The Role of the Bayes Factor in the Evaluation of Evidence

Citation for published version:
Aitken, C, Taroni, F & Bozza, S 2024, 'The Role of the Bayes Factor in the Evaluation of Evidence', *Annual Review of Statistics and its Application*, vol. 11, pp. 203-226. https://doi.org/10.1146/annurev-statistics-040522-101020

Digital Object Identifier (DOI):
10.1146/annurev-statistics-040522-101020

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Annual Review of Statistics and its Application

ANNUAL REVIEWS

Annual Review of Statistics and Its Application

The Role of the Bayes Factor in the Evaluation of Evidence

Colin Aitken,[1] Franco Taroni,[2] and Silvia Bozza[3,2]

[1]School of Mathematics and the Maxwell Institute of Mathematical Sciences, The University of Edinburgh, Edinburgh, United Kingdom; email: cgga@ed.ac.uk

[2]School of Criminal Justice, The University of Lausanne, Lausanne, Switzerland

[3]Department of Economics, Ca' Foscari University of Venice, Venice, Italy

Annu. Rev. Stat. Appl. 2024. 11:203–26

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-040522-101020

Keywords

Bayes factor, Bayesian network, forensic science, likelihood ratio, probabilistic reasoning, scientific evidence, value of evidence

Abstract

The use of the Bayes factor as a metric for the assessment of the probative value of forensic scientific evidence is largely supported by recommended standards in different disciplines. The application of Bayesian networks enables the consideration of problems of increasing complexity. The lack of a widespread consensus concerning key aspects of evidence evaluation and interpretation, such as the adequacy of a probabilistic framework for handling uncertainty or the method by which conclusions regarding how the strength of the evidence should be reported to a court, has meant the role of the Bayes factor in the administration of criminal justice has come under increasing challenge in recent years. We review the many advantages the Bayes factor has as an approach to the evaluation and interpretation of evidence.

1. THE BAYES FACTOR AND ITS ROLE IN THE ADMINISTRATION OF JUSTICE

1.1. Definition and Example

In a civil or criminal court case, the trier of fact—the judge or jury—has to make a decision. In a civil case, the decision is to find in favor of the plaintiff or defendant, and in a criminal case, the decision is to find the defendant guilty or not guilty. In this article, the focus is on criminal justice. In a trial, evidence E is presented to aid the trier of fact. Each trier of fact brings to the court their own background knowledge I, which, consciously or otherwise, affects their opinion on, or belief in, the guilt or otherwise of the defendant. Beliefs have differing strengths; there can be a strong belief, a weak belief, or beliefs of other strengths. The strength of belief may be thought of as a measure of belief. A numerical equivalence as a measure of belief, to the verbal measure as a strength, is given by subjective probability. A strong belief may be represented by a probability close to one, and a weak belief may be represented by a probability close to zero. In a trial, the trier of fact's belief in the guilt of the defendant prior to the presentation of evidence may be represented by the probability $\Pr(H_p \mid I)$, where H_p is the proposition favored by the prosecution that the defendant is guilty. It is to be hoped that, under the presumption of innocence until proven guilty, this probability is close to zero. After the presentation of the evidence E, it is reasonable to assume that the probability will have changed. It may now be represented by $\Pr(H_p \mid E, I)$, a probability posterior to the presentation of E at the trial. This latter probability may now be so large (i.e., close to one) that it is beyond a threshold probability that has to be achieved for guilt to be found beyond reasonable doubt.

The rules of probability show how the change from prior probability, given I, $\Pr(H_p \mid I)$ to the posterior probability $\Pr(H_p \mid E, I)$ may be made. The change is described in an uncontroversial theorem, known as Bayes' theorem. The defense has its own proposition, H_d, which may just be the negation of H_p. Bayes' theorem relates the prior odds, $\Pr(H_p \mid I)/\Pr(H_d \mid I)$, to the posterior odds, $\Pr(H_p \mid E, I)/\Pr(H_d \mid E, I)$, by the equation

$$\frac{\Pr(H_p \mid E, I)}{\Pr(H_d \mid E, I)} = \frac{\Pr(E \mid H_p, I)}{\Pr(E \mid H_d, I)} \times \frac{\Pr(H_p \mid I)}{\Pr(H_d \mid I)}.$$

The factor

$$\frac{\Pr(E \mid H_p, I)}{\Pr(E \mid H_d, I)}$$

that converts the prior odds in favor of H_p to posterior odds in favor of H_p is known as the Bayes factor. When the competing propositions take the form of simple hypotheses, the Bayes factor simplifies to a likelihood ratio. It is nonnegative with no upper bound.

Many practical situations are encountered where available measurements are in the form of realizations of experiments that assume only two mutually exclusive outcomes. Consider a scenario involving shoe marks recovered from a crime scene. A simplistic derivation of the value of the evidence illustrates the role of the Bayes factor. A person of interest owns a pair of shoes that produce prints, of type T, that are indistinguishable from marks recovered from the crime scene. The evidence E has two components, so that E may be written as $E = (E_r, E_c)$, where $E_r = T$, the recovered material, and $E_c = T$, the material of the pair of shoes owned by the person of interest, and thus whose source is known and that may be called the control material. Assume the probability that a shoe print is of type T is θ. The prosecution proposition, H_p, is that the shoe prints and shoe marks originate from the same source. The defense proposition, H_d, is that the shoe prints and shoe marks originate from different sources. If H_p is true, and conditioning on θ

and ignoring I for the moment, the numerator of the Bayes factor is

$$\Pr(E \mid H_{\mathrm{p}}, \theta) = \Pr(E_{\mathrm{r}}, E_{\mathrm{c}} \mid H_{\mathrm{p}}, \theta) = 1,$$

as the probability of a match between the shoe print and the shoe mark if the shoe mark is made by the shoe from which the shoe print is taken is 1, under the assumptions that there are no transcription or contamination issues (not discussed here due to space constraints). If H_{d} is true, and still conditioning on θ and ignoring I for the moment, the denominator of the Bayes factor is

$$\Pr(E \mid H_{\mathrm{d}}, \theta) = \Pr(E_{\mathrm{r}}, E_{\mathrm{c}} \mid H_{\mathrm{d}}, \theta) = \theta,$$

as the probability of a match between the shoe print and the shoe mark if the shoe mark is not made by the shoe from which the shoe print is taken is the probability of a random match, which is θ.

The likelihood ratio is then

$$\frac{\Pr(E_{\mathrm{r}}, E_{\mathrm{c}} \mid H_{\mathrm{p}}, \theta)}{\Pr(E_{\mathrm{r}}, E_{\mathrm{c}} \mid H_{\mathrm{d}}, \theta)} = \frac{1}{\theta}.$$

Any uncertainty in θ needs to be modeled. A beta prior distribution $\mathrm{Be}(\alpha, \beta)$ can be used for this purpose. A police database reports a total number of n prints of type T out of a database of N shoe prints; this may be taken as the background information I. If H_{p} is true, there are then $n + 1$ shoe prints of type T, out of $N + 1$ distinct shoe prints. Similarly, if H_{d} is true, then there are $n + 2$ shoe prints of type T, out of $N + 2$ distinct shoe prints. If the recovered shoe mark and the control shoe print originate from the same source (i.e., if H_{p} holds), and available measurements are in the form of realizations of independent counts that can be well modeled by a binomial distribution, the probability of the evidence can be obtained as

$$\begin{aligned}\Pr(E_{\mathrm{r}} = T, E_{\mathrm{c}} = T \mid H_{\mathrm{p}}, I) &= \Pr(E_{\mathrm{r}}, E_{\mathrm{c}} \mid H_{\mathrm{p}}, n, N - n) \\ &= \int_{\Theta} \theta \binom{N}{n} \theta^{n} (1-\theta)^{N-n} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}/B(\alpha, \beta) \mathrm{d}\theta \\ &= \int_{\Theta} \theta \theta^{\alpha+n-1}(1-\theta)^{\beta+N-n-1}/B(\alpha+n, \beta+N-n) \mathrm{d}\theta,\end{aligned}$$

where $\Theta = [0, 1]$, $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$, and $\Pr(E_{\mathrm{r}} = E_{\mathrm{c}})$ if H_{p} is true equals θ in this context.

If the recovered shoe mark and the control shoe print originate from different sources (i.e., if H_{d} holds), the probability of the evidence can be obtained as

$$\begin{aligned}\Pr(E_{\mathrm{r}} = T, E_{\mathrm{c}} = T \mid H_{\mathrm{d}}, I) &= \Pr(E_{\mathrm{r}}, E_{\mathrm{c}} \mid H_{\mathrm{d}}, n, N - n) \\ &= \int_{\Theta} \theta^2 \binom{N}{n} \theta^{n} (1-\theta)^{N-n} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}/B(\alpha, \beta) \mathrm{d}\theta \\ &= \int_{\Theta} \theta^2 \theta^{\alpha+n-1}(1-\theta)^{\beta+N-n-1}/B(\alpha+n, \beta+N-n) \mathrm{d}\theta,\end{aligned}$$

where $\Pr(E_{\mathrm{r}} = E_{\mathrm{c}})$ if H_{d} is true equals θ^2 in this context. The value of the evidence is then

$$\begin{aligned}\frac{\int_{\Theta} \theta^{\alpha+n}(1-\theta)^{\beta+N-n-1} \mathrm{d}\theta}{\int_{\Theta} \theta^{\alpha+n+1}(1-\theta)^{\beta+N-n-1} \mathrm{d}\theta} &= \frac{B(\alpha+n+1, \beta+N-n)}{B(\alpha+n+2, \beta+N-n)} \\ &= \frac{\alpha+\beta+N+1}{\alpha+n+1}.\end{aligned}$$

This is the reciprocal of the posterior mean of θ. In such a case, the beta posterior distribution for θ is $\text{Be}(\alpha + n + 1, \beta + N - n)$, where the prior parameter α is increased by the number of items $(n + 1)$ that match the recovered shoe mark, and the prior parameter β is increased by the number of items $(N - n)$ that do not match the recovered shoe mark. It has been argued (e.g., Sjerps et al. 2016, Morrison & Enzinger 2016) that a report of such a value (the reciprocal of the posterior mean of θ) would deprive the legal system of relevant information about the case. However, all available information, including prior uncertainty about the unknown value of θ, is encapsulated in the reported value of the Bayes factor, as explained in more detail in Section 5. Clearly, different prior probability distributions may be used, or different databases might be available, and in such cases a different value will be reported.

