies of different phenotypes have been constructed. The populations are usually racial groups, because the frequencies of different phenotypes do differ for these groups. If the alleged father belongs unequivocally to a racial group for which frequencies of phenotypes have been recorded, then we add up the frequencies for that racial group of all the phenotypes that are inconsistent with his being the father, given the known phenotypes of the mother and child. This total frequency is the probability that the alleged father will be excluded, given that he is not the father, and assuming that his phenotype is drawn at random from the phenotypes of his racial group.

When we talk about the alleged father's phenotype being drawn at random from the phenotype of his racial group, we invoke a thought experiment. The alleged father is not a random man; he is a particular man in a dispute marked by many other particulars. But before we have typed his antigens, assuming he is not the father, our only knowledge concerning his phenotype is provided by frequencies for his racial group. Thus, we can compare our state of knowledge to what it would be if we knew his phenotype were drawn at random from a population having these frequencies. This comparison is the basis for the thought experiment. We imagine how surprised we would be and how much doubt would be cast on the randomness of the drawing if the phenotype drawn were from a subset singled out beforehand as having a very small total frequency. We argue that failure to exclude the alleged father casts similar doubt on the hypothesis that he is not the real father.

In this thought experiment, the phenotype of the mother and child are taken as fixed, and we imagine the phenotype of the alleged father being drawn at random. An obvious alternative is to imagine instead that the phenotypes of all three people—a mother and child and an unrelated man—are drawn at random. Or we might fix the phenotype of the alleged father and imagine those of the mother and child being drawn at random. The difficulty with these latter thought experiments is that they cannot be performed without a further assumption. In order to specify the probability with which a mother-child pair of phenotypes will be drawn, we must specify the racial group of the real father, and in order to draw our analogy to a game of chance, we must think of the real father as being drawn at random from this group. The strength of the thought experiment in which the phenotypes of the mother and child are held fixed lies in its simplicity. It assumes only that the phenotypes of the mother and child have no causal connection with the phenotype of the alleged father if he is not the real father, and it requires only that we think of this alleged father's phenotype as random.

There are situations, of course, in which we would question even a thought experiment in which the phenotypes of the mother and child are fixed and the phenotype of the father is drawn at random. If the alleged father is related to the mother or to the real father, then there is a causal connection between his phenotype and the phenotypes of the mother and child, and the thought experiment loses its cogency.

C. *A Better Argument*

It turns out that there is a cogent thought experiment that accounts for more than the mere fact that the father is not excluded. This thought experiment accounts for the particular non-excludable phenotype the alleged father turns out to have. Since the probability of this particular phenotype will be small whether the man is the real father or not, accounting for it will involve not just the contemplation of one small probability, but also the comparison of two small probabilities. This requires considering their ratio, which is a likelihood ratio.

It will be convenient, in the exposition of this more inclusive probability argument, to use a notation that makes explicit the fact that we are holding the phenotypes of the mother and child fixed. Let M, C, and S denote the phenotypes of the mother, child, and alleged father, respectively, and let A again denote the event that the alleged father is the real father. The argument I suggest compares two probabilities, Pr[S|M&C&A] and Pr[S|M&C&(not-A)].

To calculate Pr[S|M&C&(not-A)], we use the same thought experiment as before. We imagine determining the father's phenotypes by drawing a phenotype at random from the population of phenotypes for his racial group. This requires taking Pr[S|M&C&(not-A)] as the frequency of S in that group. Since we are assuming that the alleged father is unrelated to the mother and is not the child's real father, M and C are irrelevant.

The calculation of Pr[S|M&C&A] is more complicated. Here we assume that the alleged father is the real father, so the child's phenotype is relevant. The connection between the phenotypes of parent and child can only be understood, however, in terms of their genotypes. An appropriate thought experiment must consider the distribution of genotypes in the racial groups of the mother and alleged father. Fortunately, these distributions can be estimated from the distributions of phenotypes.[21] We consider the population of triples of genotypes that result from drawing genotypes for the mother and father at random (still assuming they are unrelated), and then crossing these two genotypes to obtain the genotype for a child. We then consider the subset of this population consisting of those triples of genotypes that give the observed phenotypes for mother and child. We imagine determining the father's genotype by drawing at random from this subset; Pr[S|M&C&A] is the frequency in this subset of genotypes for the father that give the phenotype S.

