

Legal Probabilism (final version)

Marcello Di Bello
Arizona State University
mdibello@asu.edu

Rafal Urbaniak
University of Gdańsk
rfl.urbaniak@gmail.com

December 7, 2020

Legal probabilism is a research program that relies on probability theory to analyze, model and improve the evaluation of evidence and the process of decision-making in trial proceedings. While the expression ‘legal probabilism’ seems to have been coined by Haack (2014b), the underlying idea can be traced back to the early days of probability theory (see, for example, Bernoulli, 1713). Another term that is sometimes encountered in the literature is ‘trial by mathematics’ coined by Tribe (1971). Legal probabilism remains a minority view among legal scholars, but attained greater popularity in the second half of the 20th century in conjunction with the law and economics movement (Calabresi, 1961; Becker, 1968; Posner, 1973).

To illustrate the range of applications of legal probabilism, consider a stylized case. Alice is charged with murder. Traces at the crime scene, which the perpetrator probably left, genetically match with Alice. An eyewitness testifies that Alice ran away from the scene after the crime was committed. Another witness asserts that Alice had previously threatened the victim multiple times. Alice, however, has an alibi. Mark claims he was with her for the entire day. This case raises several questions. How should the evidence be evaluated? How to combine conflicting pieces of evidence, such as the incriminating DNA match and the alibi testimony? If the standard of decision in a murder case is proof beyond a reasonable doubt, how strong should the evidence be to meet this standard?

As this entry will show, legal probabilism is a theoretical framework that helps to address these different questions. Finkelstein and Fairley (1970) gave one of the first systematic analyses of how probability theory, and Bayes’ theorem in particular, can help to evaluate evidence at trial (see Section 1). After the discovery of DNA fingerprinting, many legal probabilists focused on how probability theory could be used to quantify the strength of a DNA match (see Section 2). Recent work in Artificial Intelligence made it possible to use probability theory—in the form of Bayesian networks—to evaluate complex bodies of evidence consisting of multiple components (see Section 3). Following the work of Lempert (1977), likelihood ratios are now commonly used as probabilistic measures of the relevance of the evidence presented at trial (see Section 4). Other authors, starting with a seminal paper by Kaplan (1968), deployed decision theory to model decision-making and standards of proof at trial (see Section 5).

Legal probabilism is no doubt controversial. Many legal theorists and philosophers, starting with Tribe (1971), levelled several critiques against it. These critiques range from

difficulties in assessing the probability of someone’s criminal or civil liability to the dehumanization of trial decisions.

The most difficult challenge for legal probabilism—one that has galvanized philosophical attention in recent years—comes from the paradoxes of legal proof or puzzles of naked statistical evidence. In a number of seminal papers, Nesson (1979), Cohen (1981) and Thomson (1986) formulated scenarios in which, despite a high probability of guilt or civil liability based on the available evidence, a verdict against the defendant seems unwarranted (see Section 6). Other challenges for legal probabilism include the problem of conjunction and the reference class problem (see Section 7).

Legal probabilism can also be understood as a far reaching research program that aims to analyze—by means of probability theory—the trial system as a whole, including institutions such as the jury system and trial procedure. Some early French probability theorists examined the relationship between jury size, jury voting rules and the risk of convicting an innocent (de Condorcet, 1785; Laplace, 1814; Poisson, 1837; for more recent discussions, see Kaye, 1980; Nitzan, 2010; Suzuki, 2015).

At the root of this more radical version of legal probabilism lies the dream of discerning patterns in the behavior of individuals and improving legal institutions (Hacking, 1990). Today’s rapid grow of data—paired with machine learning and the pervasiveness of cost-benefit analysis—has rendered this dream more alive than ever before (Ferguson, 2020). For a critique of mathematization, quantification and cost-benefit analysis applied to the justice system, see (Harcourt, 2018). This entry will not, however, discuss this far reaching version of legal probabilism.

Contents

1	Probabilistic toolkit	3
1.1	Probability and its interpretation	3
1.2	Probabilistic fallacies	4
1.3	Odds version of Bayes’ theorem	7
1.4	Where do the numbers come from?	8
1.5	Source, activity and offense level hypotheses	9
2	The strength of evidence	10
2.1	Bayesian factor v. likelihood ratio	10
2.2	Cold-hit DNA matches	11
2.3	Choosing competing hypotheses	14
2.4	The two-stain problem	16
3	Bayesian networks for legal applications	17
3.1	Bayesian networks to the rescue	17
3.2	Idioms	19
3.3	Modeling an entire case	22
4	Relevance	23
4.1	Likelihood ratios	24
4.2	The Small Town Murder objection	24
5	Standards of proof	25
5.1	Legal background	25
5.2	Probability thresholds	26
5.3	Minimizing expected costs	27

5.4	Alternatives to probabilistic thresholds	29
6	Naked statistical evidence	29
6.1	Blue Bus, Gatecrasher, Prisoner	30
6.2	Cold-hits	31
6.3	Revisionist responses	31
6.4	Non-revisionist responses	32
7	Further objections	33
7.1	The difficulty about conjunction	33
7.2	Cohen's other objections	34
7.3	The problem of priors	35
7.4	The reference class problem	36
8	Relation to other entries in SEP	39

1 Probabilistic toolkit

This section begins with a review of the axioms of probability and its interpretations, and then shows how probability theory helps to spot mistakes that people may fall prey to when they assess evidence at trial, such as the prosecutor's fallacy, the base rate fallacy and the defense attorney's fallacy. This section also examines how probabilities are assigned to hypotheses and how hypotheses are formulated at different levels of granularity.

1.1 Probability and its interpretation

Standard probability theory consists of three axioms:

In words	In symbols
Non-negativity: The probability of any proposition A is greater than or equal to 0.	$\Pr(A) \geq 0$
Normality: The probability of any logical tautology is 1.	If $\models A$, then $\Pr(A) = 1$
Additivity: The probability of the disjunction of two propositions A and B is the sum of their respective probabilities, provided the two propositions are logically incompatible.	If $\models \neg(A \wedge B)$, then $\Pr(A \vee B) = \Pr(A) + \Pr(B)$

An important notion in probability theory is that of conditional probability, that is, the probability of a proposition A conditional on a proposition B , in symbols, $\Pr(A|B)$. Although it is sometimes taken as a primitive notion, conditional probability is usually defined as the probability of the conjunction $\Pr(A \wedge B)$ divided by the probability of the proposition being conditioned on, $\Pr(B)$, or in other words,

$$\Pr(A|B) = \Pr(A \wedge B) / \Pr(B) \text{ assuming } \Pr(B) \neq 0.$$

This notion is crucial in legal applications. The fact-finders at trial might want to know the probability of 'The defendant was at the scene when the crime was committed' conditional on 'Mrs. Dale asserts she saw the defendant run away from the scene.' Or they might want to know the probability of 'The defendant is the source of the traces found at the crime scene' conditional on 'The DNA expert asserts that the defendant's DNA matches the traces at the

scene.’ In general, the fact-finders are interested in the probability of a given hypothesis H about what happened conditional on the available evidence E , in symbols, $\Pr(H|E)$.

Most legal probabilists agree that the probabilities ascribed to statements that are disputed in a trial—such as ‘The defendant is the source of the crime traces’ or ‘The defendant was at the crime scene when the crime was committed’—should be understood as evidence-based degrees of belief (see, for example, Cullison, 1969; Kaye, 1979a; Nance, 2016).¹ This interpretation addresses the worry that since past events did or did not happen, their probabilities should be 1 or 0. Even if the objective chances are 1 or 0, statements about past events could still be assigned different degrees of belief given the evidence available. In addition, degrees of belief are better suited than frequencies for applications to unrepeatable events, such as actions of individuals, which are often the focus of trial disputes.

Some worry that, except for obeying the axioms of probability, degrees of belief are in the end assigned in a subjective and arbitrary manner (Allen and Pardo, 2019). This worry can be alleviated by noting that degrees of belief should reflect a conscientious assessment of the evidence available which may also include empirical frequencies (see 1.2 below and the examples in ENFSI, 2015). In some cases, however, the relevant empirical frequencies will not be available. When this happens, degrees of belief can still be assessed by relying on common sense and experience. Sometimes there will be no need to assign exact probabilities to every statement about the past.

The relevant probabilities can be expressed approximately with sets of probability measures (Walley, 1991; Shafer, 1976), probability distributions over parameter values, or intervals (see later in 1.4).

1.2 Probabilistic fallacies

Setting aside the practical difficulties of assigning probabilities to different statements, probability theory is a valuable tool in detecting misinterpretations of the evidence and reasoning fallacies that may otherwise go unnoticed. Reasoning without probabilities falls prey to even more subjectivity and error.

Assuming independence A common error that probability theory helps to identify consists in assuming without justification that two events are independent of one another. A theorem of probability theory states that the probability of the conjunction of two events, $A \wedge B$, equals the product of the probabilities of the conjuncts, A and B , that is,

$$\Pr(A \wedge B) = \Pr(A) \times \Pr(B),$$

provided A and B are independent of one another in the sense that the conditional probability $\Pr(A|B)$ is the same as the unconditional probability $\Pr(A)$. More formally, note that $\Pr(A|B) = \Pr(A \wedge B) / \Pr(B)$, so $\Pr(A \wedge B) = \Pr(A) \times \Pr(B)$ provided $\Pr(A) = \Pr(A|B)$. The latter equality means that the occurrence of B does not correlate with the occurrence of A .

The trial of Janet and Malcolm Collins, a couple accused of robbery in 1964 Los Angeles, illustrates how the lack of independence between events can be overlooked. The couple was identified based on features at the time considered unusual. The prosecutor called an expert witness, a college mathematician, to the stand, and asked him to consider the following features and assume they had the following probabilities: black man with a beard (1 in 10), man with a mustache (1 in 4), white woman with blond hair (1 out of 3), woman with a ponytail (1 out of 10), interracial couple in a car (1 out of 1000), couple driving a

¹Further engagement with these issues lies beyond the scope of this entry. For a more extensive discussion, see the entry on the interpretations of probability (Hájek, 2012) as well as (Skyrms, 1968; Gillies, 2000; Mellor, 2004; Childers, 2013).

yellow convertible (1 out of 10). The mathematician, correctly, calculated the probability of a random couple displaying all these features on the assumption of independence: 1 in 12 million (assuming the individual probability estimates were correct). Relying on this argument, the jury convicted the couple. If those features are so rare in Los Angeles and the robbers had them—the jury must have reasoned—the Collins must be the robbers.

The conviction was later reversed by the Supreme Court of California in *People v. Collins* (68 Cal.2d 319, 1968). The Court pointed out the mistake of assuming that multiplying the probabilities of each feature would give the probability of their joint occurrence. This assumption holds only if the features in question are probabilistically independent. But this is not the case since the occurrence of, say, the feature ‘man with a beard’ might very well correlate with the feature ‘man with a mustache.’ The same correlation might hold for the features ‘white woman with blond hair’ and ‘woman with a ponytail.’ Besides the lack of independence, another problem is the fact that the probabilities associated with each feature were not obtained by any reliable method.

The British case *R. v. Clark* (EWCA Crim 54, 2000) is another example of how the lack of independence between events can be easily overlooked. Sally Clark had two sons. Her first son died in 1996 and her second son died in similar circumstances a few years later in 1998. They both died within a few weeks after birth. Could it just be a coincidence? At trial, the paediatrician Roy Meadow testified that the probability that a child from an affluent family such as the Clark’s would die of Sudden Infant Death Syndrome (SIDS) was 1 in 8,543. Assuming that the two deaths were independent events, Meadow calculated that the probability of both children dying of SIDS was $1/8,543 \times 1/8,543$, approximately $1/73 \times 10^6$ or 1 in 73 million. This impressively low number no doubt played a role in the outcome of the case. Sally Clark was convicted of murdering her two infant sons (though the conviction was ultimately reversed on appeal). The $1/73 \times 10^6$ figure rests on the assumption of independence. This assumption is seemingly false since environmental or genetic factors may predispose a family to SIDS (for a fuller discussion of this point, see Dawid, 2002; Sesardic, 2007; Barker, 2017).

The prosecutor’s fallacy Another mistake that people often make while assessing the evidence presented at trial consists in conflating the two directions of conditional probability, $\Pr(A|B)$ and $\Pr(B|A)$. For instance, if you toss a die, the probability that the result is 2 given that it is even (which equals $1/3$) is different from the probability that the result is even given that it is 2 (which equals 1).

In criminal cases, confusion about the two directions of conditional probability can lead to exaggerating the probability of the prosecutor’s hypothesis. Suppose an expert testifies that the blood found at the crime scene matches the defendant’s and it is 5% probable that a person unrelated to the crime—someone who is *not* the source of the blood found at the scene—would match by coincidence. Some may be tempted to interpret this statement as saying that the probability that the defendant is not the source of the blood is 5% and thus it is 95% probable that the defendant is the source. This flawed interpretation is known as the *prosecutor’s fallacy* (Thompson and Schumann, 1987). The 5% figure is the conditional probability $\Pr(\text{match}|\neg\text{source})$ that, assuming the defendant is not the source of the crime scene blood ($\neg\text{source}$), he would still match (match). The 5% figure is not the probability $\Pr(\neg\text{source}|\text{match})$ that, if the defendant matches (match), he is not the source ($\neg\text{source}$). By conflating the two directions and thinking that $\Pr(\neg\text{source}|\text{match}) = \Pr(\text{match}|\neg\text{source}) = 5\%$, one is led to erroneously conclude that $\Pr(\text{source}|\text{match}) = 95\%$.