Implementation of Bayes' theorem assumes that the evidence E is known without uncertainty (such evidence is often termed unequivocal evidence). The philosopher Richard Jeffrey developed a generalization of Bayes' theorem, known as Jeffrey conditionalization, for the situation where there is doubt about the truth of the evidence E (Jeffrey & Hendrickson 1989, Jeffrey 1992). This development is not discussed here as it is beyond the scope of this article. The interested reader is referred to Taroni et al. (2020), who introduce a Bayes factor for situations where there is uncertainty about the knowledge of the evidence, so-called equivocal evidence.

The European Network of Forensic Science Institutes (ENFSI 2015, p. 6) recommends the use of probability to quantify uncertainty and the use of what it calls the likelihood ratio:

> Evaluation. . .is based on the assignment of a likelihood ratio. Reporting practice should conform to these logical principles. This framework for evaluative reporting applies to all forensic science disciplines. The likelihood ratio measures the strength of support the findings provide to discriminate between propositions of interest. It is scientifically accepted, providing a logically defensible way to deal with inferential reasoning.

In applications to forensic science, the Bayes factor is usually referred to as the likelihood ratio, though we emphasize that the Bayes factor does not always simplify in this way.

1.2. The Weight of Evidence and Multiple Items of Evidence

The Bayes factor has intuitively pleasing properties. A value greater (less) than one in which the evidence has a higher (lower) probability if H_p is true than if H_d is true increases (decreases) the odds in favor of H_p. There is a good analogy with the scales of justice if logarithms are used. Consider a logarithmic transformation of the odds form of Bayes' theorem:

$$\begin{aligned}\log\left\{\frac{\Pr(H_\text{p} \mid E, I)}{\Pr(H_\text{d} \mid E, I)}\right\} &= \log\left\{\frac{\Pr(E \mid H_\text{p}, I)}{\Pr(E \mid H_\text{d}, I)}\right\} + \log\left\{\frac{\Pr(H_\text{p} \mid I)}{\Pr(H_\text{d} \mid I)}\right\} \\ \Rightarrow \log\{\Pr(H_\text{p} \mid E, I)\} - \log\{\Pr(H_\text{d} \mid E, I)\} &= \log\{\Pr(E \mid H_\text{p}, I)\} - \log\{\Pr(E \mid H_\text{d}, I)\} \\ &\quad + \log\{\Pr(H_\text{p} \mid I)\} - \log\{\Pr(H_\text{d} \mid I)\} \\ \Rightarrow \log\{\Pr(H_\text{p} \mid E, I)\} - \log\{\Pr(H_\text{d} \mid E, I)\} &= \log\{\Pr(E \mid H_\text{p}, I)\} + \log\{\Pr(H_\text{p} \mid I)\} \\ &\quad - \left(\log\{\Pr(E \mid H_\text{d}, I)\} + \log\{\Pr(H_\text{d} \mid I)\}\right). \qquad 1.\end{aligned}$$

The logarithms of the probabilities may be considered as weights in the scales of justice, those involving H_p on one scale and those involving H_d on the other. The expression on the left side of Equation 1 is the difference in weights on the scales after the presentation of the evidence:

$\log\{\Pr(H_p \mid E, I)\}$ on the prosecution side and $\log\{\Pr(H_d \mid E, I)\}$ on the defense side. This difference is equal to the difference in the sums of the weights for the prior probabilities and conditional probabilities of the evidence, conditional on H_p and H_d, respectively: $\log\{\Pr(E \mid H_p, I)\} + \log\{\Pr(H_p \mid I)\}$ on the prosecution side and $\log\{\Pr(E \mid H_d, I)\} + \log\{\Pr(H_d \mid I)\}$ on the defense side.

The Bayes factor also satisfies logical requirements such as adequacy, logicality, and symmetry. The interested reader can refer, for example, to Fitelson (1999, 2011), Eells (2000), and Eells & Fitelson (2002) for a detailed list of measurements of evidential value and a critical analysis. The satisfaction by the Bayes factor of all the reasonable logical requirements put forward in the philosophical literature justifies its use as a measure for the value of evidence in general and supports its use scientifically in forensic science.

Another pleasing property is the ability to consider more than one piece of evidence and to update the posterior odds sequentially as each piece of evidence is presented. Consider two pieces of evidence, E_1 and E_2. The posterior odds after presentation of E_1 may be used as the prior odds before presentation of E_2:

$$\frac{\Pr(H_p \mid E_1, E_2, I)}{\Pr(H_d \mid E_1, E_2, I)} = \frac{\Pr(E_2 \mid H_p, E_1, I)}{\Pr(E_2 \mid H_d, E_1, I)} \times \frac{\Pr(H_p \mid E_1, I)}{\Pr(H_d \mid E_1, I)}, \tag{2}$$

where the possible dependency of E_1 and E_2 is reflected in the Bayes factor $\Pr(E_2 \mid H_p, E_1, I)/\Pr(E_2 \mid H_d, E_1, I)$. The Bayes factor and the logarithm of the Bayes factor may be thought of as the value and the weight of the evidence, respectively. Good (1989, 1991) showed that, under reasonable assumptions, the Bayes factor is the best measure of the value of evidence.

Since a seminal paper by Lindley (1977) in *Biometrika*, much work has been done in the development of statistical models, such as Bayesian hierarchical multivariate models (e.g., Aitken & Lucy 2004, Bozza et al. 2008, Zadora et al. 2014), for the evaluation of evidence. This work was given added impetus in the late 1980s with the introduction of DNA profiling, which led to greater appreciation in the criminal justice system of the benefits of a probabilistic approach based on the Bayes factor.

1.3. Early Ideas

Separate from this mathematical formulation, there was an analogous debate in the legal literature about the role of probabilistic reasoning in legal cases; see, for the sake of illustration, Anglo-American legal articles by Finkelstein & Fairley (1970), Kaye (1979, 1986), Fienberg & Schervish (1986), and Fienberg (1986). Robertson & Vignaux (1993, p. 457) clarified the interest of a large majority of jurists in probabilistic methods:

> One of the main areas of interest of the so-called "New Evidence Scholarship" is the application of probability theory to arguments about facts in legal cases. As a preliminary to making [a] decision, courts have to "find facts" which requires them to reason under uncertainty. In some cases it may be the reasoning process itself which is examined in an appeal. The result may be a statement by the court about how facts ought to be thought about. Alternatively the way facts are thought about in a particular case may be seized upon as a precedent for future cases. Should there be "rules" about how facts are to be thought about? And, if so, does probability theory offer a prescription for those rules?

Philosophers of science and forensic scientists also promote the probabilistic (and Bayesian) method, for example, with reference to subjective degrees of belief (Salmon 1967, Howson & Urbach 1993, Taroni et al. 2014):

> [Assign] numbers, but these numbers are not important by themselves: what really matters is the fact that numbers allow us to use powerful rules of reasoning which can be implemented by computer programs. What is really important is not whether the numbers are "precise," whatever the meaning of "precision" may be in reference to subjective degrees of belief based upon personal knowledge. What is really important is that we are able to use sound rules of reasoning to check the logical consequences of our propositions, that we are able to answer questions like: what are the consequences with respect to the degree of belief in A if assuming that the degree of belief in B is high? And how the degree of belief in A does change, if we lower the degree of belief in B? (Taroni et al. 2014, pp. 1–2)

This approach allows scientists to assign values to their probabilities, not only by certified knowledge and experience but also by any data relevant for the event of interest, such as knowledge of an event that may be available in terms of a relative frequency. Frequency is taken here to be a term that relates to data, and probability is taken to be a term that relates to personal belief. This perspective of probability as a term that relates to personal belief is relevant to forensic science, where there are unique events (e.g., aspects of a crime) or propositions (e.g., the guilt of a defendant in a criminal trial). This idea was mentioned by de Finetti (1930, 1989), who insisted that probability is conditional on the status of the information available to the subject who assesses it. Note that alternative definitions of probability do exist (see, e.g., the discussion in Lindley 1991).

de Finetti (1931) also showed that coherence, a simple economic behavioral criterion, implies that a given individual should avoid a combination of probability assignments that is guaranteed to lead to a loss. All that is needed to ensure such avoidance is for uncertainty to be represented and manipulated using the laws of probability. In this context, the possibility of representing subjective degrees of belief in terms of betting odds is often put forward in a two-part line of argument to require that subjective degrees of belief should satisfy the laws of probability. The first part is that betting odds should be coherent, in the sense that they should not be open to a sure-loss contract. The second is that a set of betting odds is coherent if and only if it satisfies the laws of probability.

1.4. Challenges and Philosophical Properties

The role of the Bayes factor in the evaluation of evidence has been challenged in recent years. First, the role of probability (in the subjective paradigm) in legal reasoning as the best measure of uncertainty has been disputed in a discussion of the concept of relative plausibility (e.g., Allen & Pardo 2019, 2023; Aitken et al. 2022). Second, the Bayes factor quantifies the evidential value as a single number. It has been argued that a single number is insufficiently informative to provide enough nuance to the value of the evidence and that an interval is a better measure (see Section 5 for further details).

The purpose of evidence evaluation is the provision of support for a proposition, in the context of two or more propositions. Support may be qualified as weak, moderate, strong, etc. Such qualitative interpretations were proposed by Jeffreys (1983) and, more recently, in the context of forensic science (e.g., Nordgaard et al. 2012). The evidence under consideration here is taken to be scientific, in the form of continuous measurements or discrete data. Such evidence could include the elemental composition of glass for continuous measurements (Aitken & Lucy 2004) or the number of gunshot residue particles collected on the surface of individuals suspected to be involved in the discharge of a firearm for discrete data (Biedermann et al. 2009, 2011).

Consideration of the value of the evidence has to be coherent and logical. Also, there is variation associated with the evidence, and representing this variation with a probability function for categorical evidence or with a probability density function for evidence in the form of measurements is an important part of the evaluation.