In this thought experiment, we are accounting for more than the mere fact that the alleged father is not excluded. We are using more genetic theory and more detailed statistical data. Assuming we have the data, the fact that we are taking it into account makes this a better thought experiment and a stronger probability argument.

The ratio Pr[S|M&C&A]/Pr[S|M&C&(not-A)] is widely used; it is often called the "paternity index" in the literature. If it is very large, we say that the alleged father's phenotype is much more likely if he is the real father than if not, and we construe this as a strong argument for his being the real father.

Writing Pr[S|M&C&A]/Pr[S|M&C&(not-A)] for the paternity index makes it explicit that we are holding the phenotypes of the mother and child fixed in the thought experiment. If we are content to leave this implicit, then we can write Pr[S|A]/Pr[S|not-A] for the index. This makes it clear that it is a likelihood ratio; it is the likelihood ratio L[S|A], calculated by holding the phenotypes of the mother and child fixed.

D. *A Yet More Inclusive Likelihood*

It might be argued that the thought experiment of the preceding section does not yet account for all the evidence blood tests provide. When we compare Pr[S|M&C&A] with PR[S|M&C&(not-A)], we are accounting for the phenotypes M and C to some extent—we are considering them fixed in the thought experiment--but we are not taking into account that M and C may suggest genotypes for the real father that are more common in some racial groups than others. This point becomes clear when we recognize that the total evidence provided by the three phenotypes S, M, and C would be fully accounted for if we could calculate joint probabilities for these phenotypes under the two hypotheses, A and not-A. Thus, instead of considering only the ratio

Pr[S|M&C&A]/Pr[S|M&C&(not-A)]

Figure 4

we should also consider the ratio

$$\Pr[S\&M\&C|A]/\Pr[S\&M\&C|\text{not-}A] = [\Pr[M\&C|A]/\Pr[M\&C|\text{not-}A]]\,[\Pr[S|M\&C\&A]/\Pr[S|M\&C\&(\text{not-}A)]]$$

Figure 5

In order to calculate this ratio, we need thought experiments that will yield Pr[M&C|A] and Pr[M&C|not-A]. The probability Pr[M&C|A] is obtainable from the thought experiment already used to calculate Pr[S|M&C&A]. We imagine genotypes for the mother and alleged father being drawn at random from their racial groups and then crossed to produce the genotype of the child, and calculate the probability that the phenotypes M and C will result from this process. But the probability Pr[M&C|not-A] poses problems. A thought experiment for this probability must specify the racial group of the real father, or at least specify a probability distribution for his racial group. If we assume that the real father belongs to the same racial group as the alleged father, then Pr[M&C|not-A] = Pr[M&C|A], and Figure 5 reduces to Figure 4. But if we cannot confidently assume this, then we probably will not be able to construct a convincing thought experiment for the selection of the real father's racial group.

In view of the difficulties encountered in carrying out the thought experiment represented by Figure 5, we should not necessarily prefer it to the one represented by Figure 4. The former does try to take more evidence into account, but it does not always succeed. A lawyer trying to construct the strongest possible argument against the alleged father might be better advised to rest content with a large ratio for Figure 4 than to weaken her argument by resting it on a questionable thought experiment about the racial group of the real father.

This, of course, leaves a possible role for Figure 5 in rebuttal. The defense, wishing to discredit the argument based on Figure 4, may be able to suggest a plausible racial group for the real father for which Pr[M&C|not-A] is much greater than Pr[M&C|A], resulting in a low value for Figure 5. If the racial group is only plausible, then this thought experiment might not count as a direct argument against the allegation of paternity, but it would effectively rebut the plaintiff's argument based on Figure 4.

E. *A Bayesian Thought Experiment*

A full Bayesian thought experiment would go one step further than the thought experiment represented by Figure 5. It would multiply that likelihood by prior odds for paternity based on other evidence such as the mother's allegation, the man's denial, and evidence bearing on the trustworthiness of their respective testimonies.

Some suggest that only such a full Bayesian thought experiment is truly useful because other thought experiments do not yield a number properly labelled "the probability of paternity." From the constructive viewpoint, this is unconvincing. The Bayesian thought experiment does not yield *the* probability of paternity in any absolute sense. It is merely one argument: other thought experiments may be more persuasive arguments, even though on the surface they address slightly different questions. In fact, full Bayesian arguments in paternity cases are usually weaker than arguments based solely on likelihoods from blood tests because the analogy to the picture of chance using the evidence from the blood tests is stronger than the analogy using other evidence.