The same conflation occurred in the Collins case discussed earlier. Even if the calculations were correct, the 1 in 12 million probability that a random couple would have the specified characteristics should be interpreted as $\Pr(\text{match}|\text{innocent})$, not as the probability that the Collins were innocent given that they matched the eyewitness description,

$\Pr(\text{innocent}|\text{match})$. Presumably, the jurors convicted the Collins because they thought it was virtually impossible that they were not the robbers. They thought $\Pr(\text{innocent}|\text{match})$ was very low. But the 1 in 12 million figure, assuming it is correct, only shows that $\Pr(\text{match}|\text{innocent})$ equals $1/12 \times 10^6$, not that $\Pr(\text{innocent}|\text{match})$ equals $1/12 \times 10^6$.

Bayes' theorem and the base rate fallacy The relation between the probability of the hypothesis given the evidence, $\Pr(H|E)$, and the probability of the evidence given the hypothesis, $\Pr(E|H)$, is captured by Bayes' theorem:

$$\Pr(H|E) = \frac{\Pr(E|H)}{\Pr(E)} \times \Pr(H), \text{ assuming } \Pr(E) \neq 0.$$

The probability $\Pr(H)$ is called the *prior* probability of H (it is prior to taking evidence E into account) and $\Pr(H|E)$ is the *posterior* probability of H . This terminology is standard, but is slightly misleading because it suggests a temporal ordering which does not have to be there. Next, consider the ratio $\Pr(E|H)/\Pr(E)$, sometimes called the *Bayesian factor*.² This is the ratio between the probability $\Pr(E|H)$ of observing evidence E assuming H (often called, confusingly, the likelihood) and the probability $\Pr(E)$ of observing E independently of H . By the *law of total probability*, $\Pr(E)$ results from adding the probability of observing E assuming H and the probability of observing E assuming $\neg H$, each weighted by the prior probabilities of H and $\neg H$ respectively:

$$\Pr(E) = \Pr(E|H) \times \Pr(H) + \Pr(E|\neg H) \times \Pr(\neg H).$$

As is apparent from Bayes' theorem, multiplying the prior probability by the Bayesian factor $\Pr(E|H)/\Pr(E)$ yields the posterior probability $\Pr(H|E)$. Other things being equal, the lower the prior probability $\Pr(H)$, the lower the posterior probability $\Pr(H|E)$. The *base rate fallacy* consists in ignoring the effect of the prior probability on the posterior (Kahneman and Tversky, 1973). This leads to thinking that the posterior probability of a hypothesis given the evidence is different than it actually is (Koehler, 1996).

Consider the blood evidence example discussed previously. By Bayes' theorem, the two conditional probabilities $\Pr(\neg\text{source}|\text{match})$ and $\Pr(\text{match}|\neg\text{source})$ are related as follows:

$$\Pr(\neg\text{source}|\text{match}) = \frac{\Pr(\text{match}|\neg\text{source})}{\Pr(\text{match}|\neg\text{source}) \times \Pr(\neg\text{source}) + \Pr(\text{match}|\text{source}) \times \Pr(\text{source})} \times \Pr(\neg\text{source}).$$

Absent any other compelling evidence to the contrary, it should initially be very likely that the defendant, as anyone else, had little to do with the crime. Say, for illustrative purposes, that the prior probability $\Pr(\neg\text{source}) = .99$ and so $\Pr(\text{source}) = .01$ (more on how to assess prior probabilities later in 1.4). Next, $\Pr(\text{match}|\text{source})$ can be set approximately to 1, because if the defendant were the source of the blood at the crime scene, he should match the blood at the scene (setting aside the possibility of false negatives). The expert estimated the probability that someone who is not the source would coincidentally match the blood found at the scene — that is, $\Pr(\text{match}|\neg\text{source})$ — as equal to .05. By Bayes' theorem,

$$\Pr(\neg\text{source}|\text{match}) = \frac{.05}{.05 \times .99 + 1 \times .01} \times .99 \approx .83.$$

The posterior probability that the defendant is the source, $\Pr(\text{source}|\text{match})$ is roughly $1 - .83 = .17$, much lower than the exaggerated value of .95. This posterior probability would have dropped even lower if the prior probability were lower. A similar analysis applies to the Collins case. If the prior probability of the Collins's guilt is sufficiently low—say in 1 in 6 million—the posterior guilt probability given the match with the eyewitness description would be roughly .7, much less impressive than $(1 - 1/12 \times 10^6) \approx .9999999$.

²Some call it 'normalized likelihood' and reserve the expression 'Bayesian factor' for what later in 1.3 is called the 'likelihood ratio'

Defense attorney's fallacy Although base rate information should not be overlooked, paying excessive attention to it and ignoring other evidence leads to the so-called *defense attorney's fallacy*. As before, suppose the prosecutor expert testifies that the defendant matches the traces found at the crime scene and there is a 5% probability that a random person, unrelated to the crime, would accidentally match, so $\Pr(\text{match}|\neg\text{source}) = .05$. To claim that it is 95% likely that the defendant is the source of traces, or in symbols, $\Pr(\text{source}|\text{match}) = .95$, would be to commit the prosecutor's fallacy described earlier. But suppose the defense argues that, since the population in town is 10,000 people, 5% of them would match accidentally, that is, $10,000 \times 5\% = 500$ people. Since the defendant could have left the traces as any of the other 499 people, he is $1/500 = .2\%$ likely to be the source, a rather unimpressive figure. The match—the defense concludes—is worthless evidence.

This analysis portrays the prosecutor's case as weaker than it need be. If the investigators narrowed down the set of suspects to, say, 100 people—a potential piece of information the defense ignored—the match would make the defendant 16% likely to be the source. This fact can be verified using Bayes' theorem (left as an exercise for the reader). Even assuming, as the defense claims, that the pool of suspects comprises as many as 10,000 people, the prior probability that the defendant is the source would be $1/10,000$, and since the defendant matches the traces, this probability should rise to (approximately) $1/500$. Given the upward shift in the probability from $1/10,000$ to $1/500$, the match should not be considered worthless evidence (more on this later in 2.1).

1.3 Odds version of Bayes' theorem

The version of Bayes' theorem considered so far is informationally demanding since the law of total probability—which spells out $\Pr(E)$, the denominator in the Bayesian factor $\Pr(E|H)/\Pr(E)$ —requires one to consider H and the catch-all alternative hypothesis $\neg H$, comprising all possible alternatives to H . A simpler version of Bayes' theorem is the so-called odds formulation:

$$\frac{\Pr(H|E)}{\Pr(H'|E)} = \frac{\Pr(E|H)}{\Pr(E|H')} \times \frac{\Pr(H)}{\Pr(H')},$$

or in words:

$$\text{posterior odds} = \text{likelihood ratio} \times \text{prior odds}.$$

The ratio $\Pr(H)/\Pr(H')$ represents the *prior odds*, where H and H' are two competing hypotheses, not necessarily one the complement of the other. The *likelihood ratio* $\Pr(E|H)/\Pr(E|H')$ compares the probability of the same item of evidence E given the two hypotheses. The *posterior odds* $\Pr(H|E)/\Pr(H'|E)$ compares the probabilities of the hypotheses given the evidence. This ratio is different from the probability $\Pr(H|E)$ of a specific hypothesis given the evidence.

As an illustration, consider the Sally Clark case (previously discussed in 1.2). The two hypotheses to compare are that Sally Clark's sons died of natural causes (natural) and that Clark killed them (kill). The evidence available is that the two sons died in similar circumstances one after the other (two deaths). By Bayes' theorem (in the odds version),

$$\frac{\Pr(\text{kill}|\text{two deaths})}{\Pr(\text{natural}|\text{two deaths})} = \frac{\Pr(\text{two deaths}|\text{kill})}{\Pr(\text{two deaths}|\text{natural})} \times \frac{\Pr(\text{kill})}{\Pr(\text{natural})}.$$

The likelihood ratio $\Pr(\text{two deaths}|\text{kill})/\Pr(\text{two deaths}|\text{natural})$ compares the likelihood of the evidence (two deaths) under each hypothesis (kill v. natural). Since under either hypothesis both babies would have died—by natural causes or killing—the ratio should equal to one (Dawid, 2002). What about the prior odds $\Pr(\text{kill})/\Pr(\text{natural})$? Recall the $1/73 \times 10^6$ figure given by the pediatrician Roy Meadow. This figure was intended to convey how unlikely it was that two

babies would die of natural causes, one after the other. If it so unlikely that both would die of natural causes—one might reason—it must be likely that they did not actually die of natural causes and quite likely that Clark killed them. But did she?

The prior probability that they died of natural causes should be compared to the prior probability that a mother would kill them. To have a rough idea, suppose that in a mid-size country like the UK 1 million babies are born every year of whom 100 are murdered by their mothers. So the chance that a mother would kill one baby in a year is 1 in 10,000. What is the chance that the same mother kills *two* babies? Say we appeal to the (controversial) assumption of independence. On this assumption, the chance that a mother kills two babies equals 1 in 100 million. Assuming independence, $\Pr(\text{natural}) = 1/73 \times 10^6$ and $\Pr(\text{kill}) = 1/100 \times 10^6$. This means that the prior odds $\Pr(\text{kill})/\Pr(\text{natural})$ would equal .73. With a likelihood ratio of one, the posterior odds $\Pr(\text{kill}|\text{two deaths})/\Pr(\text{natural}|\text{two deaths})$ would also equal .73. On this analysis, that Clark killed her sons would be .73 times less likely than they died of natural causes, or in other words, the natural cause hypothesis is 1.37 times more likely than the hypothesis that Clark killed her sons.

A limitation of this analysis should be kept in mind. The .73 ratio is a measure of the probability of one hypothesis compared to another. From this ratio alone, the posterior probability of the individual hypotheses cannot be deduced. Only if the competing hypotheses H and H' are exclusive and exhaustive, say one is the negation of the other, can the posterior probability be derived from the posterior odds $PO = P(H|E)/P(H'|E)$ via the equality $\Pr(H|E) = PO/(1+PO)$. But the hypotheses kill and natural are not exhaustive, since the two babies could have died in other ways. So the posterior odds $\Pr(\text{kill}|\text{two deaths})/\Pr(\text{natural}|\text{two deaths})$ cannot be translated into the posterior probabilities of individual hypotheses. (A more sophisticated probabilistic analysis using a Bayesian network is presented in 3.3.)

1.4 Where do the numbers come from?

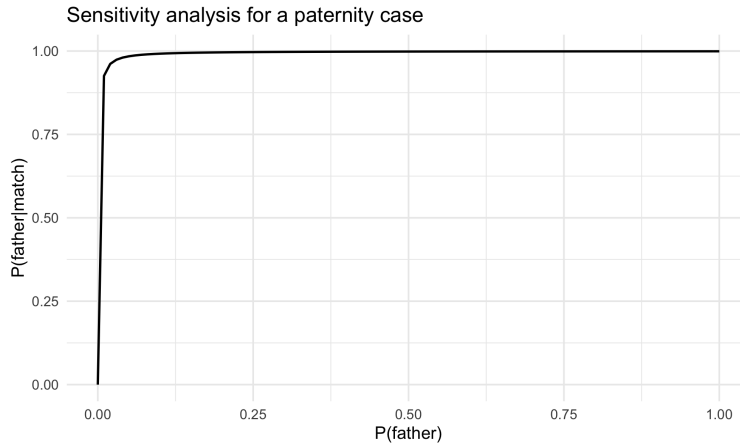
In the examples considered so far, probabilities were assigned on the basis of empirical frequencies and expert opinions. For example, an expert testimony that only 5% of people possess a particular blood type was used to set $\Pr(\text{match}|\neg\text{source}) = .05$. Probabilities were also assigned on the basis of common sense, for example, $\Pr(\text{match}|\text{source})$ was set to 1 assuming that if someone is the source of the crime traces, the person must match the crime traces (setting aside the possibility of framing, fabrication of the traces or false negatives in the test results). But probabilities need not always be assigned exact values. Agreeing on an exact value can in fact be extremely difficult, especially for prior probabilities. This is a challenge for legal probabilism (see later in 7.3). One way to circumvent this challenge is to avoid setting exact values and adopt plausible intervals. This method is based on *sensitivity analysis*, an assessment of how prior probabilities affect other probabilities.

Consider the paternity case *State v. Boyd* (1983), 331 N.W. 2d 480. Expert witness Dr. Polesky testified that 1,121 unrelated men would have to be randomly selected from the general population before another man could be found with all the appropriate genes to have fathered the child in question. In other words, the probability that a random person who is not the father would still genetically match is $1/1121$, or in symbols, $\Pr(\text{match}|\neg\text{father}) = 1/1121$. As explained earlier in 1.2, this probability cannot be easily translated into the probability that a person whose genetic profile matches is the father, or in symbols, $\Pr(\text{father}|\text{match})$. The latter should be calculated using Bayes' theorem. Suppose the prior probability that the defendant (without yet considering the genetic match) would be the father is as low as .01, or in symbols, $\Pr(\text{father}) = .01$. The probability that the defendant is actually the father (after taking into account the genetic match) is about .92. By Bayes' theorem:

$$\begin{aligned}\Pr(\text{father}|\text{match}) &= \frac{\Pr(\text{match}|\text{father}) \times \Pr(\text{father})}{\Pr(\text{match})} \\ &= \frac{1 \times .01}{1 \times .01 + 1/1121 \times .99} = .9256813\end{aligned}$$

But why take the prior probability $\Pr(\text{father})$ to be .01 and not something else? Note that this prior probability plays a part in more than one place in the calculations. It occurs in the numerator as well as the denominator.