There are certain properties that a measure for the value of evidence should satisfy to be coherent and logical, and there are also concepts to which it should not be related (Buckleton et al. 2020).

It should not comment on a proposition under consideration by the court. The role of the scientist evaluating the evidence should be distinct from that of the trier of fact. The measure should not be related to the presumption of innocence. The propositions for which the measure is providing support should be mutually exclusive. All of these properties are satisfied by the Bayes factor.

The distinction between the role of a factfinder and that of a witness or an expert was specified by one of the fathers of forensic science, Edmond Locard (1940, pp. 286–87):

> The physical certainty provided by scientific evidence rests upon evidential values of different orders. These are measurable and can be expressed numerically. Hence the expert knows and argues that he knows the truth, but only within the limits of the risks of error inherent to the technique. This numbering of adverse probabilities should be explicitly indicated by the expert. The expert is not the judge: he should not be influenced by facts of a moral sort. His duty is to ignore the trial. It is the judge's duty to evaluate whether or not a single negative evidence, against a sextillion of probabilities, can prevent him from acting. And finally, it is the duty of the judge to decide if the evidence is in that case, proof of guilt.

It is becoming more common to read papers published in scientific and legal journals that criticize the role of Bayes' theorem for probabilistic reasoning in the interpretation of evidence and the use of the Bayes factor for the assessment of the value of the evidence to which a scientist reports (e.g., Stiffelman 2019). Kaye & Sensabaugh (2011, p. 173) presented some examples and responded to criticisms expressed by others by affirming that it appears that "the major objection to likelihoods is not statistical but psychological."

The use of the Bayes factor is generally supported by the affirmations (Buckleton et al. 2020) that (*a*) the Bayes factor does not infringe on the ultimate issue (i.e., it does not express an opinion on the proposition of judicial interest), (*b*) the Bayesian approach clearly separates the role of the scientist from that of the decision-makers (e.g., the judge and jury) so that the scientist is distanced from comment on the hypotheses put forward by parties at trial, (*c*) the Bayes factor does not affect the reasonable doubt standard and it does not infringe on the presumption of innocence, and (*d*) the Bayes factor can be easily deduced from the ratio between posterior odds and prior odds. Those desiderata in evidential assessment clarify why one can give preference to some evaluative methods and views rather than to others. Justification of the use of Bayes' theorem in a forensic context has also been provided by, for example, Finkelstein & Fairley (1970), Lempert (1977), and Evett & Weir (1998).

More generally, desirable properties of the Bayes factor are balance, transparency, robustness, added value, flexibility, and logic. For an inferential process to be balanced (or impartial), attention cannot be restricted to only one side of the argument (Jackson 2000). Evett (1996, p. 122) noted that "a scientist cannot speculate about the truth of a proposition without considering at least one alternative proposition. Indeed, an interpretation is without meaning unless the scientist clearly states the alternatives he has considered." The requirement to consider alternative propositions is a general one that applies to many instances in daily life (Lindley 1985), but, in a legal context, the requirement is fundamental. Evett specified that

> balance means that when I am doing anything for a court of justice, I do it in full knowledge that there are two sides represented in that court. Even though the evidence that I've found appears to favor one or the other of those sides, my view of that evidence is directed not to proving the case for that side, but to helping the court to set that evidence into the context of all the other evidence and the views of both teams, prosecution and defence. (Joyce 2005, p. 37)

There is more in this quotation than the requirement to consider at least two alternatives. It also states that forensic scientists should primarily be concerned with the evidence and not with the competing propositions that are put forward to explain it. This distinction is crucial in that (as

Locard previously wrote) it provides a demarcation of the boundaries of the expert's and court's areas of competence.

Besides balance, a forensic scientist's evaluation should also comply with the requirements of transparency, i.e., it should "explain in a clear and explicit way what we have done, why we have done it and how we have arrived at our conclusion. We need to expose the reasoning, the rationale, behind our work" (Jackson 2000, p. 84). The requirement for robustness challenges a scientist's ability to explain the ground for their opinion together with their degree of understanding of the particular evidence type (Jackson 2000). These desiderata help determine the role of the scientist with regard to the evaluation and interpretation of evidence. The degree to which the scientist succeeds in meeting these criteria depends on the chosen inferential framework, which may be judged through its flexibility (a criterion that demands a form of reasoning to be generally applicable, i.e., not limited to specific subject matter) (Robertson & Vignaux 1998) and through its logic (a set of principles that qualify as rational) (Robertson & Vignaux 1993).

2. COMPLEXITY: THE USE OF PROBABILISTIC GRAPHICAL MODELS TO DEAL WITH PHENOMENA IN EVIDENCE-BASED REASONING THROUGH A COHERENT APPROACH

The application of Bayesian networks or nets (BNs) to forensic science and the evaluation of evidence, following a foundational paper with a medical example (Lauritzen & Spiegelhalter 1988), enabled the consideration of problems of increasing complexity. Roberts & Aitken (2014) review the assistance of BNs for inferential reasoning in the administration of criminal justice, and the following three quotations illustrate the benefits of their use in legal proceedings:

> Bayes nets are able to model sets of conditional probabilities in a strictly disciplined way and to put numbers on a range of compounded possibilities. [In so doing] they can supply information that could be highly informative, perhaps even decisive, in the conduct of legal proceedings. (Roberts & Aitken 2014, p. 105)

> None of this implies that *jurors* in criminal trials need to know the first thing about Bayes nets. . . . The key challenges and relationships are entirely professional, concerning how and when forensic scientists employ Bayes nets in their analyses and how properly contextualised analytical results are successfully communicated to, and comprehended by police, prosecutors, defense lawyers and trial judges. How advocates argue cases in court, and how judges sum up cases for the benefit of the jury, remain perforce questions of professional *legal* judgement and expertise. (Roberts & Aitken 2014, p. 108; emphasis in original)

> Bayes nets assist their users (e.g., forensic scientists and lawyers) to understand the structure of complex inferential problems, to form a better appreciation of mutual dependencies between uncertain events and compound probabilities, and to express this understanding in a graphical form that both assists in deepening their own comprehension and enables them to communicate their insights to others. Bayes nets help to clarify the nature of arguments predicated on probabilistic assumptions and thus promote logical analysis and rational further discussion and evaluation of factual propositions. (Roberts & Aitken 2014, pp. 109–110)

BNs have attracted researchers in forensic science since the late 1980s (Aitken & Gammerman 1989), and this attention has intensified considerably throughout the last decade (see, e.g., Taroni et al. 2014, Dawid & Mortera 2021).

BNs can be loosely defined as a pictorial representation of the dependencies and influences (represented by arcs) among variables (represented by nodes) deemed to be relevant for a particular probabilistic inferential problem. BNs are a combination of graph theory, which is used to provide a qualitative model structure, and probability theory, which is used to characterize the nature and strength of the relationships that reign within a model. More formally, a BN covers the following elements:

- A finite collection of random variables that are represented by nodes. Each of these nodes either has a finite set of mutually exclusive states or may represent a continuous measurement.
- A set of directed arcs that connect pairs of nodes.
- A combination of the set of variables and the set of directed arcs in such a way that a directed acyclic graph is obtained, i.e., a graph where no loops are permitted.
- The association of node probability tables with each variable of the network. The probability table of a variable, say A, that receives entering arcs from variables $B_1, \ldots, B_n$ contains conditional probabilities $\Pr(A \mid B_1, \ldots, B_n)$, whereas a variable A with no entering arcs from other variables contains unconditional probabilities $\Pr(A)$. It is assumed that personal degrees of belief can be assigned to these states when, but not only when, relevant data are unavailable.

The actual state of a variable may not be known with certainty. For example, there may be uncertainty about the truth or otherwise of a proposition that explains why, for example, a crime stain has been left by the person of interest (e.g., they were guilty of the crime). Within a BN, such a proposition is conceptualized in terms of a Boolean node, whose states represent the truth and the falsity of that proposition. The degree of belief maintained in each of these states is expressed numerically, i.e., in terms of probabilities. These probabilities are organized in that node's probability table. The arcs in a BN represent relationships that correspond to a property that a modeler assumes to hold within the context of an inferential problem at hand. A directed arc from a node H to a node E signifies that variable H has a direct influence on variable E. In **Figure 1**, node H can have two states, H_p (for prosecution) and H_d (for defense). Node E may be the outcome of a comparison between the DNA profile of a bloodstain found at a crime scene and the profile of a person of interest. Then, the probability of E is dependent on the state of node H (i.e., states H_p or H_d).

A key task of BNs is to process newly acquired information—i.e., to revise the conditional probabilities of the states of the nodes in the network in which one is interested (e.g., a proposition node) given that the states of some other nodes (e.g., evidence nodes) have been observed.

The analysis of inferential interactions plays an important role in the description of the line of reasoning for a forensic evaluator in a case involving several items of evidence. Consider the following example (Taroni et al. 2014). A young girl, Lulu, has been found murdered in her home with many knife wounds. Bloodstains found at the crime scene have a DNA profile that does not match Lulu's. Jack, a friend of Lulu's, has been seen by a witness, John, near her house around the time of the murder. Jack was said to be in love with Lulu. John said that he, John, was also in love with Lulu. This information may be used to form the BN in **Figure 2**, where the Boolean nodes represent the following propositions, scientific and nonscientific evidence:

- Proposition: Jack stabbed Lulu.
- Scientific evidence: The bloodstain at the scene was determined to have come from the offender.

Figure 1

Simple two-node Bayesian network for a proposition, E, relating to a scientist's observation (i.e., evidence) of corresponding features between questioned and known materials, and propositions, H, referring to a common source (Aitken et al. 2021, p. 269).