F. *Where Do We Stop?*

We have studied a series of successively more complex thought experiments. Each takes more evidence into account, but by doing so risks weakening the analogy to the picture of chance. Where in this progression should we stop? When is the advantage of taking more evidence into account outweighed by the disadvantage of treating it less convincingly? There is no general answer; we must consider the circumstances of the individual case.

Some readers, although willing to stop short of the full Bayesian argument, will insist that the only reasonable stopping point is the argument that calculates the full likelihood—the likelihood given by Figure 5. These readers should recognize that there really is no such thing as a full likelihood because the evidence to be taken into account by the likelihood is not well-defined. We first studied a likelihood that accounted only for the fact that the alleged father was not excluded, then a likelihood that accounted for his phenotype, then a likelihood that accounted for all three phenotypes—the mother's, the child's, and the alleged father's. Why stop there? Why not consider a likelihood that accounts for all three phenotypes plus either the alleged father's denial, or some other piece of evidence introduced in court?

Shafer and Tversky point out that, outside of the realm of planned statistical experiments, problems do not come with the evidence on which likelihoods are to be based distinguished from the evidence on which prior probabilities are to be based.[22] This partitioning of the evidence must be done deliberately by the person who constructs the Bayesian argument. The standard terminology encourages a pretense that the partitioning is determined by timing: likelihoods are based on new evidence, priors on old evidence. But this is only pretense. It is clear in our example that it is pretense because the blood tests are unlikely to be the last evidence we obtain, no matter whose shoes we are in.

G. *The Relevance of Likelihood*

Several authors, including Professor Lempert[23] and Professor Kaye,[24] have advanced likelihood as a general tool for explicating and correcting legal doctrine. Professor Lempert, for example, explains judgments of relevance in terms of likelihood ratios. Given evidence is logically relevant, he explains, unless its likelihood ratio is close to one. Such evidence may be excluded, nonetheless, from a jury's consideration if the jury is ill-equipped to estimate this likelihood.

This reliance on likelihood is misguided because it overlooks the constructive nature of likelihood. A cogent likelihood argument does establish the relevance of the evidence it uses, but if we have not succeeded in constructing a cogent likelihood argument, then there is no content in talk about whether the likelihood ratio is close to one, and we are free to consider other ways of establishing the relevance of the evidence.

Professor Lempert's references to difficulties that a judge or jury might have in estimating a likelihood suggest that a likelihood, as opposed to a prior probability, has objective reality. Even in the case of blood test evidence, this is not necessarily so. If we cannot identify a population from which we can say the observed phenotypes were randomly drawn, then our likelihoods are simply meaningless symbols.

We might defend a reliance on likelihood by appealing to Savage's personalistic interpretation, according to which an individual must always have both subjective prior probabilities and subjective likelihoods.[25] We could then say that given evidence is relevant if any reasonable juror can give it a likelihood ratio different from one. This approach leads, however, to all the difficulties associated with the personalistic pretense to all-encompassing probability opinions. Moreover, it leads to the paradoxical result that evidence is relevant if it is weak. When we know little, almost any likelihood ratio may seem reasonable.[26]

V. Is There a Theory of Argument Construction?

If Bayes's Theorem is not a general recipe for constructing probability arguments, how do we construct them? Can we develop a theory that tells how to do so? This question has gained prominence in recent years because computer scientists have begun trying to incorporate the ability to construct probability arguments into expert systems.[27] Their initial efforts have had limited success, but the effort itself represents an important challenge to statisticians and probabilists.

Genuine progress towards automating the construction of probability arguments will depend in part on progress in the construction of artificial associative memories.[28] Human probability judgment, poor as it is, depends on an ability to retrieve memory of situations that are fairly similar to a given situation. It also depends on an ability to evaluate the relevance of the instances retrieved and to adjudicate between conflicting instances and analogies. How can we mimic this human ability?

It is here that the courtroom metaphor may have its greatest impact. The adversary system, the system the law has evolved to deal with conflicting arguments, may turn out to be an essential element of automated probability argument.

NOTES

* Professor of Statistics, University of Kansas School of Business. The author has benefited from conversations with Mikel Aickin, Paul Meir, and Howard Stratton. Research for the article has been partially supported by NSF grant IST-8405210.

[1] Personal communication with Lorraine Daston (1984).

[2] J. BERNOULLI, ARS CONJECTANDI (Basel 1713).