The idea in sensitivity analysis is to look at a range of plausible probability assignments and investigate the impact of such choices. In legal applications, the key question is whether assignments that are most favorable to the defendant would still return a strong evidentiary case against the defendant. In the case at hand, the expert testimony is so strong that for a wide spectrum of the prior probability $\Pr(\text{father})$, the posterior probability $\Pr(\text{father}|\text{match})$ remains high, as this plot shows:



The posterior surpasses .9 once the prior is above approximately .008. In a paternity case, given the mother's testimony and other evidence, it is clear that the probability of fatherhood before the expert testimony is taken into account should be higher than that, no matter what the prior probability should exactly be. Without settling on an exact value, the interval of values above .008 would guarantee a posterior probability of paternity to be at least .9. This is amply sufficient to meet the preponderance standard that governs civil cases such as paternity disputes (on standards of proof, see later in Section 5.)

1.5 Source, activity and offense level hypotheses

Difficulties in assessing probabilities go hand in hand with the choice of the hypotheses of interest. To some approximation, hypotheses can be divided into three levels: offence, activity, and source level hypotheses. At the offence level, the issue is whether the defendant is guilty, as in the statement 'Smith is guilty of manslaughter.' At the activity level, hypotheses describe what happened and what those involved did or did not do. An example of activity level hypothesis is 'Smith stabbed the victim.' Finally, source level hypotheses describe the source of the traces, such as 'The victim left the stains at the scene,' without specifying how the traces got there. Overlooking differences in hypothesis level can lead to serious confusions.

Consider a case in which a DNA match is the primary incriminating evidence. DNA evidence is one of the most widely used forms of quantitative evidence currently available, but it does not work much differently from blood evidence or other forms of trace evidence. In testifying about a DNA match, trial experts will often assess the probability that a random person, unrelated to the crime, would coincidentally match the crime stain profile (Foreman et al., 2003). This is called the Genotype Probability or Random Match Probability, which can be extremely low, in the order of 1 in 100 million or even lower (Donnelly, 1995; Kaye and Sensabaugh, 2000; Wasserman, 2008). It is tempting to equate the random match probability to $\Pr(\text{match}|\text{innocence})$ and together with the prior $\Pr(\text{innocence})$ use Bayes' theorem to calculate the posterior probability of innocence $\Pr(\text{innocence}|\text{match})$. This would be a mistake. Applying Bayes' theorem is of course recommended and helps to avoid the prosecutor's fallacy, the conflation of $\Pr(\text{innocence}|\text{match})$ and $\Pr(\text{match}|\text{innocence})$. The problem here lies elsewhere. Equating the random match probability with $\Pr(\text{match}|\text{innocence})$ overlooks the difference between offense, activity and source level hypotheses. A DNA match cannot speak directly to the question of guilt or innocence. Even if the suspect actually left the genetic material at the scene—source level proposition—the match does not establish guilt. Even if the defendant did visit the scene and came into contact with the victim, it does not follow that they committed the crime they were accused of.

Few forms of evidence can speak directly to offense level hypotheses. Circumstantial evidence that is more amenable to a probabilistic quantification, such as DNA matches and other trace evidence, does not. Eyewitness testimony may speak more directly to offense level hypotheses, but it is also less easily amenable to a probabilistic quantification (see however Friedman, 1987, recent results by Wixted and Wells, 2017 and a survey of related issues by Urbaniak et al., 2020). This makes it difficult to assign probabilities to offense level hypotheses. Moving beyond source level hypotheses requires a close collaboration between scientists, investigators and attorneys (see Cook et al., 1998, for a discussion).

2 The strength of evidence

The posterior probability of a hypothesis given the evidence should not be confused with the strength (or probative value, weight) of the evidence in favor of the hypothesis. The strength of an item of evidence reflects its impact on the probability of the hypothesis. Suppose the prior probability of H is extremely low, say $\Pr(H) = .001$, but taking evidence E into account brings this probability up to 35%, that is, $\Pr(H|E) = .35$. This is a dramatic upward shift. Even though the posterior probability of H given E is not high, E strongly favors H . This section examines how probability theory helps to assess the strength of evidence.

2.1 Bayesian factor v. likelihood ratio

The notion of strength of evidence—as distinct from prior and posterior probability—can be captured formally in a number of different ways. (For a comprehensive discussion, see the entry on confirmation theory.³) One measure of the strength of evidence is the Bayesian factor $\Pr(E|H)/\Pr(E)$ (already discussed in 1.2). This is an intuitively plausible measure of evidential strength. Note that by Bayes' theorem

$$\Pr(H|E) = \text{BayesianFactor}(H, E) \times \Pr(H),$$

and thus the Bayesian factor is greater than one if and only if the posterior probability $\Pr(H|E)$ is higher than the prior probability $\Pr(H)$. The greater the Bayesian factor (for

³SEP entry: <https://plato.stanford.edu/entries/confirmation/>

values above one), the greater the upward shift from prior to posterior probability, the more strongly E positively supports H . Conversely, the smaller the Bayesian factor (for values below one), the greater the downward shift from prior to posterior probability, the more strongly E negatively supports H . If $\Pr(H) = \Pr(H|E)$ the evidence has no impact, upwards or downwards, on the prior probability of H .

The Bayesian factor $\Pr(E|H)/\Pr(E)$ is an absolute measure of evidence E 's support toward H since it compares the probability of E under hypothesis H against the probability of E in general. The denominator is calculated following the law of total probability:

$$\Pr(E) = \Pr(E|H)\Pr(H) + \Pr(E|\neg H)\Pr(\neg H).$$

The catch-all alternative hypothesis $\neg H$ can be replaced by a more fine-grained set of alternatives, say H_1, H_2, \dots, H_k , provided H and its alternatives cover the entire space of possibilities. The law of total probability would then read:

$$\Pr(E) = \Pr(E|H)\Pr(H) + \sum_{i=1}^k \Pr(E|H_i)\Pr(H_i).$$

Instead of the Bayesian factor, the strength of evidence can be assessed by means of the likelihood ratio, a comparative measure of whether evidence E supports a hypothesis H more than a competing hypothesis H' , in symbols, $\Pr(E|H)/\Pr(E|H')$. An expert, for instance, may testify that the blood-staining on the jacket of the defendant is ten times more likely to be seen if the wearer of the jacket hit the victim (hypothesis H) rather than if he did not (hypothesis H') (Aitken et al., 2010, p. 38). If the evidence supports H more than H' , the ratio would be above one, and if the evidence supports H' more than H , the ratio would be below one. The greater the likelihood ratio (for values above one), the stronger the evidence in favor of H as contrasted with H' . The smaller the likelihood ratio (for values below one), the stronger the evidence in favor of the competing hypothesis H' as contrasted with H .

The relationship between likelihood ratio $\Pr(E|H)/\Pr(E|H')$ and posterior odds $\Pr(H|E)/\Pr(H'|E)$ is apparent in the odds version of Bayes' theorem (see earlier in 1.3). If the likelihood ratio is greater (lower) than one, the posterior odds will be greater (lower) than the prior odds of H . The likelihood ratio, then, is a measure of the upward or downward impact of the evidence on the odds of two hypotheses H and H' .

A competitor of the likelihood ratio as a measure of evidentiary strength is an even simpler notion, the probability $\Pr(E|H)$. It is tempting to think that, whenever $\Pr(E|H)$ is low, E should be strong evidence against H . Consider an example by Robertson et al. (2016). In a child abuse case, the prosecutor offers evidence that a couple's child rocks (a movement pattern) and that only 3% of non-abused children rock, $\Pr(\text{child rocks}|\text{no abuse}) = .03$. If it is unlikely that a non-abused child would rock, the fact that this child rocks might seem strong evidence of abuse. But this reading of the 3% figure is mistaken. It could well be that 3% of abused children rock, $\Pr(\text{child rocks}|\text{abuse}) = .03$. If rocking is unlikely under either hypothesis—which means the likelihood ratio $\Pr(\text{child rocks}|\text{abuse})/\Pr(\text{child rocks}|\text{no abuse})$ equals one—rocking cannot count as evidence of abuse. In order to avoid exaggerations of the evidence, it is best to assess it by means of the likelihood ratio rather the probability of the evidence given a hypothesis (Royall, 1997; Triggs and Buckleton, 2004; ENFSI, 2015).

2.2 Cold-hit DNA matches

To better understand likelihood ratios, it is instructive to look at DNA evidence as a case study, focusing in particular on cold-hit matches. DNA evidence may be used to corroborate other evidence in a case or as the primary incriminating evidence. Suppose different

investigative leads point to an individual, Mark Smith, as the perpetrator. Suppose the investigators also find several traces at the crime scene left by the perpetrator, and laboratory analyses show that the genetic profile associated with the traces matches Smith. In this scenario, the DNA match corroborates the other evidence against Smith. In contrast, suppose the police have no suspect and no other investigative lead except the traces left at the crime scene. The police run the genetic profile associated with the traces through a database of profiles and find a match, a so-called *cold-hit*. Since in cold-hit cases there is no other evidence, cold-hit matches are the primary item of evidence against the defendant. Some scholars believe that this circumstance weakens the evidentiary value of the match. Others disagree. What follows examines some of the main arguments in this debate.

Random match v. database match Suppose an expert testifies that the crime traces genetically match the defendant and that the random match probability is extremely low, say 1 in 100 million. The random match probability—often interpreted as the probability that someone who is not the source would coincidentally match, $\Pr(\text{match}|\neg\text{source})$ —is a common measure of the strength of a DNA match. The lower this probability, the more strongly incriminating the match. Strictly speaking, a match is strong evidence that the defendant is the source only if the likelihood ratio $\Pr(\text{DNA match}|\text{source})/\Pr(\text{DNA match}|\neg\text{source})$ is significantly greater than one. In practice, however, when the random match probability is low—that is, $\Pr(\text{match}|\neg\text{source})$ is low—the likelihood ratio should be significantly above one because the probability that the individual who is the source would match, $\Pr(\text{match}|\text{source})$, should be high so long as the test has a low false negative rate. For practical purposes, then, a low random match probability does count as strong incriminating evidence.

When it comes to cold-hit matches, however, further complications emerge. The Puckett case can serve as an illustration. In 2008, John Puckett was identified through a database search of 338,000 profiles. He was the only individual in the database who matched the traces collected from Diana Sylvester, a victim of rape in 1972. The expert witness testified that Puckett's genetic profile should occur randomly among Caucasian men with a frequency of 1 in 1.1 million. This would seem strong evidence against Puckett's innocence. But the DNA expert for the defense, Bicka Barlow, pointed out that besides the cold-hit match the evidence against Puckett was slim. Barlow argued that the correct assessment of the cold-hit match required to multiply 1/1.1 million by the size of the database. Call the result of this multiplication the *database match probability*. Multiplying 1/1.1 million by 338,000 yields a database match probability of roughly 1/3, an unimpressive figure. If someone in the database could match with a probability of 1/3, the cold-hit match should not count as strong evidence against Puckett. This was Barlow's argument.

Barlow followed a 1996 report by the National Research Council, often referred to as NRC II (National Research Council, 1996). The report recommended that in cold-hit cases the random match probability should be multiplied by the size of the database. This correction was meant to guard against the heightened risk of mistaken matches for the innocent people in the database. NRC II used an analogy. If you toss several different coins at once and all show heads on the first attempt, this seems strong evidence that the coins are biased. If, however, you repeat this experiment sufficiently many times, it is almost certain that at some point all coins will land heads. This outcome should not count as evidence that the coins are biased. According to NRC II, repeating the coin toss experiment multiple times is analogous to trying to find a match by searching through a database of profiles. As the size of the database increases, it is more likely that someone in the database who had nothing to do with the crime would match.

The aptness of the analogy has been challenged (Donnelly and Friedman, 1999). Searching a larger database no doubt increases the probability of finding a match at some point. But judges and jurors should not be concerned with the statement 'At least one of the pro-

files in the database would randomly match the crime sample.’ They should be concerned with the statement ‘The profile of the defendant on trial would randomly match the crime sample.’ The probability of finding a match between the defendant and the crime sample does not increase because other people in the database are tested. In fact, suppose everyone in the world is recorded in the database. A unique cold-hit match would be extremely strong evidence of guilt since everybody would be excluded as a suspect except one matching individual. Instead, if the random match probability were multiplied by the size of the database, the probative value of the match should be quite low. This is counter-intuitive.