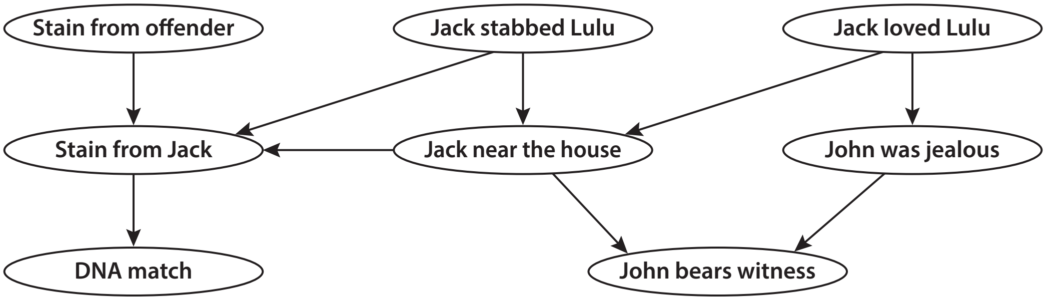

Figure 2

A Bayesian network to illustrate the interaction between scientific and nonscientific (testimony) evidence in an assault case (Taroni et al. 2014, pp. 56–57). Node descriptions are given in the text.

- Testimony: Jack loved Lulu.
- Proposition: The stain came from Jack.
- Scientific evidence: The DNA profile of the stain matches that of Jack.
- Proposition: John was jealous of Jack.
- Testimony: John bears witness that Jack was near the scene shortly after the time the crime was committed.
- Proposition: Jack was near the house shortly after the time the crime was committed.

The network could be enlarged by accounting for the testimony, and its reliability, of a witness that John was jealous of Jack, and also by accounting for an analysis of eyewitness reliability through consideration of the variables involved in the analysis of testimonial evidence (Schum 1994, pp. 100–114, 324–344).

Studies on the evidential foundations of probabilistic reasoning may be extended by using the notion of the association of weight of evidence with the measurement of evidential phenomena in the presence of a mass of evidence, which gives rise to complex reasoning patterns. Consideration of masses of evidence and their role in a given criminal case (Schum 1994) leads to methods to measure the following features: (*a*) relative contributions of items of evidence to the overall weight of evidence, (*b*) interactions among items of evidence, and (*c*) dissonances among items of evidence. Clear definitions of the measurements of all three features help with inferential tasks involving the combination of items of evidence from amid a mass of evidence. Measurements of these three features enable a detailed examination of recurrent phenomena in evidence-based reasoning, such as convergence, contradiction, redundancy, and synergy, to be made. P. Juchli, F. Taroni & C.G.G. Aitken (manuscript submitted) integrate these evidential phenomena into the development of the formulae for Bayes factors for the evaluation and interpretation of evidence.

Masses of evidence and their interaction measures have a role in the management of justice. The idea that any scientific or judicial decision should be based on the available evidence, with a potential need to gather more evidence, has been explored from various points of view (see, e.g., Raiffa & Schlaifer 1961, Good 1985)—for example, through a decision-making approach calling for a calculus of the expected value of sample information (Good 1967) and through the expectation of the weights of evidence considered as a criterion for the value of an experimental design (Good 1979). The intuitively attractive notion that it is better to use all available information is a simple consequence of Bayesian confirmation theory, where posterior probabilities can be updated with the acquisition of new information in order to discriminate better between hypotheses (Carnap 1947; F. Taroni, C.G.G. Aitken, S. Bozza & P. Juchli, manuscript submitted).

Historically, Ayer (1957) was the first to ask why new observations should be made. He related the question to Carnap's (1947) principle of total evidence, where the use of all available information is recommended in the assignment of a probability. In other words, "the requirement of total evidence. . .says that, in evaluating a hypothesis, you should take account of all evidence you have" (Barrett & Sober 2020, p. 191). In the Bayesian paradigm, the assessment of the probability of the truth of a hypothesis should be conditioned on all the available evidence. This approach represents an example of what Carnap defined as the methodology of induction.

Bayesian decision theory explains how the expected value of information may be calculated by taking advantage of the recommendation to maximize the expected utility. This approach is not new in forensic science. Consider a scenario involving information provided by finger marks yet to be processed in a forensic science laboratory. Gittelson et al. (2013) examined the question of whether or not to process a finger mark from a decision-theoretic point of view, based on theoretical work by Lindley (1985) with a series of forensic applications by Taroni et al. (2010) and Gittelson (2013). The question was answered with a quantified expression of the expected value of the information associated with the processed finger mark, which could be compared with the cost of processing the mark. A general review of Bayes factors for forensic decision analysis from an operational perspective, with practical relevance and applicability that keeps theoretical and philosophical justifications limited, is given by Bozza et al. (2022).

3. PROBABILISTIC CONFIRMATION THEORY

The impact of an item of evidence on the credibility of a hypothesis can be studied through what is known as probabilistic confirmation theory (Maher 1996, Crupi & Tentori 2016, Taroni et al. 2021). This theory provides both a qualitative and a quantitative response to the question of whether or not a piece of evidence E confirms, is neutral with respect to, or disconfirms a hypothesis of interest H:

- E confirms or supports H if and only if $\Pr(H \mid E, I) > \Pr(H \mid I)$.
- E is neutral with respect to H if and only if $\Pr(H \mid E, I) = \Pr(H \mid I)$.
- E disconfirms or undermines H if and only if $\Pr(H \mid E, I) < \Pr(H \mid I)$.

An appropriate measure $c(E, H)$ of the degree of confirmation (or degree of support) that a hypothesis H receives from information E is specified as one that quantifies the change in belief of H. Such a measure c does not initially need to be either a probability or a function of a probability, and a question of interest is whether some appropriate function of probability can be such a measure of confirmation. Philosophers of science and statisticians have expressed some reasonable requirements for a quantitative confirmation measure c and have shown that probability satisfied them. In particular, these logical requirements are satisfied by the Bayes factor and its logarithm.

A list of those requirements is provided, for example, by Crupi & Tentori (2014). As an illustration, consider the so-called compatibility requirement. Take evidence E that a DNA profile from a person of interest matches, in some sense, that of a crime stain, and the proposition H that the person of interest is the source of the crime stain. Then E may be said to confirm H. Alternatively, consider evidence F that a DNA profile from a person of interest does not match that of a crime stain, and the same previous proposition H. Then F may be said to disconfirm H and $c(E, H) > c(F, H)$. Another reasonable assumption is that the confirmation measure depends solely on the degrees of belief about the two events of interest (this is called the formality requirement), so it depends only on the probability values concerning E, F, and H.

A confirmation measure should also satisfy a classificatory requirement that confirmation implies $c(E, H) > 0$, neutrality implies $c(E, H) = 0$, and disconfirmation implies $c(E, H) < 0$, as illustrated in the DNA profile example, where evidence E confirms H and F disconfirms H, so that $\Pr(H \mid E, I) > \Pr(H \mid I)$ and $\Pr(H \mid F, I) < \Pr(H \mid I)$. This implies that $c(E, H) > 0$ and $c(F, H) < 0$.

Different measures agree with those requirements. Measures that agree are the Bayes factor and any monotonic function of the Bayes factor, such as its logarithm (Good 1950). The Bayes factor confirms a hypothesis if its value is greater than one and disconfirms the hypothesis if the value is less than one; its logarithm confirms a hypothesis if its value is greater than zero and disconfirms the hypothesis if the value is less than zero.

Any form of presentation for the evidence that is adopted by scientists must be demonstrably logical according to well-defined criteria. So-called Bayesian confirmation measures or evidential support measures offer a numerical expression for the impact of a piece of evidence on a judicial hypothesis of interest. The Bayes factor satisfies a number of necessary conditions on normative logical adequacy; the same cannot be said for alternative expressions put forward in some legal and forensic circles.

As an illustration, compare the properties of the Bayes factor with two alternative expressions for evidential value. Consider two mutually exclusive and exhaustive propositions, H_p and H_d, and evidence E. The three expressions considered are:

1. The Bayes factor: $BF = \Pr(E \mid H_p)/\Pr(E \mid H_d)$
2. The difference between the posterior probability and the prior probability of H_p, given E: $D = \Pr(H_p \mid E) - \Pr(H_p)$
3. The ratio of the posterior probability and the prior probability of H_p given E: $R = \Pr(H_p \mid E)/\Pr(H_p)$

One desirable property of a measure for the value of evidence is that it takes minimal and maximal values that are independent of E, H_p, and H_d. Otherwise, it would not be possible to compare values between different items of evidence and propositions. Fitelson (2006) defined a property of logicality that was satisfied by a measure of evidential value that was maximal when evidence E implied the proposition H_p, in that $\Pr(H_p \mid E) = 1$, and minimal when evidence E implied the complement of H_p, namely H_d, in that $\Pr(H_d \mid E) = 1$ and so $\Pr(H_p \mid E) = 0$. Consider the three expressions immediately above in this context:

1. When evidence E implies hypothesis H_p, $\Pr(H_p \mid E) = 1$ and, hence, $\Pr(H_p \mid E)/\Pr(H_d \mid E) = \infty$. The Bayes factor is the ratio of posterior odds to prior odds and is thus equal to ∞ and is at its maximum. If evidence E implies hypothesis H_d, then $\Pr(H_p \mid E) = 0$, $\Pr(H_p \mid E)/\Pr(H_d \mid E) = 0$, and the Bayes factor takes its minimal value of zero. These maximal and minimal values are independent of the evidence and the propositions.
2. When E implies H_p, $D = 1 - \Pr(H_p)$, and when E implies H_d, $D = 0 - \Pr(H_p)$. Thus, the maximal and minimal values of the measure D depend on H_p and fail the property of logicality.
3. When E implies H_p, $R = 1/\Pr(H_p)$. When E implies H_d, $\Pr(H_p \mid E) = 0$, so $R = 0/\Pr(H_p) = 0$. Thus, the minimal value of the measure R is independent of H_p and E, but the maximal value is not. Again, measure R fails the property of logicality.

Probabilistic reasoning also enables definitions of the possible interactions between the evidence transferred in each direction (criminal to scene, scene to criminal) and hence eases the interpretation of such evidence. Given that the decision to collect new evidence has to be made before the information is available, another problem is the calculation of the expected gain of this new (unknown) information, so that the gain can be compared with the cost of the search.

Provided that the utilities, or losses, of the outcomes of decisions may be quantified in such a way that they can be compared with the cost of the experiment, Bayesian decision theory explains how the expected value of information may be calculated by taking advantage of the recommendation to maximize the expected utility.