[3] Garber & Zabell, *On the Emergence of Probability*, 21 ARCHIVE FOR HIST. EXACT SCI. 33 (1979).

[4] Shafer & Tversky, *Languages and Designs for Probability Judgment*, 9 COGNITIVE SCI. 309 (1985).

[5] *See, e.g.*, Schum, *The Behavioral Richness of Cascaded Inference Models: Examples in Jurisprudence*, in 2 COGNITIVE THEORY 149 (1977).

[6] *See, e.g.*, H. SOLOMON, MEASUREMENT AND THE BURDEN OF EVIDENCE (1983).

[7] L. COHEN, THE PROBABLE AND THE PROVABLE (1977).

[8] Ekelöf, *Free Evaluation of Evidence*, 8 SCANDINAVIAN STUD. L. 45 (1964); *see also* Ekelöf, *Beweiswert*, in FESTSCHRIFT FÜR FRITZ BAUER 343 (W. Grunsky, R. Sturner, G. Walter, and M. Wolf eds. 1981); Ekelöf, *My Thoughts on Evidentiary Values*, in EVIDENTIARY VALUE: PHILOSOPHICAL, JUDICIAL AND PSYCHOLOGICAL ASPECTS OF A THEORY 8 (P. Gardenfors, B. Hansson, and N. Sahlin eds. 1983).

[9] Shafer, *Lindley's Paradox*, 77 J. AM. STATISTICAL A. 325 (1982).

[10] L. SAVAGE, THE FOUNDATIONS OF STATISTICS (1954).

[11] *See, e.g.*, M. FINKELSTEIN, QUANTITATIVE METHODS IN LAW (1978); R. BENDER & A. NACK, TATSACHENFESTSTELLUNG VOR GERICHT (1981).

[12] *See* Savage, *Elicitation of Personal Probabilities and Expectations*, 66 J. AM. STATISTICAL A. 783 (1971).

[13] *See* Shafer and Tversky, *supra* note 4.

[14] *See generally* G. SHAFER, A MATHEMATICAL THEORY OF EVIDENCE (1976); *see also* Shafer, *supra* note 9; Shafer, Belief Functions and Possibility Measures, to be printed in 1 THE ANALYSIS OF FUZZY INFORMATION (J. Bezdek ed. 1986).

[15] Freund, *Puzzle or Paradox?*, 19 AM. STATISTICIAN no. 4, 29, 44 (1965).

[16] *See* Shafer, *Conditional Probability*, 53 INT'L STATISTICAL REV. 261 (1985).

[17] *See id.*

[18] *See* Speed, *Discussion of* Conditional Probability, *by Glenn Shafer*, 53 INT'L STATISTICAL REV. 276 (1985).

[19] Essen-Möller, *Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis, theoretische Grundlagen*, 68 MITT. ANTHROP. GES. 9, 598 (1938). For an excellent review, written primarily for statisticians, see Berry & Geisser, *Inference in Cases of Disputed Paternity* (1982) (Technical Report No. 404, available at the Department of Statistics at the University of Minnesota).

[20] *See* Lee, *Estimation of the Likelihood of Paternity*, in PATERNITY TESTING 28 (H. Polesky ed. 1975); *see also* Weiner, *Likelihood of Parentage*, in PATERNITY TESTING BY BLOOD GROUPING 125 (M. Sussman 2d ed. 1976).

[21] *See* Berry & Greisser, *supra* note 19, at 3-8.

[22] *See* Shafer & Tversky, *supra* note 4.

[23] Lempert, *Modeling Relevance*, 75 MICH. L. REV. 1021 (1977).

[24] Kaye, *Probability Theory Meets Res Ipsa Loquitur*, 77 MICH. L. REV. 1456 (1979).

[25] *See* Savage, *supra* note 10.

[26] I discuss this point in more technical detail elsewhere. See G. Shafer, *Comment*, 66 B.U. L. Rev. 629 (1986).

[27] *See, e.g.*, B. BUCHANAN & E. SHORTLIFFE, RULE-BASED EXPERT SYSTEMS: THE MYCIN EXPERIMENTS OF THE STANFORD HEURISTIC PROGRAMMING PROJECT (1984).

[28] *See generally* T. KOHONEN, SELF-ORGANIZATION AND ASSOCIATIVE MEMORY (1984).

Glenn Shafer,
Professor of Statistics,
University of Kansas School of Business.