Another analogy is sometimes used to argue that the evidentiary value of a cold-hit match should be weakened. The analogy is between searching for a match in a database and multiple hypothesis testing, an objectionable research practice. In classical hypothesis testing, if the probability of type I error in a single test of a hypothesis is .05, this probability will increase by testing the same hypothesis multiple times. The database match probability—the argument goes—would correct for the increased risk of type I error. However, as Balding (2002, 2005) points out, multiple testing consists in testing the *same* hypothesis multiple times against new evidence. In cold-hit cases, multiple hypotheses—each concerning a different individual in the database—are tested only once and then excluded if a negative match occurs. From this perspective, the hypothesis that the defendant is the source is one of the many hypotheses subject to testing. The cold-hit match supports that hypothesis and rules out the others.

The likelihood ratio of cold-hit matches A more principled way to assess cold-hit matches is based on the likelihood ratio. The proposal draws from the literature on the so-called *island problem*, studied by Eggleston (1978), Dawid (1994), and Dawid and Mortera (1996). Let the prosecutor’s hypothesis H_p be ‘The suspect is the source of the crime traces’ and the defense hypothesis H_d be ‘The suspect is not the source of the crime traces.’ Let M be the DNA match between the crime stain and the suspect (included in the database) and D the information that no one among the profiles in the database matches the crime stain. The likelihood ratio associated with M and D should be (Balding and Donnelly, 1996; Taroni et al., 2014):

$$V = \frac{\Pr(M, D|H_p)}{\Pr(M, D|H_d)}.$$

Since $\Pr(A \wedge B) = \Pr(A|B) \times \Pr(B)$, for any statement A and B , this ratio can be written as

$$V = \frac{\Pr(M|H_p, D)}{\Pr(M|H_d, D)} \times \frac{\Pr(D|H_p)}{\Pr(D|H_d)}.$$

The first ratio $\Pr(M|H_p, D)/\Pr(M|H_d, D)$ is roughly $1/\gamma$, where γ is the random match probability. The second ratio $\Pr(D|H_p)/\Pr(D|H_d)$ is the *database search ratio*, defined as follows (see Balding and Donnelly (1996); Taroni et al. (2014) for details):

$$\frac{\Pr(D|H_p)}{\Pr(D|H_d)} = \frac{1}{1 - \varphi},$$

where $\Pr(S|H_d) = \varphi$ and S stands for the proposition that someone in the database is the source of the crime traces. Donnelly and Friedman (1999) derived a similar formula for this ratio. As the database gets larger, φ increases and the database search ratio increases. This ratio equals one only if no one in the database could be the source.

Since the likelihood ratio V of the cold-hit match results by multiplying the likelihood ratio of the DNA match and the database search ratio, V will always be greater than the

mere likelihood ratio of the match (except for the unrealistic case in which $\varphi = 0$). Thus, a cold-hit DNA match should count as stronger evidence than a DNA match of a previously identified suspect. Dawid and Mortera (1996) study different database search strategies and consider the possibility that information about the match is itself uncertain, but the general point remains. Under reasonable assumptions, ignoring the database search would give a conservative assessment of the evidentiary strength of the cold-hit match.

There is intuitive resistance in basing a conviction on a cold-hit match, but this resistance is less strong in case of an ordinary match (more on this later in Section 6). This preference for convictions based on an ordinary DNA match seems in tension with the claim that a cold-hit match is stronger evidence of guilt than an ordinary match. There is a way to reconcile both sides, however. The key is to keep in mind that the evidentiary strength—measured by the likelihood ratio—should not be confused with the posterior probability of guilt given the evidence. If the cold-hit match is the only evidence of guilt, the posterior probability of guilt may well be lower compared to cases in which other evidence, such as investigative leads, supplements the DNA match. This lower posterior probability would justify the intuitive resistance towards convictions in cold-hit cases despite the stronger probative value of the cold-hit match.

2.3 Choosing competing hypotheses

The likelihood ratio is helpful for assessing the strength of the evidence, as the preceding discussion about cold-hit matches shows. One major difficulty, however, is the choice of the hypotheses H and H' that should be compared. The hypotheses should compete with one another—say, in a criminal trial, H is the hypothesis put forward by the prosecution and H' is the hypothesis put forward by the defense. But this constraint leaves open the possibility for manipulations and misinterpretations of the evidence. What follows outlines some of the main controversies in the literature on this topic.

Ad hoc hypotheses and Barry George Consider a stylized DNA evidence case. Suppose the prosecutor puts forward the hypothesis that the suspect left the traces found at the crime scene. This hypothesis is well supported by laboratory analyses showing that the defendant genetically matches the traces. The defense, however, responds by putting forward the following *ad hoc* hypothesis: ‘The crime stain was left by some unknown person who happened to have the same genotype as the suspect.’ Since the probability of the DNA match given either hypothesis is 1, the likelihood ratio equals 1 (Evetts et al., 2000). The problem generalizes. For any item of evidence and any given prosecutor’s hypothesis H , there is an *ad hoc* competing hypothesis H^* such that $\Pr(E|H)/\Pr(E|H^*) = 1$. Hypothesis H^* is a just-so hypothesis, one that is selected only because it explains the evidence just as well as hypothesis H does (Mayo, 2018). If no further constraints are placed on the choice of the competing hypotheses—it would seem—no evidence could ever incriminate a defendant.

Judges and jurors will often recognize *ad hoc* hypotheses for what they are—artificial theories that should not be taken seriously. Perhaps, the common sense of the participants in a trial will suffice to constrain the choice of hypotheses in the right way. But real cases are complex, and it is not always obvious whether a choice of competing hypotheses, which are not obviously *ad hoc*, is legitimate or not.

A notable example is *R. v. George* (2007 EWCA Crim 2722). Barry George was accused of murdering British journalist and TV celebrity Jill Dando. A single particle of firearm residue was found one year later in George’s coat pocket and it matched the residue from the crime scene. This was the key incriminating evidence against him. George was convicted, and his first appeal was unsuccessful. After the first appeal, Ian Evetts from the Forensic Science Service worried that the evidence had not been properly assessed at trial. The jurors were

presented with the conditional probability of finding the firearm residue in George's coat given the defense hypothesis that George *did not* fire the gun. This probability was estimated to be quite low. But the jurors were not presented with the conditional probability of finding the same evidence given the prosecutor's hypothesis that George *did* fire the gun that shot Dando. An expert witness, Mr. Keeley, was asked to provide both conditional probabilities and estimated them to be $1/100$, which indicated that the firearm residue had no probative value. George appealed again in 2007, and relying on Keeley's estimates, won the appeal.

A study of the trial transcript shows that Keeley's choice of hypotheses lacked coherence and the likelihood ratio based on them was therefore meaningless (Fenton et al., 2014). On one occasion, Keeley compared the hypothesis that the particle found in George's pocket came from a gun fired by George himself, and the alternative hypothesis that the particle came from another source. On another occasion, Keeley took the prosecutor's hypothesis to be 'The particle found in George's pocket came from the gun that killed Dando.' But the conditional probability of the evidence given this hypothesis should not be low. It should actually be one. The most charitable reading of the trial transcript suggests that the expert had in mind the hypotheses 'George was the man who shot Dando' and 'The integrity of George's coat was corrupted.' But Keeley gave no justification for why these hypotheses should be compared in the likelihood ratio (see Fenton et al., 2014, for details).

A related complication is that competing hypotheses can concern any factual dispute, from minute details such as whether the cloth used to suffocate the victim was red or blue, to ultimate questions such as whether the defendant stabbed the victim. The likelihood ratio varies across hypotheses formulated at different levels of granularity: offense, activity and source level hypotheses (on this distinction, see earlier in 1.5). It is even possible that, at the source level, the likelihood ratio favors one side, say the prosecution, but at the offence level, the likelihood ratio favors the other side, say the defense, even though the hypotheses at the two levels are quite similar (Fenton et al., 2014).

Exclusive and exhaustive? The confusion in the Barry George case is attributable to the absence of clear rules for choosing the hypotheses in the likelihood ratio. One such rule could be: pick competing hypotheses that are exclusive (they cannot be both true) and exhaustive (they cannot be both false). In this way, the parties would not be able to pick *ad hoc* hypotheses and skew the assessment of the evidence in their own favor.

There are other principled reasons to follow the exclusive-and-exhaustive rule, such as that when the hypotheses are not exclusive or exhaustive, the likelihood ratio delivers counterintuitive results. For consider two competing hypotheses that are not mutually exclusive. Let H_p stand for 'The defendant is guilty' and H_d for 'The defendant was not at the crime scene.' Let E stand for 'Ten minutes before the crime took place the defendant—seen at a different location—was overheard on the phone saying *go ahead and kill him*.' The evidence positively supports each hypothesis, yet it is conceivable that the likelihood ratio should equal one in this context. Further, consider two competing hypotheses that are not exhaustive. Suppose Fred and Bill attempted to rob a man. The victim resisted, was struck on the head and died. Say H_p stand for 'Fred struck the fatal blow' and H_d stand for 'Bill struck the fatal blow.' The hypotheses are not exhaustive because a third hypothesis is 'The man did not die from the blow.' Suppose E is the information that the victim had a heart attack six months earlier. The likelihood ratio $\Pr(E|H_p)/\Pr(E|H_d)$ equals one since $\Pr(E|H_p) = \Pr(E|H_d)$. Yet E reduces the probability of both H_p and H_d . So, in this case, the evidence negatively supports each hypothesis, contrary to what the likelihood ratio suggests.

But relying on exclusive and exhaustive hypotheses is not without complications either. Consider an expert who decides to formulate the defense hypothesis by negating the prosecution hypothesis, say, 'the defendant did not hit the victim in the head.' This choice of defense hypothesis can be unhelpful in assessing the evidence. What is the probability

that the suspect would carry such and such blood stain if he did not hit the victim in the head? This depends on whether he was present at the scene, what he was doing at the time and many other circumstances. As Evett et al. (2000) point out, the choice of a particular hypothesis to be used in the evaluation of the strength of the evidence will depend on contextual factors. More often than not, the hypotheses chosen will not be mutually exclusive. Comparing exclusive and exhaustive hypotheses can also be unhelpful for jurors or judges making a decision at trial. In a paternity case, for example, the expert should not compare the hypotheses ‘The accused is the father of the child’ and its negation, but rather, ‘The accused is the father of the child’ and ‘The father of the child is a man unrelated to the putative father’ (Biedermann et al., 2014). Even though the relatives of the accused are potential fathers, considering such a far-fetched possibility would make the assessment of the evidence more difficult than needed (Evett et al., 2000).

Exclusive and exhaustive hypotheses guard against arbitrary comparisons and ensure a more principled assessment of the evidence. The drawback is that such hypotheses cover the entire space of possibilities, and sifting through this space is cognitively unfeasible. In this respect, comparing more circumscribed hypotheses is preferable. The danger of doing so, however, is that likelihood ratios heavily depend on the hypotheses that are compared. The more latitude in the choice of the hypotheses, the more variable the likelihood ratio as a measure of evidentiary value.

2.4 The two-stain problem

A case study that further illustrates the limitations of the likelihood ratio as a measure of evidentiary strength is the two-stain problem, originally formulated by Evett (1987).

In Evett’s original formulation, two stains from two different sources are left at the crime scene, and the suspect’s blood matches one of them. Let the first hypothesis be that the suspect was one of the two men who committed the crime and the second hypothesis the negation of the first. Evett (1987) shows (see his paper for details) that the likelihood ratio of the match relative to these two hypotheses is $1/2q_1$ where q_1 is the estimated frequency of the characteristics of the first stain. Surprisingly, the likelihood ratio does not depend on the frequency associated with the second stain.

Consider now a more complex two-stain scenario. Suppose a crime was committed by two people, who left two stains at the crime scene: one on a pillow and another on a sheet. John Smith, who was arrested for a different reason, genetically matches the DNA on the pillow, but not the one on the sheet. Meester and Sjerps (2004) argue that there are three plausible pairs of hypotheses associated with numerically different likelihood ratios (see their paper for details). The three options are listed below, where R is the random match probability of Smith’s genetic profile and δ the prior probability that Smith was one of the crime scene donors.

H_p	H_d	LR
Smith was one of the crime scene donors.	Smith was not one of the crime scene donors.	$R/2$
Smith was the pillow stain donor.	Smith was not one of the crime scene donors.	R
Smith was the pillow stain donor.	Smith was not the pillow stain donor.	$R(2-\delta)/2(1-\delta)$

Even though the likelihood ratios are numerically different, their posterior probabilities given the evidence are the same. Note that the prior odds of the three H_p ’s in the table should be written in terms of δ . Following Meester and Sjerps (2004), the prior odds of the first hypothesis in the table are $\delta/1-\delta$. The prior odds of the second hypothesis are $(\delta/2)/(1-\delta)$. The prior odds of the third hypothesis are $(\delta/2)/(1-(\delta/2))$. In each case, the posterior odds —

the result of multiplying the prior odds by the likelihood ratio — are the same: $R \times \delta/2(1-\delta)$. So despite differences in the likelihood ratio, the posterior odds of the different hypotheses are the same so long as the priors are appropriately related.

Meester and Sjerps (2004) recommend that each likelihood ratio should be accompanied by a tabular account of how a choice of prior odds (or prior probabilities) will impact the posterior odds, for a sensible range of priors (for a general discussion of this strategy, called ‘sensitivity analysis,’ see earlier discussion in 1.4). In this way, the impact of the likelihood ratio is made clear, no matter the hypotheses chosen. This strategy concedes that likelihood ratios are insufficiently informative, and that they should be combined with other information, such as a range of priors, to allow for an adequate assessment of the evidence.