4. LEGAL REASONING AND PROBABILISTIC REASONING

According to the preface to the first edition of Robertson et al. (2016) (Robertson & Vignaux 1995, p. xi), "the examination of the applicability of logical and probabilistic reasoning to evidence generally. . .has been the subject of vigorous discussion in the legal literature and is one of the main threads of the 'New Evidence Scholarship,'" a term coined by Lempert (1986). This discussion continues to the present day. The preface to the second edition (Robertson et al. 2016, p. xv) notes that there is "little sign of great increase in understanding in the legal profession or academia." Some have suggested that probabilistic reasoning is ill-suited to many aspects of the administration of civil and criminal justice (see, e.g., Allen & Pardo 2019, 2023). Common criticisms are:

- the use of numbers for the comparison of the value of evidence with the standard of proof,
- the incompatibility of probabilistic reasoning and the way in which factfinders evaluate and reason with evidence, and
- the conjunction problem,

which are discussed in turn.

4.1. The Use of Numbers for the Comparison of the Value of Evidence with the Standard of Proof

Factfinders are presented with various explanations for the contested events, and, individually, they compare these explanations, informed by both the evidence and their own background knowledge. The factfinder has to make a decision that is, in turn, informed by the standard of proof applicable in the case at hand. Probabilistic reasoning enlists probability as a measure of uncertainty to aid the comparison. Critics of probabilistic reasoning argue that it is not possible to assign numbers to the degree of uncertainty; supporters of probabilistic reasoning give the example of the use of hypothetical bets by which a person's subjective belief about evidence or an event may be elicited as a probability (see, e.g., de Finetti 1940, Edwards et al. 1963, Lindley 2014). The probabilities or subjective beliefs for the various explanations may then be compared. For consistency, the explanation with the highest probability is the one which forms the basis for any decision to be made. As a person has a subjective belief about evidence or an event that can be elicited as a probability, so they may have a subjective belief about the standard of proof. For each threshold of, for example, (*a*) belief beyond reasonable doubt, (*b*) clear and convincing evidence, or (*c*) the preponderance of the evidence, the person may assign a probabilistic threshold. Their beliefs about the evidence and events may then be compared with the appropriate threshold.

4.2. The Incompatibility of Probabilistic Reasoning and the Way in Which Factfinders Evaluate and Reason with Evidence

In a court case, evidence is presented sequentially. Before the case begins, the factfinder will have some prior subjective belief about the veracity of the case. In a civil case, this could be that the plaintiff's case and the defendant's case are equally likely to be true: Each has a probability of 0.5 of being true. In a criminal case, this could be that the defendant is as likely to be innocent as any other member of some relevant population (a population that could initially be that of the world but which would be rapidly and drastically reduced through the progression of the consideration of the evidence). Evidence is led sequentially. The posterior odds in favor of the prosecution/plaintiff's proposition, H_p say, compared with the defendant's proposition H_d after two pieces of evidence E_1 and E_2 are led, are then given by Equation 2.

Of course, as more evidence is led, the updating procedure becomes more complex. However, the complexity is independent of the mode of reasoning by the factfinder. Probabilistic reasoning is a more transparent approach to the updating than verbal reasoning. An interesting comment on this updating procedure is given by Lindley (1991). At the beginning of a trial, the members of the jury will, almost certainly, have different beliefs—perhaps as many different beliefs as there are members of the jury:

> Suppose on knowledge *K*, two people have different beliefs in the truth of an event *G*: their probabilities are not the same. Suppose, now, additional evidence *E* relevant to *G* is produced. Then it can be shown rather generally that *E* will tend to bring the two probabilities closer together, and that, for sufficiently large amounts of evidence, they will agree for all practical purposes. Briefly, additional evidence makes for agreement in beliefs. (Lindley 1991, p. 49).

There is a clear analogy with the reasoning process in a jury room. Lindley (1991, p. 49) further notes that "[t]here is nothing to force agreement, but experience shows that agreement is usually reached."

BNs, as illustrated in **Figures 1** and **2**, may be presented as a summary of the evidence, though they are usually more complex than these illustrations. These may be thought of as a development of Wigmore charts (Wigmore 1913). A review of Wigmore charts and Bayes nets is provided by Roberts & Aitken (2014).

4.3. The Conjunction Problem

Consider two elements A and B that are parts of a plaintiff's case. Both need to be "proven" on the balance of probabilities to satisfy the burden of proof. There is one item of evidence, denoted E, and A and B are deemed independent, given E. The evidence may be thought to support both elements if it can be shown that $\Pr(A \mid E)$ and $\Pr(B \mid E)$ are both greater than 0.5. Assume $\Pr(A \mid E)$ and $\Pr(B \mid E)$ are both equal to 0.6; both prove the plaintiff's case on the balance of probabilities. However, A and B, given E, are independent, and $\Pr(A, B \mid E) = \Pr(A \mid E) \times \Pr(B \mid E) = 0.36$; the case would fail on the balance of probabilities. Examples such as this one have been raised [e.g., by Cohen (1977) and Allen & Pardo (2019)] to show that probabilistic reasoning is incompatible with legal reasoning.

An example is given by Cohen (1977, p. 59):

> Perhaps a car driver is suing his insurance company because it refuses to compensate him after an accident. Suppose the two component issues that are disputed are first, what were the circumstances of the crash, and secondly, what were the terms of the driver's insurance contract. Then, if each of these two issues is determined with a probability of 0.71, their joint outcome can be determined with a sufficiently high probability, since 0.71^2 is greater than 0.501. But if one of the component issues is determined with only a 0.501 probability, then the other component issue must be determined with a probability of very nearly 1. Otherwise the product of the two probabilities would not be high enough to satisfy the requirements of justice. Or, in other words, if one of the component issues is determined on the balance of probability (whether this balance be understood to lie at 0.501, 0.51, 0.6 or even a higher figure), the other must, in effect, be determined beyond reasonable doubt.

However, further analysis (see, e.g., Nesson 1985, Aitken et al. 2022) reveals that the plaintiff's case (argument) has a higher probability than any other case (argument). Denote the complement (negation) of A as $\bar{A}$ and the complement of B as $\bar{B}$ such that $\Pr(A \mid E) = 0.6 \Rightarrow \Pr(\bar{A} \mid E) = 0.4$ and $\Pr(B \mid E) = 0.6 \Rightarrow \Pr(\bar{B} \mid E) = 0.4$. Then, we have

- $\Pr(A, B \mid E) = \Pr(A \mid E) \times \Pr(B \mid E) = 0.36$,
- $\Pr(A, \bar{B} \mid E) = \Pr(A \mid E) \times \Pr(\bar{B} \mid E) = 0.24$,
- $\Pr(\bar{A}, B \mid E) = \Pr(\bar{A} \mid E) \times \Pr(B \mid E) = 0.24$, and
- $\Pr(\bar{A}, \bar{B} \mid E) = \Pr(\bar{A} \mid E) \times \Pr(\bar{B} \mid E) = 0.16$.

The case with the highest probability, conditional on E, is $\{A, B\}$. A holistic evaluation would assign a probability less than 0.5 to the joint occurrence of these two events. However, the joint occurrence of A and B has a higher probability than the joint occurrence of any other combination of the events or their negations. Whatever argument is put forward by the defendant will have a lower probability than that of the plaintiff. It is a matter for the lawyers as to which argument to support. This raises the question of whether the argument of the plaintiff has to be of higher probability than each individual alternative or of all alternatives considered as one, but this is a legal argument, not a failure of probability.

A similar approach to this conjunction problem is illustrated by what is known as Linda's example (Sides et al. 2002). Here, the evidence E is the following information about Linda: "she is 31 years old, single, outspoken, and very bright; she majored in philosophy and, as a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations" (p. 191).

The elements are:

- A: Linda is a bank teller;
- B: Linda is active in the feminist movement; and
- (A, B): Linda is a bank teller and is active in the feminist movement.

A majority of respondents across a variety of studies ranked the conjunction of A and B as more probable than A. Hertwig & Chase (1998) provide a review of findings; the original report is by Tversky & Kahneman (1983). This ranking is in contradiction to the laws of probability that state that, for two events A and B, $\Pr(A, B) = \Pr(B \mid A)\Pr(A) \leq \Pr(A)$, since $\Pr(B \mid A) \leq 1$.

This paradoxical finding can be understood by considering relative values of unconditional and conditional probabilities. Consider a definition that evidence E favors (A, B) more than it favors A if and only if

$$\frac{\Pr(A, B \mid E)}{\Pr(A, B)} > \frac{\Pr(A \mid E)}{\Pr(A)}.$$

It can be shown that this inequality holds if and only if

$$\Pr(E \mid A, B) > \Pr(E \mid A).$$

This inequality asserts that Linda is more likely to be single, outspoken, etc. on the assumption that she is a feminist bank teller than on the assumption that she is a bank teller. This is a reasonable assertion.

5. THE PRESENTATION OF EVIDENCE: SINGLE NUMBERS AND INTERVALS

The measurement and reporting of the value of evidence have been widely debated in the forensic literature. Following Taroni et al. (2016), many articles have been published on these topics, including those in a special issue of the journal *Science & Justice* (Morrison 2016). While the use of a Bayes factor to assess the probative value of evidence is now supported by recommended standards in different forensic disciplines (ENFSI 2015), there is not a widespread consensus on how conclusions regarding the strength of the evidence should be reported to a court. On the one hand, there is a school of thought that a forensic expert should report a single value for a Bayes factor (e.g., Berger & Slooten 2016, Ommen et al. 2016, Taroni et al. 2016). On the other hand, a different school of thought fears that reporting a single value would deprive the legal system of essential information needed to assess the reliability of the evidence and advocates that a reported value should be accompanied by an interval that would allow account to be taken of inherent uncertainties characterizing the assessment of the evidence in order to maximize scientific objectivity and avoid personal opinions (e.g., Morrison & Enzinger 2016, Sjerps et al. 2016).

The Bayes factor is often expressed as a ratio between two conditional probabilities, and the task of the expert is focused on the assessment of these probabilities. An important issue that is often debated concerns the interpretation of probabilities as personal expressions of belief held by an individual and the reasonableness of such an interpretation in applications in forensic science. Subjective probabilities are often thought to be an arbitrary expression of an individual's belief, with the consequence that such assessments are labeled as unfounded guesses (Morrison & Enzinger 2016) or as bare assertions of belief (Martire et al. 2017). Such arguments are, however, unsound. Subjective probabilities are conditioned on all task-relevant information; subjectivism is not unconstrained, and it is the duty of the forensic scientist to express their probability (measure of belief) responsibly. Different experts may present different values, as the knowledge bases of different people on which assessments of the truth or otherwise of a given proposition are based may differ. Such differences reflect the capacity of the Bayesian framework to account for differences in personal knowledge bases between individuals. Any claimed accuracy of the reported value (Morrison 2016) is unlikely be achieved.

The Bayes factor, like probability, is not a quantity that could be well approximated if only there were sufficient data. It is, rather, a construction of the human mind for reasoning under uncertainty. It can be argued that objective probabilities do not exist, but are, at best, assignments of probability on which several individuals may find agreement (Biedermann et al. 2017). Leaving aside philosophical matters concerning the soundness of a personal interpretation of probability, the objection is often raised (see, e.g., Kafadar 2015) that such a personal interpretation does not match the daily case work of a forensic scientist, which often relies on the combination of data on the occurrence of target features, summarized in terms of relative frequencies. However, nothing in the Bayesian paradigm prevents the use of relative frequencies, whenever available information may be expressed in such a form, from a contribution to the process of the assignment of subjective probabilities, a process known as elicitation. This is not only acknowledged as reasonable but is the subject of the representation theorem of de Finetti, as discussed by Taroni et al. (2018). The assumptions that underlie the assignment of a probability are always open to question. Different assumptions may lead to different assignments and, perhaps, different interpretations. These potential differences have led to the suggestion that the provision of a lower and an upper bound for the assigned probability may better reflect uncertainty arising from these differences than reliance on a single value. Assignment of an interval for a probability is problematic, however, as it provides no guidance to the recipients of such expert information on how the pair of values represented by the lower and upper bounds of the interval may be used (Biedermann et al. 2016).

5.1. Sources of Uncertainty

The assessment of a value for a Bayes factor can indeed be a challenging task that can be subjected to many sources of uncertainty in addition to the elicitation of probabilities. These include the formulation of hypotheses, the model choice, the selection of the control samples and recovered items to be analyzed, the elicitation of prior probability distributions, or the computational impasses that may be encountered. In some cases, the forensic results may be seen as the outcome of a process about which there is enough knowledge to formulate a probabilistic model, and it will be a matter of judgement by the expert as to whether this is adequate. Different experts may provide different models, even starting from the same data, and it is desirable that information be provided to help recipients of the models' output understand how these experts have reached their conclusions. Since the expression of a Bayes factor often involves model parameters (say, θ) that are unknown, the existence of a true value of the Bayes factor that can be estimated, and thus accompanied with some interval, has been widely debated. However, that premise is unsound, as there is no true value that can be estimated. Take the case where an analytical feature F is observed, and the probability $\Pr(F \mid \theta)$ for an unknown individual to be associated with this analytical feature is to be assessed. Personal beliefs concerning the analytical feature for the unknown individual can be formulated as $\Pr(F) = \int \Pr(F \mid \theta) f(\theta) \mathrm{d}\theta$, where $f(\theta)$ describes the available knowledge about θ. Clearly, a change in the available knowledge about θ will lead to a different $f(\theta)$.

A further complication originates from the fact that in some situations, the marginal likelihoods are unavailable in closed form. The error that may result from the implementation of numerical techniques is an important source of information about which the scientist should be transparent. Following the ideas of Tanner (1996), reconsidered by Ommen et al. (2017) in a forensic context, numerical precision can be estimated by an associated Monte Carlo standard error. With reference to the computational impasses that can make an analytical solution unachievable, some scientists prefer to adopt a so-called Bayesian-likelihood approach, according to which the likelihoods in the numerator and in the denominator are considered to be functions of the parameter θ (e.g., van den Hout & Alberink 2016). Though such a proposal has an apparent appeal, objections can be raised either from a philosophical point of view, as its compatibility with the Bayesian perspective can be questioned (see Gelman et al. 2014), or from a practical point of view, as it must be decided which distribution for θ should be used for the numerator and which for the denominator. A forensic scientist may alternatively choose a frequentist approach. However, a difficulty with such an approach, in which a parametric estimate $\hat{\theta}$ is plugged into the Bayes factor (which in this case is a likelihood ratio), is the decision as to which estimate of θ to use for the numerator and which for the denominator, and there is not necessarily a unique answer (Dawid 2017). The Bayesian approach requires simply the calculation of a Bayes factor. Uncertainty about a population parameter θ is taken into account in its computation; this is the integration of the available information into the evaluation of evidence to provide the best assessment of its value.

It is advisable and insightful to consider different values for input parameters and investigate the impact on the Bayes factor, as well as that of different data sets. This would amount to the determination of a distribution for the Bayes factor, independent of observations made in a given case, that can be informative about the sensitivity of the Bayes factor to variation in the input parameters. Marginal distributions can be highly sensitive to the choice of the prior distribution, and situations characterized by an abundance of information for prior elicitation purposes are rare. Such information is obtainable before findings are made, and it informs about the potential of misleading evidence (Aitken et al. 2021). Conversely, when deciding whether to allocate more resources to obtain more information, it is relevant whether or not more information would lead to a different Bayes factor. The collection of more data represents a potentially key aspect of a criminal investigation, but a study of whether or not to gather more data to assist further in the determination of the input parameters is a separate issue that should not distract from the evidential interpretation of the current value of the Bayes factor (Berger & Slooten 2016, Taroni et al. 2016, Taylor et al. 2016, Meester & Slooten 2020).

5.2. Combination of Values

Another argument against an interval-based formulation of the Bayes factor is that even if the argument were accepted that an interval should be determined, it is not clear how such an interval should act as a multiplication factor in the odds-form of Bayes' theorem, and therefore how it should be used by the trier of fact for the purpose of making a decision (Berger & Slooten 2016). The combination of supports for different pieces of evidence with the Bayes factor is multiplicative. It is difficult to envisage how interval supports for different pieces of evidence can be combined meaningfully (Biedermann et al. 2016). The choice of various endpoints may result in an incoherent or biased process (Ommen et al. 2016).

The discovery of a transparent and logical explanation of the different forms of evidence, such as scientific, testimonial, and circumstantial, relied upon by forensic experts and the use to which these different forms may be put for the elicitation of probabilities are major challenges for the use of probabilistic reasoning in the administration of justice. If a single value of the Bayes factor is thought to represent an incomplete expression of the value of the evidence, then all possible sources of uncertainty should be considered. However, an interval would be inadequate, and a multidimensional representation would need to be chosen. Much work is still required concerning all the important aspects raised in this section and the communication of results so that the best representation of the value of the evidence under consideration can be achieved.

6. CALIBRATION

There is a mathematical result that the likelihood ratio of the likelihood ratio is the likelihood ratio:

$$\frac{\Pr(\mathrm{LR} \mid H_{\mathrm{p}})}{\Pr(\mathrm{LR} \mid H_{\mathrm{d}})} = \mathrm{LR} \qquad 3.$$

(van Leeuwen & Brümmer 2013, Aitken et al. 2021), for which an outline proof is given in the Appendix.

A criticism of some approaches to the evaluation of evidence is that their results do not satisfy this result. This result is given by van Leeuwen & Brümmer (2013) as a definition of calibration. Methods of obtaining Bayes factors that do not satisfy Equation 3 are said to be poorly calibrated. Critics, for example, Vergeer et al. (2020) [written in response to Aitken et al. (2019), with a reply by Aitken et al. (2020)], use the likelihood ratio as a score. They argue that "LR-values coming from assumed statistical model families. . .often cannot be interpreted as such and require a so-called posthoc calibrating step" (p. 1). Consider the standard approach for the statistical evaluation of evidence. Statistical models are developed based on training data. Likelihood ratios follow. Their performance is assessed, ideally with validation data or, failing their availability, a cross-validation analysis of the training data. These likelihood ratios are likelihood ratios by definition. The post hoc calibrating step does not then produce a likelihood ratio, in the sense that evaluation of evidence defines a likelihood ratio. The resultant statistic is not the ratio of the probabilities (loosely defined) of the evidence given the prosecution proposition and the evidence given the defense proposition. The original likelihood ratio can be defined as a score, as Vergeer et al. (2020) suggest, but this is an artificial construct based on a post hoc desire to obtain a better result. Once a likelihood ratio is determined, it cannot be adjusted in the manner suggested by calibration, namely to determine its likelihood ratio.

The results from a model that is not well calibrated cannot be adjusted to obtain something that is well calibrated. This would be an adjustment made after the analysis to obtain a result that looks better in some sense. The correct response to poor calibration is to look for a better model. Consider the weather forecaster whose 90% predictions tend to be right only 70% of the time. The correct response is not an automatic adjustment of 90% to 70%. It is to obtain a better model for forecasting.

Calibration is a measure for the assessment of performance of a model. It is not a method for the evaluation of evidence. Some further comments about the role of calibration in the evaluation and interpretation of evidence are given by Aitken et al. (2021).

7. CONCLUSION

The role of the Bayes factor as the factor that converts prior odds in favor of a proposition to posterior odds in favor of the proposition after consideration of evidence is well known and is intuitively very attractive. Historically, it has been shown that, with certain reasonable assumptions, the Bayes factor is the best way to evaluate evidence. More generally, the probabilistic line of reasoning in the administration of criminal justice has other attractive features in addition to that provided by the Bayes factor. It satisfies a number of necessary conditions on normative logical adequacy. It may also be used to assess, with versatile applicability, evidential phenomena for the combination of evidence and in the presence of a mass of evidence, which gives rise to complex reasoning patterns.

The role of the Bayes factor in the evaluation of evidence has been challenged in recent years. For example, the role of probability (in the subjective paradigm) in legal reasoning as the best measure of uncertainty has been disputed in a discussion of the concept of relative plausibility. It has been shown that careful use of probabilistic reasoning counters the criticisms of the proponents of relative plausibility.