3 Bayesian networks for legal applications

In legal cases, different lines of evidence may converge, such as two witnesses who testify that the defendant was seen at the crime scene, or they may diverge, such as a witness who asserts the defendant was seen at the crime scene while DNA testing shows no genetic match between the defendant and the scene. Another source of complexity is that the hypotheses put forward by the parties in a trial are often complex structures of statements. How can different statements, and their supporting evidence, be combined and the overall prosecutor’s case (or the defense’s case) be evaluated?

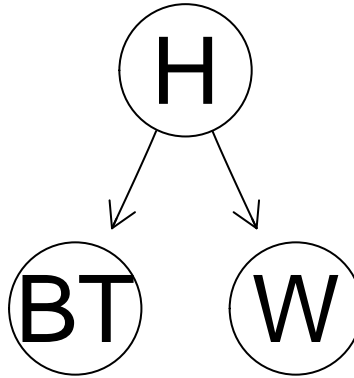
The probability of a hypothesis given multiple pieces of evidence can, in principle, be assessed by sequential applications of Bayes’ theorem. Consider, for example, a case in which the defendant faces two pieces of incriminating evidence: a DNA match and hair evidence found at the scene matching the defendant’s hair color. Assume—as is often done—that someone’s hair color is independent of someone’s genetic profile. Say the likelihood ratio of hair evidence is 40 and the likelihood ratio of the DNA match is 200, that is, it is 40 times more likely (and 200 times more likely) to find the evidence given the guilt hypothesis than the innocence hypothesis. If the prior odds of guilt to innocence are $1/1000$, the posterior odds would be $1/1000 \times 40 \times 200 = 8$.

These calculations are straightforward, but in more realistic cases, there will be complications. The parties at trial will often put forward several piecemeal claims that need to be combined together to form a theory of what happened. For example, the prosecutor may present eyewitness testimony to argue that the defendant ran away from the crime scene, along with documentary evidence as proof of a motive. The different piecemeal claims, each supported by distinct pieces of evidence, must be combined to form structured hypotheses about what happened. Since different claims and different pieces of evidence may depend on one another, direct calculations would soon become unmanageable. Fortunately, a tool exist to make the task easier: Bayesian networks. This section identifies guidelines for deploying Bayesian networks in the presentation, aggregation and evaluation of complex bodies of evidence and hypotheses.

3.1 Bayesian networks to the rescue

The idea that Bayesian networks can be used for probabilistic reasoning in legal fact-finding started gaining traction in late eighties (Friedman, 1986) and early nineties (Edwards, 1991). Two recent books on the topic with an emphasis on legal applications are (Fenton and Neil, 2018) and (Taroni et al., 2014). A Bayesian network comprises two components: first, a directed acyclic graph of relations of dependence (represented by arrows) between variables (represented by nodes); second, conditional probability tables. Consider the graphical component first. The graph is acyclic because the arrows connecting the nodes do not form

loops. As an illustration, let H be the claim that the suspect committed the murder, BT the presence of a blood type B match with a crime scene stain, and W the fact that an eye-witness observed the suspect near the scene around the time of the crime. The graphical component of the Bayesian network would look like this:



The *ancestors* of a node X are all those nodes from which we can reach X by following the arrows going forwards. The *parents* of a node X are those for which we can do this in one step. The *descendants* of X are all which can be reached from X by following the arrows going forward. The *children* are those for which we can do this in one step. In the example, H is the parent (and ancestor) of both W and BT , which are its children (and descendants). There are no non-parent ancestors or non-children descendants.

The variables, which are represented by nodes and are connected by arrows, stand in relation of probabilistic dependence. To describe these relations, the graphical model is accompanied by conditional probability tables. For parentless nodes such as H , the tables specify the prior probabilities of all their possible states. Assuming H stands for a binary random variable, with two possible states, the prior probabilities could be:

	Prior
H=murder	.01
H=no.murder	.99

The .01 figure for H=murder rests on the assumption that, absent any incriminating evidence, the defendant is unlikely to be guilty. For children nodes, the tables specify their conditional probability given combinations of their parents' states. If the variables are binary, an assignment of values for them could be:

	H=murder	H=no.murder
W=seen	.7	.4
W=not.seen	.3	.6

	H=murder	H=no.murder
BT=match	1	.063
BT=no.match	0	.937

According to the tables above, even if the defendant is not the culprit, the eyewitness testimony would still incriminate him with probability of .4, while the blood evidence with probability equal to only .063. The blood type frequency estimate is realistic (Lucy, 2013, 141), and so are the conditional probabilities for the eyewitness identification. As expected, eyewitness testimony is assumed to be less trustworthy than blood match evidence (but for complications about assessing eyewitness testimony, see Wixted and Wells, 2017; Urbaniak et al., 2020).

The three probability tables above are all that is needed to define the probability distribution. The tables do not specify probabilistic dependencies between nodes that are not in a relation of child/parent, such as BT and W . Since there is no arrow between them, nodes BT and W are assumed to be independent conditional on H , that is, $\Pr(W|H) = \Pr(W|H \wedge BT)$. This fact represents, as part of the structure of the network, the independence between eyewitness testimony and blood evidence. A generalization of this fact is the so-called Markov condition (see the textbook by Neapolitan (2004) and the supplement on Bayesian networks of the entry on Artificial Intelligence.⁴)

While the Bayesian network above—comprising a directed acyclic graph along with probability tables—is simple, a correct intuitive assessment of the probability of the hypothesis given the evidence is already challenging. Try to guess the probability that the defendant committed the murder ($H=\text{murder}$) given the following states of the evidence:

- The suspect’s blood type matches the crime stain but information about the witness is unavailable.
- The suspect’s blood type matches the crime stain but the witness says they did not see the suspect near the crime scene.
- The suspect’s blood type matches the crime stain and the witness says they saw the suspect near the crime scene.

Already at this level of complexity, calculations by hand become cumbersome. In contrast, software for Bayesian networks (see, for example, the R package `bnlearn` developed by Marco Scutari and described in Scutari and Denis, 2015) will easily give the following results:

	H=murder
BT=match,W=?	.138
BT=match,W=not.seen	.074
BT=match, W=seen	.219

Perhaps surprisingly the posterior probability of H is about .22 even when both pieces of evidence are incriminating ($BT=\text{match}$, $W=\text{seen}$).

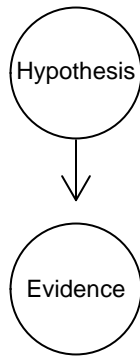
3.2 Idioms

Simple graphical patterns (called *idioms*) often recur while modeling the relationships between evidence and hypotheses. Complex graphical models can be created by combining these basic patterns in a modular way. Discussion of general methods for Bayesian network constructions can be found in (Neil et al., 2000), (Hepler et al., 2007) and general idioms are discussed in (Fenton et al., 2013). A few of the basic idioms are illustrated below.

The *evidence idiom* is the most basic graphical representation of the relation between a hypothesis and a piece of evidence:

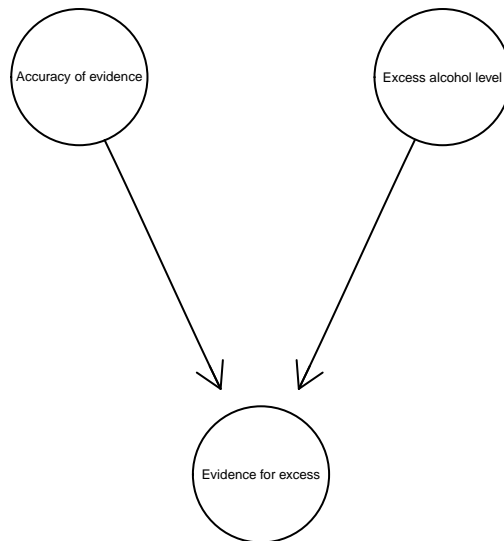
⁴Available at <https://plato.stanford.edu/entries/artificial-intelligence/bayesian-nets.html>

Reminder: Upload the R files with the calculations to a github folder if the entry gets published.



This directed graph suggests that the direction of influence—which could, but need not, be interpreted as causal influence—goes from hypothesis to evidence (though the probabilistic dependence goes both ways). The hypothesis node and the evidence node can be binary variables, such as ‘The defendant was the source of the crime scene traces’ (hypothesis) and ‘The defendant genetically matches the crime traces’ (evidence). But the variables need not be binary. The hypothesis node might take values from the range of 1–40, say the distance in meters from which the gun was shot, and the evidence node might be a continuous variable representing the density of gun shot residues (Taroni et al., 2014).

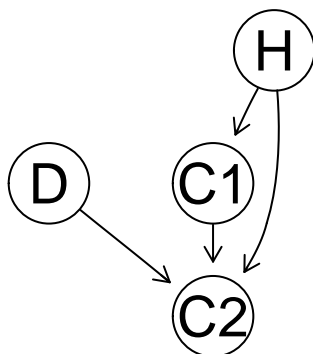
A more complex idiom, called the *evidence accuracy idiom*, consists of two arrows going into the evidence node (Bovens and Hartmann, 2004; Friedman, 1974). One incoming arrow comes from the hypothesis node and the other from the accuracy node. This idiom can be used to model, say, an alcohol test:



The directions of the arrows indicate that the accuracy of the evidence (accuracy node) and the alcohol level (hypothesis node) influence the outcome of the test (evidence node). The graphical model represents different sources of uncertainty. The uncertainty associated with the sensitivity and specificity of the test—that is, the probability that the tests reports excessive alcohol level when the level is excessive (sensitivity) and the probability that the test reports normal alcohol level when the level is normal (specificity)—is captured by the arrow going from the hypothesis node (Excess alcohol level) to the evidence node (Evidence for excess). Other sources of uncertainty comprise the possibility that the police officer lied about the test report or the possibility that the driver took medications which then affected

the alcohol level. These possibilities can be taken into consideration by adding an accuracy node (or multiple accuracy nodes, if each factor is kept separate from the other).

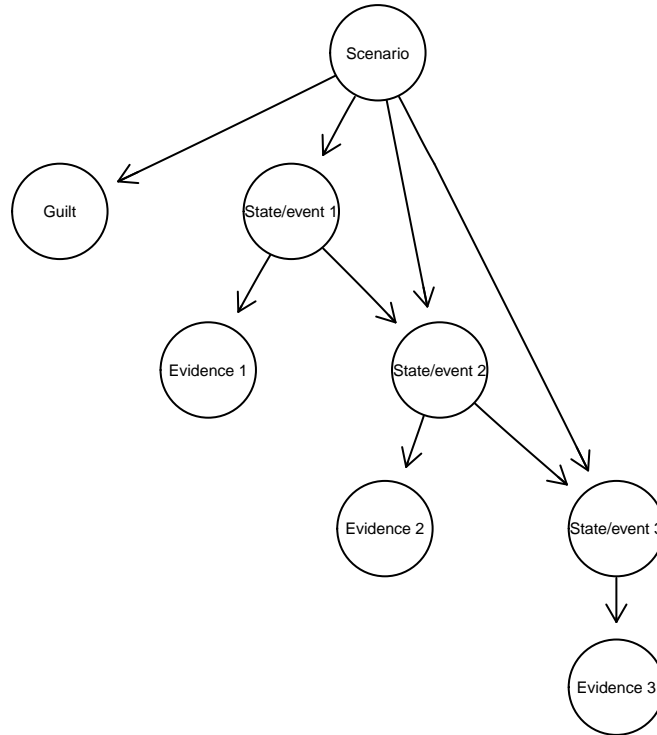
When multiple items of evidence depend on one another—as may happen in many legal cases—this situation is modelled by the *evidence dependency idiom*. Following an example by Fenton and Neil (2018), if one of two security cameras directed at the same location captured an image of someone who looks like the defendant but isn’t him, it is likely that the same person walked by the second camera, which also captured the same image. In such cases, presenting the second recording as independent from the first would lead to overestimating the strength of the evidence.



Node	Proposition
H	Defendant present at the scene
C1	Camera 1 captures image of a matching person
C2	Camera 2 captures image of a matching person
D	What cameras capture is dependent

The network structure is quite natural. The truth of the hypothesis, say, the defendant was present at the crime scene, influences whether the cameras capture an image of someone who looks like the defendant. However, if the two camera recordings are dependent on one another (for instance, they are directed at the same spot with a similar angle), the fact that the second camera captured the same image as the first does not make the hypothesis more likely once the first camera recording is known.