The benefits to the administration of criminal justice system of the roles of probabilistic reasoning in general and the Bayes factor in particular have been reviewed. Other proposals for the evaluation of evidence have been discussed, and it has been shown that appropriate use of the Bayes factor is able to counter the criticisms by these other proposals of its use.

APPENDIX

A speaker recognition system has as input two speech segments, denoted X and Y. Let $s = f(X, Y)$ be a single, scalar score. The likelihood ratio, here denoted r for consistency with van Leeuwen & Brümmer (2013), is a function of s:

$$r = \frac{\Pr(s \mid H_\mathrm{p}, M)}{\Pr(s \mid H_\mathrm{d}, M)}, \qquad 4.$$

where H_p is the proposition that X and Y originate from the same speaker, H_d is the proposition that X and Y are from different speakers, and M is a probabilistic model for s.

Let $\Pr(H_\mathrm{p} \mid M) = \pi$. Then, we have

$$\begin{aligned}\Pr(H_\mathrm{p} \mid s, M, \pi) &= \frac{\Pr(s \mid H_\mathrm{p}, M, \pi)\pi}{\Pr(s \mid H_\mathrm{p}, M, \pi)\pi + \Pr(s \mid H_\mathrm{d}, M, \pi)(1-\pi)}\\ &= \frac{r\pi}{r\pi + (1-\pi)},\end{aligned}$$

and a similar argument holds for H_d with

$$\Pr(H_\mathrm{d} \mid s, M, \pi) = \frac{(1-\pi)}{r\pi + (1-\pi)}.$$

Thus, the posterior probability may be written as

$$\Pr(h \mid s, M, \pi) = \Pr(h \mid r, M', \pi), \; h \in \{H_\mathrm{p}, H_\mathrm{d}\}, \qquad 5.$$

where M' has been introduced to denote M, augmented with Equation 4. Then, we have

$$\frac{\Pr(H_p \mid s, M, \pi)}{\Pr(H_d \mid s, M, \pi)} = \frac{\pi}{(1-\pi)} \frac{\Pr(s \mid H_p, M)}{\Pr(s \mid H_d, M)} = \frac{\pi}{(1-\pi)} r;$$

$$\begin{aligned}
\frac{\Pr(H_p \mid r, M', \pi)}{\Pr(H_d \mid r, M', \pi)} &= \frac{\pi}{(1-\pi)} \frac{\Pr(r \mid H_p, M')}{\Pr(r \mid H_d, M')} \\
\Rightarrow \frac{\pi}{(1-\pi)} r &= \frac{\Pr(H_p \mid s, M, \pi)}{\Pr(H_d \mid s, M, \pi)} \\
&= \frac{\Pr(H_p \mid r, M', \pi)}{\Pr(H_d \mid r, M', \pi)} \quad \text{(from Equation 5)} \\
&= \frac{\pi}{(1-\pi)} \frac{\Pr(r \mid H_p, M')}{\Pr(r \mid H_d, M')} \\
\Rightarrow r &= \frac{\Pr(r \mid H_p, M')}{\Pr(r \mid H_d, M')}.
\end{aligned}$$

SUMMARY POINTS

1. Probability has a central role in the measurement of uncertainty.
2. The Bayes factor is a coherent and logical measure to assess the value of evidence; it therefore plays an important role in the evaluation and interpretation processes.
3. Simplification of complexity and consideration of masses of evidence are much aided with the use of BNs.
4. Subjective probability can be helpful in a discussion about thresholds for the standard of proof.
5. Probabilistic reasoning can be presented in a way that is compatible with how factfinders evaluate and reason with evidence.
6. A resolution of the conjunction problem is achieved with consideration of relative probabilities rather than absolute probabilities.
7. Evidence is best evaluated and interpreted with a single number rather than an interval of numbers.
8. Calibration is an assessment of the performance of a statistical procedure for evaluation, not a part of the procedure.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank the Swiss National Science Foundation for its support through grant number 100011_204554/1 (The Anatomy of Forensic Inference and Decision) and an anonymous reviewer for very helpful comments.

LITERATURE CITED

Aitken CGG, Chang YT, Buzzini P, Zadora G, Massonnet G. 2019. The evaluation of evidence for microspectrophotometry data using functional data analysis. *Forensic Sci. Int.* 305:110007

Aitken CGG, Chang YT, Buzzini P, Zadora G, Massonnet G. 2020. Letter to the editor in response to Vergeer et al. (2020) FSI 314 (2020) 110388. *Forensic Sci. Int.* 316:110478

Aitken CGG, Gammerman A. 1989. Probabilistic reasoning in evidential assessment. *J. Forensic Sci. Soc.* 29:303–16

Aitken CGG, Lucy D. 2004. Evaluation of trace evidence in the form of multivariate data. *Appl. Stat.* 53:109–22

Aitken CGG, Taroni F, Bozza S. 2021. *Statistics and the Evaluation of Evidence for Forensic Scientists*. Chichester, UK: John Wiley & Sons. 3rd ed.

A review of the role of the Bayes factor for evaluation and interpretation of evidence.

Aitken CGG, Taroni F, Bozza S. 2022. Evidence, probability and relative plausibility. *Int. J. Evid. Proof* 26(4):309–24

Allen RJ, Pardo MS. 2019. Relative plausibility and its critics. *Int. J. Evid. Proof* 23(1/2):5–59

Allen RJ, Pardo MS. 2023. Evidence, probability, and relative plausibility: a response to Aitken, Taroni, and Bozza. *Int. J. Evid. Proof* 27(2):126–42

Ayer AJ. 1957. The conception of probability as a logical relation. In *Observation and Interpretation*, ed. S Körner, pp. 12–30. London: Butterworths

Barrett M, Sober E. 2020. The requirement of total evidence: a reply to Epstein's critique. *Philos. Sci.* 87:191–203

Berger CEH, Slooten K. 2016. The LR does not exist. *Sci. Justice* 56:388–91

Biedermann A, Bozza S, Taroni F. 2009. Probabilistic evidential assessment of gunshot residue particle evidence (part I): likelihood ratio calculation and case pre-assessment using Bayesian networks. *Forensic Sci. Int.* 191:24–35

Biedermann A, Bozza S, Taroni F. 2011. Probabilistic evidential assessment of gunshot residue particle evidence (part II): Bayesian parameter estimation for experimental count data. *Forensic Sci. Int.* 206:103–10

Biedermann A, Bozza S, Taroni F, Aitken CGG. 2016. Reframing the debate: a question of probability, not of likelihood ratio. *Sci. Justice* 56:392–96

Biedermann A, Bozza S, Taroni F, Aitken CGG. 2017. The consequences of understanding expert probability reporting as a decision. *Sci. Justice* 57:80–85

Bozza S, Taroni F, Biedermann A. 2022. *Bayes Factors for Forensic Decision Analyses with R*. New York: Springer

A review of Bayesian decision analysis in forensic science with *R* code.

Bozza S, Taroni F, Marquis R, Schmittbuhl M. 2008. Probabilistic evaluation of handwriting evidence: likelihood ratio for authorship. *Appl. Stat.* 57:329–41

Buckleton JS, Robertson B, Curran J, Berger C, Taylor D, et al. 2020. A review of likelihood ratios in forensic science based on a critique of Stiffelman "No longer the gold standard: probabilistic genotyping is changing the nature of DNA evidence in criminal trials." *Forensic Sci. Int.* 310:110251

Carnap R. 1947. On the application of inductive logic. *Philos. Phenomenol. Res.* 8:133–48

Cohen LJ. 1977. *The Probable and the Provable*. Oxford, UK: Clarendon

Crupi V, Tentori K. 2014. State of the field: measuring information and confirmation. *Stud. Hist. Philos. Sci.* 47:81–90

Crupi V, Tentori K. 2016. Confirmation theory. In *Oxford Handbook of Philosophy and Probability*, ed. A Hajek, C Hitchcock, pp. 650–65. Oxford, UK: Oxford Univ. Press

Dawid AP. 2017. Forensic likelihood ratio: statistical problems and pitfalls. *Sci. Justice* 57:73–75

Dawid AP, Mortera J. 2021. Bayesian networks in forensic science. In *Handbook of Forensic Statistics*, ed. D Banks, K Kafadar, DH Kaye, M Tackett, pp. 165–97. Boca Raton, FL: Chapman & Hall/CRC

de Finetti B. 1930. Fondamenti logici del ragionamento probabilistico. *Boll. Unione Matem. Ital.* 9:258–61