Finally, the *scenario idiom* can model complex hypotheses, consisting of a sequence of events organized in space and time (a scenario). A graphical model that uses the scenario idiom would consist of the following components: first, nodes for the states and events in the scenario, with each node linked to the supporting evidence; second, a separate scenario node that has states and events as its children; finally, a node corresponding to the ultimate hypothesis as a child of the scenario node. The graphical model would look like this (Vlek et al., 2014):

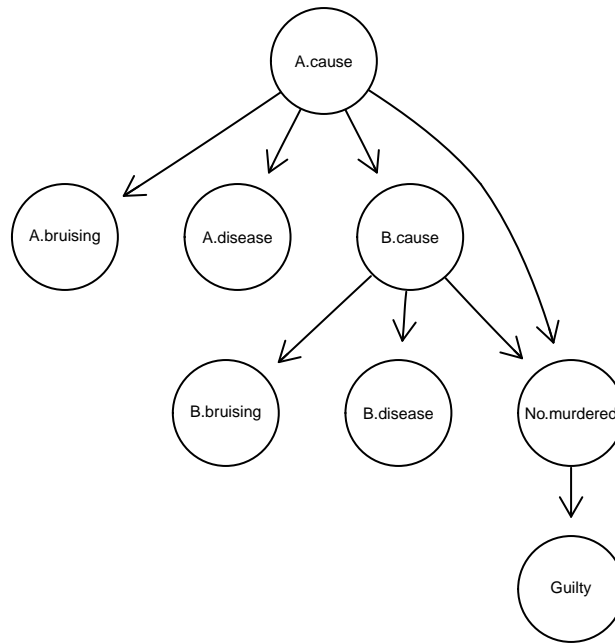


The scenario node unifies the different events and states. Because of this unifying role, increasing the probability of one part of the scenario (say State/event 2) will also increase the probability of the other parts (State/event 1 and State/event 3). This captures the fact that the different components of a scenario form an interconnected sequence of events.

A discussion of modelling crime scenarios by means of other graphical devices (called structural scenario spaces) mixed with probabilities can be found in the work of Shen et al. (2007), Bex (2011, 2015) and Verheij (2017). See also the survey by Di Bello and Verheij (2018). Dawid and Mortera (2018) give a treatment of scenarios in terms of Bayesian networks. Lacave and Díez (2002) show how Bayesian Network can be used to construct explanations.

3.3 Modeling an entire case

Kadane and Schum (2011) made one the first attempts to model an entire criminal case, Sacco & Vanzetti from 1920, using probabilistic graphs. More recently, Fenton and Neil (2018) constructed a Bayesian network for the Sally Clark case (discussed earlier in 1.3), reproduced below:



The arrows depict relationships of influence between variables. Whether Sally Clark’s sons, call them *A* and *B*, died by SIDS or murder (A.cause and B.cause) influences whether signs of disease (A.disease and B.disease) and bruising (A.brushing and B.brushing) were present. Since son A died first, whether A was murdered or died by SIDS (A.cause) influences how son B died (B.cause). How the sons died determines how many sons were murdered (No.murdered), and how many sons were murdered decides whether Sally Clark is guilty (guilty).

According to the calculation by Fenton and Neil (2018) (see their paper for details), the prior probability of Guilty = Yes should be .0789. After taking into account the incriminating evidence presented at trial, such as that there were signs of bruising but no signs of a preexisting disease affecting the children, the posterior probabilities are as follows:

Evidence (cumulative)	Pr(Clark guilty)
A bruising	.2887
A no signs of disease	.3093
B bruising	.6913
B no signs of disease	.7019

The incriminating evidence, combined, brings the probability of guilt from .0789 to .7019. This is a significant increase, but not quite enough for a conviction. If one wishes to perform sensitivity analysis—see earlier discussion in 1.4—by modifying some of the probabilities, this can be easily done. During the appeal trial, new evidence was discovered, in particular, evidence that son A was affected by a disease. Once this evidence is taken into account, the probability of guilt drops to .00459 (and if signs of disease were also present on B, the guilt probability would drop even further to .0009). For a general discussion on how to elicit probabilities, see (Renooij, 2001) and (van der Gaag et al., 2013).

4 Relevance

The preceding sections modeled evidence assessment, using Bayes’ theorem (Section 1), likelihood ratios (section 2) and Bayesian networks (Section 3). Evidence assessment, how-

ever, begins with a preliminary decision, the identification of relevant evidence. Once a piece of evidence is deemed relevant, the next step is to assess its strength (probative value, weight). This section discusses how probability theory helps to identify relevant evidence.

4.1 Likelihood ratios

The U.S. Federal Rules of Evidence define relevant evidence as evidence that has ‘any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence’ (rule 401). This definition is formulated in a probabilistic language. Legal probabilists interpret it using the likelihood ratio, a standard probabilistic measure of evidential relevance (Lempert, 1977; Lyon and Koehler, 1996; Aitken and Taroni, 2004; Aitken et al., 2010; Sullivan, 2019). The likelihood ratio (discussed in Section 2) is the probability of observing the evidence given the prosecutor’s or plaintiff’s hypothesis, divided by the probability of observing the same evidence given the defense’s hypothesis.

Let E be the evidence, H the prosecutor’s or plaintiff’s hypothesis, and H' the defense’s hypothesis. The likelihood ratio is defined as follows:

$$LR(E, H, H') = \frac{P(E|H)}{P(E|H')}$$

On the likelihood ratio interpretation, relevance depends on the choice of the competing hypotheses. A piece of evidence is relevant—in relation to a pair of hypotheses H and H' —provided the likelihood ratio $LR(E, H, H')$ is different from one and irrelevant otherwise. For example, the bloody knife found in the suspect’s home is relevant evidence in favor of the prosecutor’s hypothesis because we think it is far more likely to find such evidence if the suspect committed the crime (prosecutor’s hypothesis) than if he didn’t (defense hypothesis) (Finkelstein, 2009). In general, for values greater than one, $LR(E, H, H') > 1$, the evidence supports the prosecutor’s or plaintiff’s hypothesis H , and for values below one, $LR(E, H, H') < 1$, the evidence supports the defense’s hypothesis H' . If the evidence is equally likely under either hypothesis, $LR(E, H, H') = 1$, the evidence is irrelevant.

4.2 The Small Town Murder objection

This account of relevance has been challenged by cases in which the evidence is intuitively relevant and yet its likelihood ratio, arguably, equals one. Here is one problematic case:

Small Town Murder: A person accused of murder in a small town was seen driving to the small town at a time prior to the murder. The prosecution’s theory is that he was driving there to commit the murder. The defense theory is an alibi: he was driving to the town to visit his mother. The probability of this evidence if he is guilty equals that if he is innocent, and thus the likelihood ratio is 1 ... Yet, every judge in every trial courtroom of the country would admit it [as relevant evidence]. (The difficulty has been formulated by Ronald Allen, see the discussion in Park et al., 2010)

Counterexamples of this sort abound. Suppose a prisoner and two guards had an altercation because the prisoner refused to return a food tray. The prisoner had not received a package sent to him by his family and kept the tray in protest. According to the defense, the prisoner was attacked by the guards, but according to the prosecution, he attacked the guards. The information about the package sent to the prisoner and the withholding of the tray fails to favor either version of the facts, yet it is relevant evidence (Pardo, 2013).

It is true that if a piece of evidence E fits equally well with two competing hypotheses H and H' , then $P(E|H) = P(E|H')$ and thus $LR(E, H, H')$ will equal 1. But the likelihood ratio may change depending on the selection of hypotheses. Rule 401 makes clear that relevant evidence should have ‘any tendency to make the existence of *any fact that is of consequence* [emphasis ours] to the determination of the action more probable or less probable.’ So the range of hypotheses to compare should be broad. Just because the likelihood ratio equals one for a specific selection of H and H' , it does not follow that it equals one for *any* selection of H and H' which are of consequence to the determination of what happened. In *Small Town Murder*, whether the suspect was in town is of consequence for determining what happened (if he was not in town, he could not have committed the crime). The fact that he was seen driving is helpful information for establishing whether he was in town.

But if the range of hypotheses H and H' to compare in the likelihood ratio $LR(E, H, H')$ is broad, this may raise another concern. The choice of hypotheses needed to determine the relevance of an item of evidence might depend on other items of evidence, and so it might be difficult to determine relevance until one has heard all the evidence. This fact — Ronald Allen and Samuel Gross argue in (Park et al., 2010) — makes the probabilistic account of relevance impractical. In response, David Kaye points out that deciding whether a reasonable juror would find evidence E helpful requires only looking at what hypotheses or stories the juror would reasonably consider. Since the jurors will rely on several clues about which stories are reasonable, this task is computationally easier than going over all possible combinations of hypotheses (Park et al., 2010).

The problem with the paradoxes of relevance is that in complex situations there is no single likelihood ratio that corresponds to a piece of evidence. The problematic cases focus on a single likelihood ratio based on non-exclusive or non-exhaustive hypotheses. However, evidence can be relevant so long as it has a probabilistic impact on a pertinent sub-hypothesis, even without having a probabilistic impact on the prosecutor’s or defense’s ultimate hypotheses. When this happens, the evidence is relevant, in agreement with Rule 401 of the Federal Rules of Evidence. Bayesian networks (discussed in the preceding section) help to see how pieces of evidence can increase or decrease the probability of different sub-hypotheses (for more details, see de Zoete et al., 2019).

5 Standards of proof

After the evidence has been presented, examined and cross-examined at trial, trained judges or lay jurors must reach a decision (see Laudan, 2010, for a few caveats on what the decision should be about). The decision criterion is defined by law and consists of a standard of proof, also called the burden of persuasion. If the evidence against the defendant is sufficiently strong to meet the requisite proof standard, the defendant should be found liable. This section begins with a description of standards of proof in the law, then outlines a probabilistic account of standards of proof, and finally discusses some objections to this account.

5.1 Legal background

In criminal proceedings, the governing standard is ‘proof beyond a reasonable doubt.’ In civil cases, the standard is typically ‘preponderance of the evidence.’ The latter is less demanding than the former, so the same body of evidence may be enough to meet the preponderance standard, but not enough to meet the beyond a reasonable doubt standard. A vivid example of this difference is the 1995 trial of O.J. Simpson who was charged with murdering his wife. He was acquitted of the criminal charges, but when the family of the victim

brought a lawsuit against him, they prevailed. O.J. Simpson did not kill his wife according to the beyond a reasonable doubt standard, but he did according to the preponderance standard. An intermediate standard, called 'clear and convincing evidence', is sometimes used for civil proceedings in which the decision is particularly weighty, for example, a decision whether someone should be involuntarily committed to a hospital facility.

How to define standards of proof, or whether they should be even defined in the first place, remains contentious (Diamond, 1990; Newman, 1993; Horowitz and Kirkpatrick, 1996; Laudan, 2006; Walen, 2015). Judicial opinions offer different paraphrases, sometimes conflicting, of what these standards mean. The meaning of 'proof beyond a reasonable doubt' is the most controversial. It has been equated to 'moral certainty' or 'abiding conviction' (Commonwealth v. Webster, 59 Mass. 295, 320 (1850)), or to 'proof of such a convincing character that a reasonable person would not hesitate to rely and act upon it in the most important of his own affairs' (US Federal Jury Practice and Instructions, 12.10, at 354, 4th ed. 1987). But courts have also cautioned that there is no need to define the term because 'jurors know what is reasonable and are quite familiar with the meaning of doubt' and attempts to define it only 'muddy the water' (U.S. v. Glass, 846 F.2d 386, 1988).

Probability theory can bring conceptual clarity to an otherwise heterogeneous legal doctrine, or at least this is the position of legal probabilists.

5.2 Probability thresholds

Legal probabilists have proposed to interpret proof beyond a reasonable doubt as the requirement that the defendant's probability of guilt, given the evidence presented at trial, meet a threshold, say, $>.95$. Variations of this view are common (see, for example, Bernoulli, 1713; Laplace, 1814; Kaplan, 1968; Dekay, 1996; Kaye, 1979a; Laudan, 2006). This interpretation is, in some respects, plausible. From a legal standpoint, the requirement that guilt be established with high probability, still short of 1, accords with the principle that proof beyond a reasonable doubt is the most stringent standard of all but 'does not involve proof to an absolute certainty' and thus 'it is not proof beyond any doubt' (R v Lifchus, 1997, 3 SCR 320, 335).

Reliance on probabilistic ideas is even more explicit in the standard 'preponderance of the evidence'—also called 'balance of probabilities'—which governs decisions in civil disputes. This standard can be interpreted as the requirement that the plaintiff—the party making the complaint against the defendant—establish its version of the facts with greater than probability of .5. The .5 threshold, as opposed to a more stringent threshold of .95 for criminal cases, reflects the fact that preponderance is less demanding than proof beyond a reasonable doubt. The intermediate standard 'clear and convincing evidence' is more stringent than the preponderance standard but not as stringent as the beyond a reasonable doubt standard. Since it lies in between the other two, it can be interpreted as the requirement that the plaintiff establish its versions of the facts with, say, probability at the level of .75-.8.

Some worry that a mechanical application of numerical thresholds would undermine the humanizing function of trial decision-making. As Tribe (1971) put it, 'induced by the persuasive force of formulas and the precision of decimal points to perceive themselves as performing a largely mechanical and automatic role, few jurors ... could be relied upon to recall, let alone to perform, [their] humanizing function.' Thresholds, however, can vary depending on the costs and benefits at stake in each case (see later discussion). So they need not be applied mechanically without considering the individual circumstances (Hedden and Colyvan, 2019). Furthermore, if jurors are numerically literate, they should not lose sight of their humanizing function as they would no longer be intimidated by numbers. So the worry suggests the need to ensure that jurors are numerically literate, not dispensing

with probabilistic thresholds altogether.

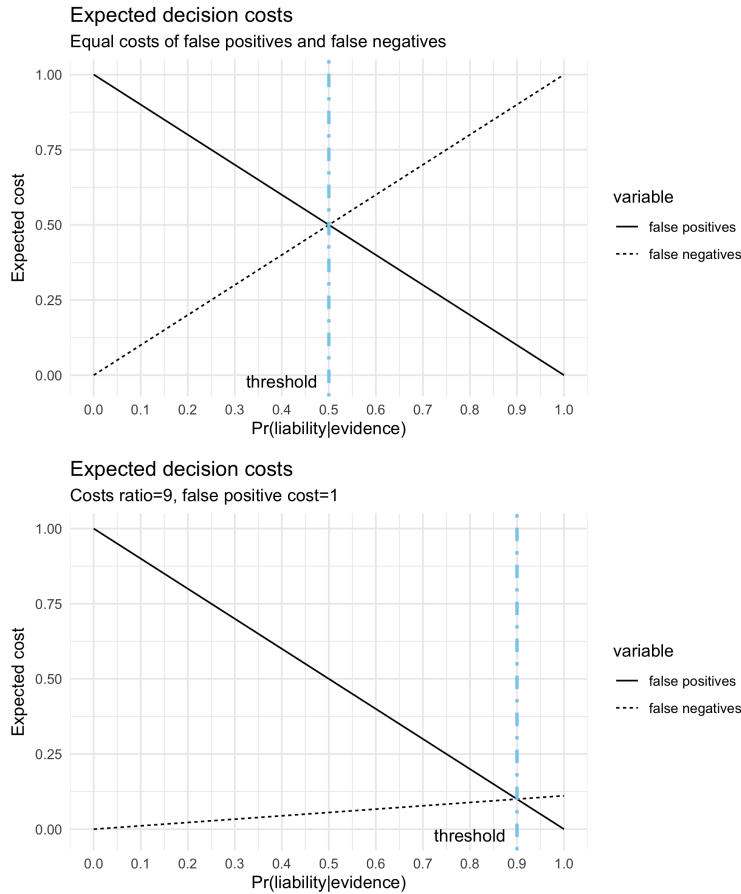
Even if numerical thresholds cannot be used in the daily business of trial proceedings, they can still serve as theoretical concepts for understanding the role of proof standards in the justice system, such as regulating the relative frequency of false positive and false negative decisions or minimizing expected costs. A more stringent threshold will decrease the number of false positives (say false convictions) at the cost of increasing the number of false negatives (say false acquittals), and a less stringent threshold will increase the number of false positives while decreasing the number of false negatives. This trade-off has been described, among others, by Justice Harlan in his concurring opinion in *re Winship*, 397 U.S. 358, 397 (1970). As shown below, the formal apparatus of probability theory, in combination with expected utility theory, can make this point more precise.

5.3 Minimizing expected costs

Expected utility theory recommends agents to take the course of action that, among the available alternatives, maximizes expected utility. On this view, the standard of proof is met whenever the expected utility (or cost) of a decision against the defendant (say, a conviction) is greater (or lower) than the expected utility (or cost) of a decision in favor of the defendant (say, an acquittal) (Kaplan, 1968; Dekay, 1996; Hamer, 2004). Let $c(CI)$ be the cost of convicting a factually innocent defendant and $c(AG)$ the cost of acquitting a factually guilty defendant. For a conviction to be justified, the expected cost of convicting an innocent—that is, $c(CI)$ discounted by the probability of innocence $[1 - \Pr(G|E)]$ —must be lower than the expected cost of acquitting a guilty defendant—that is, $c(AG)$ discounted by the probability of guilt $\Pr(G|E)$. This holds just in case

$$\frac{\Pr(G|E)}{1 - \Pr(G|E)} > \frac{c(CI)}{c(AG)}.$$

This inequality captures how high the probability of guilt or civil liability must be to justify a verdict against the defendant. When the cost ratio $\frac{c(CI)}{c(AG)}$ is set to 9—which might be appropriate in a criminal case since convicting an innocent is often considered more harmful than acquitting a guilty defendant (however, see Laudan, 2016)—the inequality holds only if $\Pr(G|E)$ meets the .9 threshold. The same analysis *mutatis mutandis* applies to civil cases in which mistaken decisions comprise mistaken attributions of liability (false positives) and mistaken failures to attribute liability (false negatives). If the cost ratio is one—as might be appropriate in a civil case in which false positives and false negatives are equally harmful—the inequality holds only if the probability that the defendant is liable meets the .5 threshold.



This analysis only considers the costs of mistaken decisions, but leaves out the benefits associated with correct decisions. More comprehensive analyses would consider both, but the basic insight would remain the same. Trial decision-making is viewed as one instrument among others for maximizing overall social welfare (Posner, 1973). On this account of proof standards, the stringency of the threshold depends on costs and benefits, and thus different cases may require different thresholds. Cases in which the charge is more serious than others—say, murder compared to petty theft—may require higher thresholds so long as the cost of a mistaken decision against the defendant is more significant. Whether or not standards of proof should vary in this way is debated (Kaplow, 2012; Picinali, 2013).

The standard ‘proof beyond a reasonable doubt’ is often paired with the Blackstone ratio, the principle that it is better that ten guilty defendants go free rather than even just one innocent be convicted. The exact ratio is in fact a matter of controversy (Volkov, 1997). It is tempting to think that, say, the .9 threshold guarantees a 1:9 ratio between false convictions and false acquittals. But this would be hasty for at least two reasons. First, probabilistic thresholds affect the expected rate of mistaken decisions. The actual rate may deviate from its expected value (Kaye, 1999). Second, if the threshold is .9, *at most* 10% of decision against defendants are expected to be mistaken (false convictions) and *at most* 90% of the decisions in favor of the defendant are expected to be mistaken (false acquittals). The exact ratio will depend on the probabilities assigned to defendants and how they are distributed (Allen, 2014). What can be said, in general, is that the threshold that minimizes the expected rate of incorrect decisions overall, no matter the underlying distribution, lies at .5 (see Kaye, 1982, 1999; Cheng and Pardo, 2015, for a proof).

5.4 Alternatives to probabilistic thresholds

There exist several theoretical alternatives to the probabilistic interpretation of proof standards in the scholarly literature. Pennington and Hastie (1991, 1993) have proposed the *story model* according to which judges and jurors first make sense of the evidence by constructing stories of what happened, and then select the best story on the basis of multiple criteria, such as coherence, fit with the evidence and completeness. Along similar lines, Pardo and Allen (2008) argue that the version of the facts that best explains the evidence should prevail in a court of law. For a discussion of inference to the best explanation in legal reasoning, see Schwartz and Sober (2019); Hastie (2019); Ho (2019); Nance (2019).

Another approach is due to Gordon et al. (2007) and Prakken and Sartor (2009) who view the trial as a place in which arguments and counterarguments confront one another. The party that has the best arguments, all things considered, should prevail. On this view, probability estimates can themselves be the target of objections and counterarguments. Along these lines, Stein (2008) argues that, in order to warrant a verdict against the defendant, the evidence should have withstood objections and counterarguments, not merely supporting a high probability.

Philosophers and legal theorists have also levelled distinctively epistemological critiques. Ho (2008) and Haack (2014a) hold that degrees of epistemic warrant for a claim, which depend on multiple factors – such as the extent to which the evidence supports the claim and it is comprehensive – cannot be equated to probabilities. Gardiner (2019) argues that standards of proof should rule out all error possibilities that are relevant and these need not coincide with error possibilities that are probable. Finally, some epistemologists argue that a probabilistic belief, no matter how high, is not enough to warrant knowledge, and knowledge should be the standard for trial verdicts (Duff et al., 2007; Littlejohn, 2017; Blome-Tillmann, 2017; Levanon, 2019; Moss, 2020).

Scholars and commentators have also voiced more specific objections that need not invalidate the probabilistic framework but rather call for refinements. Nance (2016) argues that the evidence on which to base a trial decision should be reasonably complete—it should be all the evidence that one would reasonably expect to see from a conscientious investigation of the facts. A similar argument can be found in (Davidson and Pargetter, 1987). Arguably, probability-based decision thresholds can accommodate these considerations, for example, by lowering the probability of civil or criminal liability whenever the body of evidence is one-sided or incomplete (Kaye, 1979b, 1986; Friedman, 1996). Another strategy is to give a probability-based account of the notion of completeness of the evidence and other seemingly non-probabilistic criteria (Urbaniak, 2018).

There are a plethora of other objections. The puzzles of naked statistical evidence and the conjunction paradox are two of the most widely debated in the literature. These and other objections are examined in the sections that follow.

6 Naked statistical evidence

The puzzles of naked statistical evidence consist of hypothetical scenarios in which the probability of the defendant's civil or criminal liability, given the evidence, is above the requisite threshold. Yet many have the intuition that the defendant should not be found liable. The question is how to justify this intuition despite the fact that the probability of liability meets the threshold. The puzzles of naked statistical evidence concern the question of *sufficiency*, namely, whether a body of evidence is enough to meet the proof standard applicable in a case. They are not about whether some evidence should be *admissible* at trial (on the distinction, see Picinali, 2016),

6.1 Blue Bus, Gatecrasher, Prisoner

Blue Bus (Tribe, 1971). Mrs. Brown is run down by a bus. It is common knowledge that 80% of the buses in town are owned by Blue Bus, and the remaining 20% by Red Bus. There was no witness to the accident except Brown who is, however, color-blind. Since Blue Bus owns 80% of the buses in town, it is 80% likely that Brown was run down by a bus operated by Blue Bus, well above the 50% threshold for civil liability. Yet, merely presenting the 80% naked statistics should not be sufficient for Brown to prevail in a civil lawsuit against Blue Bus.

Gatecrasher (Cohen, 1977). It is known that 499 people paid for admission to a rodeo, but the total number of spectators was 1000. Suppose no paper tickets were issued and no witness could identify those who paid. For any spectator picked at random, it is more than 50% likely that they did not pay for admission. But even if this probability is above the .5 threshold for civil liability, it would be odd that the rodeo organizers could win a lawsuit against any of the spectators simply by presenting the 449-to-1000 naked statistics.

Prisoner (Nesson, 1979). 100 prisoners are exercising in a prison yard. Suddenly 99 of them attack and kill the only guard on duty. One prisoner played no role whatsoever in the assault. These are the undisputed facts in the case, and there is no further information about what happened. If a prisoner is picked at random, his probability of guilt would be as high as .99. Yet the intuition is that this cannot be enough to establish guilt beyond a reasonable doubt.

These scenarios are like lottery cases in which the probability of a proposition such as ‘my ticket is a loser’ is high, yet intuitively the proposition cannot count as knowledge (see e.g. Harman, 1968; Nelkin, 2000; Hawthorne, 2004; Lawlor, 2013; Ebert et al., 2018). The evidence in these scenarios—in particular, *Gatecrasher* and *Prisoner*—does not single out an individual specifically, but applies to any member of a group. Just as in lottery cases any ticket is very likely to lose, anyone who attended the rodeo or any prisoner in the yard is very likely to be liable. In this sense, naked statistical evidence is sometimes contrasted with individualized or case-specific evidence such as trace evidence or eyewitness testimony (Williams, 1979; Thomson, 1986; Colyvan et al., 2001). The distinction is contested, however. Any form of evidence, after all, relies on a categorization that places the defendant in a class with others, being the class of those who have such-and-such facial features or were in such-and-such a place (Saks and Kidd, 1980; Schoeman, 1987; Wright, 1988; Schauer, 2003; Harcourt, 2006). Tillers (1997, 2005) notes that it is not always objectionable to base inferences about the behavior of an individual on the behavior of others, for example, membership in a gang or in the Ku Klux Klan can be indicative of someone’s beliefs.

Still, there is intuitive resistance against verdicts of liability based on naked statistics, but this resistance is less pronounced for verdicts based on more traditional forms of evidence, such as trace or eyewitness testimony. The asymmetry might just be an artifact of history, since the testimony of an honest witness has been the bedrock of the Anglo-American trial system. However, the resistance toward naked statistical evidence along with the preference for other forms of evidence has also been verified empirically (Wells, 1992; Niedermeier et al., 1999; Arkes et al., 2012) and is not limited to the legal context (Sykes and Johnson, 1999; Friedman and Turri, 2015; Ebert et al., 2018).

Some scholars have expressed reservations about the relevance of hypothetical scenarios involving naked statistical evidence. Since these scenarios are removed from trial practice, they might be an unreliable guide for theorizing about the trial (Schmalbeck, 1986; Dant, 1988; Allen and Leiter, 2001). These scenarios, however, are partly modeled on real cases. For example, *Blue Bus* is modeled on *Smith v. Rapid Transit, Inc.* 317 Mas. 469 (1945). The

hypothetical also bears a similarity with the landmark case *Sindell v. Abbott Laboratories*, 26 Cal. 3d 588 (1980) in which different companies marketed the same drug that was later shown to have caused cancer to a particular individual. Since the drug was sold by multiple companies, it was impossible to determine which company was responsible. Statistics about the market share of the two companies were used to attribute liability in absence of better, more individualized evidence.