de Finetti B. 1931. Sul significato soggettivo delle probabilitià. *Fundam. Math.* 17:298–329
de Finetti B. 1940. La teoria delle probabiliti nei suoi rapporti con i'analisi. In *Relazioni della XXVIII Riunione della Società Italiana per il Progresso delle Scienze*, pp. 27–35. Rome: Soc. Ital. Prog. Sci.
de Finetti B. 1989. La probabilità: Guardarsi dalle contraffazioni! In *La Logica dell'Incerto*, ed. M Mondadori, pp. 149–88. Milan: Saggiatore
Edwards W, Lindman H, Savage LJ. 1963. Bayesian statistical inference for psychological research. *Psychol. Rev.* 70:193–242
Eells E. 2000. Review: *The Foundations of Causal Decision Theory*, by James M. Joyce. *Br. J. Philos. Sci.* 51:893–900
Eells E, Fitelson B. 2002. Symmetries and asymmetries in evidential support. *Philos. Stud.* 107:129–42
ENFSI (Eur. Netw. Forensic Sci. Inst.). 2015. *ENFSI guideline for evaluative reporting in forensic science.* Rep., ENFSI, Wiesbaden, Ger. **https://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf**
Evett IW. 1996. Expert evidence and forensic misconceptions of the nature of exact science. *Sci. Justice* 36:118–22
Evett IW, Weir BS. 1998. *Interpreting DNA Evidence*. Sunderland, MA: Sinauer
Fienberg SE. 1986. Gatecrashers, blue buses, and the Bayesian representation of legal evidence. *Boston Univ. Law Rev.* 66:693–99
Fienberg SE, Schervish MJ. 1986. The relevance of Bayesian inference for the presentation of statistical evidence and for legal decision making. *Boston Univ. Law Rev.* 66:771–98
Finkelstein MO, Fairley WB. 1970. A Bayesian approach to identification evidence. *Harvard Law Rev.* 83:489–517
Fitelson B. 1999. The plurality of Bayesian measures of confirmation and the problem of measure sensitivity. *Philos. Sci.* 66:362–78
Fitelson B. 2006. Logical foundations of evidential support. *Philos. Sci.* 73:500–12
Fitelson B. 2011. Favoring, likelihoodism, and Bayesianism. *Philos. Phenomenol. Res.* 83:666–71
Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2014. *Bayesian Data Analysis*. Boca Raton, FL: CRC. 3rd ed.
Gittelson S. 2013. *Evolving from inferences to decisions in forensic science.* PhD Thesis, Sch. Crim. Justice, Univ. Lausanne, Lausanne, Switz.
Gittelson S, Bozza S, Biedermann A, Taroni F. 2013. Decision-theoretic reflections on processing a fingermark. *Forensic Sci. Int.* 226:e42–47
Good IJ. 1950. *Probability and the Weighing of Evidence*. London: Griffin
Good IJ. 1967. On the principle of total evidence. *Br. J. Philos. Sci.* 17:319–21
Good IJ. 1979. Studies in the history of probability and statistics. XXXVII. A.M. Turing's statistical work in World War II. *Biometrika* 66:393–96
Good IJ. 1985. Weight of evidence: a brief survey (with discussion). In *Bayesian Statistics 2*, ed. JM Bernardo, MH DeGroot, DV Lindley, AFM Smith, pp. 249–70. Amsterdam: North-Holland
Good IJ. 1989. Weight of evidence and a compelling metaprinciple. *J. Stat. Comput. Simul.* 31:121–23
Good IJ. 1991. Weight of evidence and the Bayesian likelihood ratio. In *The Use of Statistics in Forensic Science*, ed. CGG Aitken, DA Stoney, pp. 85–106. Chichester, UK: Ellis Horwood
Hertwig R, Chase VM. 1998. Many reasons or just one: how response mode affects reasoning in the conjunction problem. *Think. Reason.* 4:319–52
Howson C, Urbach P. 1993. *Scientific Reasoning: The Bayesian Approach*. Chicago: Open Court. 2nd ed.
Jackson G. 2000. The scientist and the scales of justice. *Sci. Justice* 40:81–85
Jeffrey RC. 1992. *Probability and the Art of Judgment*. Cambridge, UK: Cambridge Univ. Press
Jeffrey RC, Hendrickson M. 1989. Probabilizing pathology. *Proc. Aristot. Soc.* 89:211–25
Jeffreys H. 1983. *Theory of Probability*. Oxford, UK: Clarendon. 3rd ed.
Joyce H. 2005. Career story: Consultant forensic statistician: Ian Evett. *Significance* 2:34–37
Kafadar K. 2015. Statistical issues in assessing forensic evidence. *Int. Stat. Rev.* 83(1):111–34
Kaye DH. 1979. Probability theory meets *res ipsa loquitur*. *Mich. Law Rev.* 77:1456–84
Kaye DH. 1986. Do we need a calculus of weight to understand proof beyond a reasonable doubt? *Boston Univ. Law Rev.* 66:657–72
Kaye DH, Sensabaugh G. 2011. Reference guide on DNA identification evidence. In *Reference Manual on Scientific Evidence*, pp. 129–210. Washington, DC: Natl. Acad. Press. 3rd ed.

Lauritzen SL, Spiegelhalter DJ. 1988. Local computations with probabilities on graphical structures and their application to expert systems. *J. R. Stat. Soc. B* 50:157–224

Lempert RO. 1977. Modeling relevance. *Mich. Law Rev.* 75:1021–57

Lempert RO. 1986. The new evidence scholarship—analysing the process of proof. *Boston Univ. Law Rev.* 66:439–77

Lindley DV. 1977. A problem in forensic science. *Biometrika* 64(2):207–13

Lindley DV. 1985. *Making Decisions*. Chichester, UK: John Wiley & Sons. 2nd ed.

Lindley DV. 1991. Probability. In *The Use of Statistics in Forensic Science*, ed. CGG Aitken, DA Stoney, pp. 27–50. New York: Ellis Horwood

Lindley DV. 2014. *Understanding Uncertainty*. Chichester, UK: John Wiley & Sons. 2nd ed.

Locard E. 1940. *Traité de criminalistique. L'enquête criminelle*. Lyon, Fr.: Desvigne

Maher P. 1996. Probability captures the logic of scientific confirmation. In *Contemporary Debates in Philosophy of Science*, ed. CR Hitchcock, pp. 69–93. Oxford, UK: Blackwell

Martire KA, Edmond G, Navarro DJ, Newell BR. 2017. On the likelihood of "encapsulating all uncertainty." *Sci. Justice* 57:76–79

Meester R, Slooten K. 2020. An epistemic interpretation of the posterior likelihood ratio distribution. *Law Probab. Risk* 19:139–55

Morrison GS. 2016. Special issue on measuring and reporting the precision of forensic likelihood ratios: introduction to the debate. *Sci. Justice* 56:371–73

Morrison GS, Enzinger E. 2016. What should a forensic practitioner's likelihood ratio be? *Sci. Justice* 56:374–79

Nesson C. 1985. The evidence or the event? On judicial proof and the acceptability of verdicts. *Harvard Law Rev.* 98:1357–92

Nordgaard A, Ansell R, Drotz W, Jaeger L. 2012. Scale of conclusions for the value of evidence. *Law Probab. Risk* 11:1–24

Ommen DM, Saunders CP, Neumann C. 2016. An argument against presenting interval quantifications as a surrogate for the value of evidence. *Sci. Justice* 56:383–87

Ommen DM, Saunders CP, Neumann C. 2017. The characterization of Monte Carlo errors for the quantification of the value of forensic evidence. *J. Stat. Comput. Simul.* 87:1608–43

Raiffa H, Schlaifer R. 1961. *Applied Statistical Decision Theory*. Cambridge, MA: MIT Press

Roberts P, Aitken CGG. 2014. *The logic of forensic proof.* Pract. Guide 3, R. Stat. Soc. Work. Group Stat. Law, R. Stat. Soc., London

A discursive review of charting (Wigmore and Bayes nets) for the consideration of evidence.

Robertson B, Vignaux GA. 1993. Taking fact analysis seriously. *Mich. Law Rev.* 91:1442–64

Robertson B, Vignaux GA. 1995. *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. Chichester, UK: John Wiley & Sons

Robertson B, Vignaux GA. 1998. Explaining evidence logically. *New Law J.* 148:159–62

Robertson B, Vignaux GA, Berger CEH. 2016. *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. Chichester, UK: John Wiley & Sons. 2nd ed.

Salmon WC. 1967. *The Foundations of Scientific Inference*. Pittsburgh, PA: Univ. Pittsburgh Press

Schum DA. 1994. *Evidential Foundations of Probabilistic Reasoning*. Chichester, UK: John Wiley & Sons

Sides A, Osheron D, Bonini N, Viale R. 2002. On the reality of the conjunction fallacy. *Mem. Cogn.* 30:191–98

Sjerps M, Alberink I, Bolck A, Stoel RD, Vergeer P, van Zanten JH. 2016. Uncertainty and LR: to integrate or not to integrate, that's the question. *Law Probab. Risk* 15:23–29

Stiffelman B. 2019. No longer the gold standard: probabilistic genotyping is changing the nature of DNA evidence in criminal trials. *Berkeley J. Crim. Law* 2:110–46

Tanner MA. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. New York: Springer. 3rd ed.

Taroni F, Biedermann A, Bozza S, Garbolino P, Aitken CGG. 2014. *Bayesian Networks for Probabilistic Inference and Decision Analysis in Forensic Science*. Chichester, UK: John Wiley & Sons. 2nd ed.

A more technical review than that by Roberts & Aitken (2014) of Bayes nets in forensic science.

Taroni F, Bozza S, Biedermann A, Aitken CGG. 2016. Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law Probab. Risk* 15:1–16

Taroni F, Bozza S, Biedermann A, Garbolino P, Aitken CGG. 2010. *Data Analysis in Forensic Science: A Bayesian Decision Perspective*. Chichester, UK: John Wiley & Sons

A review of statistical decision theory with applications in forensic science.

Taroni F, De March I, Garbolino P, Bozza S. 2018. Prova genetica del DNA e risultati dissonanti: come valutare congiuntamente gli elementi scientifici di prova. *Dirit. Penale Contemp.* 11:77–94

Taroni F, Garbolino P, Bozza S. 2020. Coherently updating degrees of belief: radical probabilism, the generalization of Bayes' theorem and its consequences on evidence evaluation. *Law Probab. Risk* 19:293–316

Taroni F, Garbolino P, Bozza S, Aitken CGG. 2021. The Bayes factor: the coherent measure for hypothesis confirmation. *Law Probab. Risk* 20:15–36

Taylor D, Hicks T, Champod C. 2016. Using sensitivity analyses in Bayesian networks to highlight the impact of data paucity and direct future analyses: a contribution to the debate on measuring and reporting the precision of likelihood ratios. *Sci. Justice* 56:402–10

Tversky A, Kahneman D. 1983. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol. Rev.* 90:293–315

van den Hout A, Alberink I. 2016. Posterior distributions for likelihood ratios in forensic science. *Sci. Justice* 56:397–401

van Leeuwen DA, Brümmer N. 2013. The distribution of calibrated likelihood ratios in speaker recognition. arXiv:1304.1199 [stat.AP]

Vergeer P, Alberink I, Sjerps M, Ypma R. 2020. Why calibrating LR-systems is best practice. A reaction to "The evaluation of evidence for microspectrophotometry data using functional data analysis," in FSI 305. *Forensic Sci. Int.* 314:110388

Wigmore JH. 1913. The problem of proof. *Ill. Law Rev.* 8:77–103

Zadora G, Martyna A, Ramos D, Aitken CGG. 2014. *Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data*. Chichester, UK: John Wiley & Sons

Bayesian hierarchical multivariate models with *R* code.