6.2 Cold-hits

Legal scholars have drawn parallels between naked statistical evidence and DNA evidence in cold-hit cases (Roth, 2010). The peculiarity of cold-hit cases is that the defendant is identified through a database search of several different genetic profiles, and as a consequence, the evidence in cold-hit cases consists almost exclusively of a DNA match between the crime traces and the defendant (see earlier discussion in 2.2). The match—as is customary—is complemented by a statistical estimate of the frequency of the profile, say, one in one hundred million people share the same matching profile. Given the largely statistical nature of the evidence, cold-hit cases can be seen as realistic examples of scenarios such as *Prisoner* or *Gatecrasher*. But whether we should think about cold-hit cases in this way is contested. Some authors place naked statistical evidence and cold-hit matches on a par (Smith, 2017) and others do not (Enoch et al., 2012; Enoch and Fisher, 2015).

Some appellate courts in the United States have ruled that cold-hit matches, even though they are uncorroborated by other evidence, constitute sufficient ground for verdicts of criminal liability (Malcom, 2008). Judge Hardwick, for example, writes that if ‘DNA material is found in a location, quantity, and type inconsistent with casual contact and there is one in one quintillion likelihood that some else was the source of the material, the evidence is legally sufficient to support a guilty verdict’ (*Missouri v. Abdelmalik*, 273 S.W.3d, 61, 66 (Mo. Ct. App. 2008)). Such pronouncements by appellate court lend support to the view that the statistics underlying cold-hit DNA matches are unlike naked statistical evidence (Cheng and Nunn, 2016; Di Bello, 2019).

6.3 Revisionist responses

The puzzles of naked statistical evidence are some of the most difficult for legal probabilism. They directly challenge the claim that a high probability should suffice for a judgment of criminal or civil liability. One line of response that legal probabilists can pursue is to recommend revising our intuitions in hypothetical cases and questioning their relevance as a guide for theorizing about standards of proof. Legal probabilists can argue that the preference for traditional forms of evidence over statistical evidence is an unwarranted bias (Laudan, 2006; Papineau, 2019). They can point out that research in psychology and cognitive science has shown that eyewitness testimony and fingerprint evidence are often unreliable (Simons and Chabris, 1999), prone to manipulation (Loftus, 1996), and influenced by subjective considerations and matters of context (Dror et al., 2006; Zabel, 2005). Relying on statistical evidence, on the other hand, should improve the overall accuracy of trial decisions (Koehler and Shavero, 1990). Legal probabilists can also argue that the puzzles of naked statistical evidence are confined to hypotheticals, and our judgments about statistical evidence may well differ in more realistic cases (Hedden and Colyvan, 2019; Ross, 2020).

Few have defended such revisionist views, however. Even when it advances one variant or another of legal probabilism, the literature has predominately tried to vindicate the intuitive difference between naked statistical evidence and other forms of evidence. What follows is a discussion of some of the proposals in the literature, focusing specifically on

the probabilistic ones. An examination of the non-probabilistic solutions to the paradoxes of naked statistical evidence falls outside the scope of this entry (for critical surveys, see Redmayne, 2008; Gardiner, 2018; Pardo, 2019).

6.4 Non-revisionist responses

Below are five non-revisionist moves legal probabilists can make in response to the paradoxes of naked statistical evidence.

First, legal probabilists can deny that in scenarios such as *Gatecrasher* the probability of liability is high enough to meet the required threshold. Kaye (1979c) argues that whenever there is no other evidence besides naked statistical evidence, this is not enough for the plaintiff to win because the evidence is suspiciously thin. This circumstance should lead the fact-finders to lower the probability of liability below the requisite threshold. Along similar lines, Nance (2016) argues that when the evidence presented at trial is incomplete—that is, evidence that one would reasonably expect to see at trial is missing—the defendant should not be found liable. The difficulty with this strategy is that it might work in scenarios such as *Blue Bus* in which the paucity of the evidence could well be the plaintiff's fault. But it is unclear whether this strategy works for scenarios such as *Gatecrasher* or *Prisoner* in which the paucity of the evidence is a characteristic of the scenarios themselves, not anyone's fault.

Second, legal probabilists can appeal to the reference class problem (Colyvan et al., 2001). An individual may fall under different reference classes. If 3% of those who smoke have lung cancer and 0.1% of those who exercise regularly have lung cancer, what about someone who smokes and exercises regularly? In *Gatecrasher*, it is arbitrary to single out the group of people at the stadium and not the group of people with a history of trespassing. There is no clear reason why one or the other reference class was chosen. The choice of another reference class could have led to a different conclusion about the defendant's probability of liability. This approach has also been endorsed by scholars who are decidedly opposed to legal probabilism (Allen and Pardo, 2007). Critics of this approach note that the reference class problem affects any evidence-based judgment. In assessing the strength of any piece of evidence, different reference classes can be used, say the class of witnesses who give detailed descriptions of what they saw or the class of witnesses who are nervous (Redmayne, 2008).

Third, legal probabilists can point out how probabilistic claims based on naked statistical evidence are not resilient enough in light of possible countervailing evidence (on the notion of resilience and stability of belief, see Skyrms, 1980; Leitgeb, 2014). If an eyewitness were to assert that the prisoner did not participate in the riot, the probability that he would be guilty of killing the guard should be lowered significantly. Presumably, trial verdicts cannot be so volatile. They should aim to some degree of stability even in light of possible further evidence (Bolinger, 2020). A problem with this approach is that more evidence could always be found that would change one's probabilistic assessment. A further problem is that the puzzles of naked statistical evidence are cases in which no further evidence is—or could possibly be—available. But adding the assurance that the probabilistic claims could not be revised does not make the naked statistics less problematic.

Bolinger reference to be updated. Paper is still forthcoming.

Fourth, legal probabilists can insist that verdicts solely based on naked statistics do not promote the goal of expected utility maximization. This is not so much because the naked statistics are bad evidence, but rather, because reliance on them may have a number of unintended costs, such as suboptimal allocation of the burden of error or lack of deterrence. In *Blue Bus*, for example, a verdict against the company with the largest market share might create a perverse economic incentive against larger companies (Posner, 1973). It is unclear how this explanation could be extended to other cases such as *Gatecrasher* or *Prisoner*, and there are even variants of *Blue Bus* not susceptible to this objection (Wells, 1992). Alter-

natively, Dahlman (2020) argues that verdicts based on naked statistical evidence do not provide any added incentive for lawful conduct because they would not make a distinction between lawful and unlawful behaviour. (On deterrence and naked statistics, see also Enoch et al., 2012; Enoch and Fisher, 2015).

Finally, another line of response for legal probabilists is to concede that the paradoxes of naked statistical evidence demonstrate the inadequacy of simple probabilistic thresholds as proof standards (Urbaniak, 2019). Instead of the posterior probability of liability, a number of scholars have focused on the likelihood ratio $P(E|H)/P(E|H')$. Their argument is that, even though naked statistical evidence can support a high posterior probability of liability, the likelihood ratio of this evidence equals one because it could be presented against a defendant no matter what the defendant did. If so, naked statistical evidence should have no evidential value (Cheng, 2012; Sullivan, 2019). However, Dahlman (2020) has criticized this argument noting that, under plausible assumptions, the likelihood ratio of naked statistical evidence is significantly greater than one. Di Bello (2019) argues that in cases such as *Prisoner* and *Gatecrasher* the likelihood ratio could take a range of different values depending on the background information that is taken into account. The likelihood ratio of naked statistical evidence in those scenarios is therefore neither one nor greater than one, but strictly speaking unknown (for a critique of this argument, see Urbaniak et al., 2020).

7 Further objections

Aside from the paradoxes of naked statistical evidence, the conjunction paradox is one of the most widely discussed objections against legal probabilism. This section examines this paradox together with a few other objections. Many of these objections can be traced back to the seminal work of Cohen (1977) who also levelled criticisms against Bayesian epistemology more generally (for further discussion, see Earman, 1992; Bovens and Hartmann, 2004; Bradley, 2015).⁵

7.1 The difficulty about conjunction

Suppose the plaintiff is to prove two separate claims, *A* and *B*, according to the governing standard of proof, say, preponderance of the evidence (which the legal probabilist may interpret as the requirement that the facts be established with greater probability than .5). If the plaintiff has proven each claim with probability .7, the burden of proof should be met. And yet, if the two claims are independent, the probability of their conjunction is only $.7 \times .7 = .49$, below the required threshold. Arguably, common law systems subscribe to a conjunction principle which states that if *A* and *B* are established according to the governing standard of proof, so is their conjunction. Probability theory, the criticism goes, cannot capture this principle. This is the so-called conjunction paradox or difficulty about conjunction. It was originally formulated by Cohen (1977) and has enjoyed great popularity every since (Allen, 1986; Stein, 2005; Allen and Stein, 2013; Haack, 2014b; Schwartz and Sober, 2017; Allen and Pardo, 2019).

Without rejecting the conjunction principle of the common law, legal probabilists can respond in a few different ways. Dawid (1987) argues that the difficulty about conjunction disappears if evidential support is modeled probabilistically by likelihood ratios instead of posterior probabilities. He writes that ‘suitably measured, the support supplied by the conjunction of several independent testimonies exceeds that supplied by any of its constituents.’ Although the original paradox pertains to posterior probabilities of liability, Dawid argues that the paradox does not arise for likelihood ratios. Garbolino (2014) also recommends

⁵SEP entry: <https://plato.stanford.edu/entries/epistemology-bayesian/>

switching to likelihood ratios. Yet the conjunction paradox still arises for likelihood ratios when the items of evidence, say a and b , are assessed relative to a composite hypothesis, say $A \wedge B$. Suppose a and b provide positive support for A and B , respectively. Urbaniak (2019) shows that the combined likelihood ratio $\Pr(a \wedge b | A \wedge B) / \Pr(a \wedge b | \neg(A \wedge B))$ can be lower than the individual likelihood ratios $\Pr(a | A) / \Pr(a | \neg A)$ and $\Pr(b | B) / \Pr(b | \neg B)$.

Cheng (2012) provides another probability-based solution to the conjunction paradox. He argues that the standard of proof in civil cases should require that the plaintiff's hypothesis be comparatively more probable on the evidence than the defendant's hypothesis. On this account, the probability of $A \wedge B$ given the overall evidence in a case should be compared with the probability of the alternative hypotheses given the same evidence. The alternatives to be considered are as follows: $A \wedge \neg B$, $\neg A \wedge B$, and $\neg A \wedge \neg B$. Given suitable assumptions, the probabilities of these alternatives will fall below the probability of $A \wedge B$ provided A and B , individually, are supported by the evidence. So whenever the standard of proof is met for individual claims A and B , the standard should also be met for the composite claim $A \wedge B$. Kaplow (2014) advances a similar argument. Urbaniak (2019) points out that Cheng splits the defense hypothesis into three sub-cases, $A \wedge \neg B$, $\neg A \wedge B$, and $\neg A \wedge \neg B$, but does not consider the alternative $\neg(A \wedge B)$. The probability of $\neg(A \wedge B)$ may actually exceed that of $A \wedge B$. If $\neg(A \wedge B)$ is the alternative hypothesis, the standard might not be met for the composite claim $A \wedge B$, even when the standard is met for individual claims A and B (for another critique of Cheng's approach, see Allen and Stein, 2013).

Finally, legal probabilists can pursue a holistic approach in response to the conjunction paradox. This solution has been defended by opponents of legal probabilism (Allen, 1986; Allen and Pardo, 2019), but can equally be adopted by legal probabilists. Instead of assessing the posterior probabilities of individual claims given individual pieces of evidence, the holistic approach recommends assessing the probability of the overall composite claim, say $A \wedge B$, in light of all the evidence available (Di Bello, 2019). Along these lines, Hedden and Colyvan (2019) suggest that we should adopt the holistic approach even if the legal formulation of proof standards is ambiguous between a holistic reading and a piecemeal reading in which the support for each individual claim is assessed separately. Bayesian networks (see the earlier discussion in Section 3) can be of service here (de Zoete and Sjerps, 2018; de Zoete, 2017; Neil et al., 2019).

7.2 Cohen's other objections

Cohen (1977) levelled a few other objections against legal probabilism. They are less well-known than the paradoxes of naked statistical evidence or the conjunction paradox, but are still worth examining.

Completeness The statement $\Pr(\neg H | E) = 1 - \Pr(H | E)$ is a theorem of the probability calculus. If the probability of H given E is low, the probability of $\neg H$ given the same evidence must be high. This fact may seem to create evidence from ignorance. If E is meager evidence of H (that is, $\Pr(H | E)$ is low), it must be strong evidence for the negation of H (that is, $\Pr(\neg H | E)$ is high). This seems wrong. Intuitively, some evidence can weakly support both a hypothesis and its negation. For example, suppose one has heard a rumor that the defendant spent the night at the bar when the crime was committed. The rumor is weak evidence for the claim that the defendant spent the night at the bar. It does not follow that the rumor is strong evidence that the defendant did *not* spend the night at the bar. Evidence may actually have no bearing whatsoever on a hypothesis. Probability seems unable to capture this fact, or at least this is the objection.

This difficulty motivated the development of a non-classical theory of probability and evidential support by Dempster (1968) and Shafer (1976). Legal probabilists, however, need

not reject classical probability theory. They can respond that the difficulty just described arises only because one is inclined to measure the strength of evidence in terms of the posterior probability $\Pr(H|E)$, rather than by means of the likelihood ratio (on this distinction, see earlier in Section 2). If E weakly supports H —that is, $\Pr(E|H)/\Pr(E|\neg H)$ is barely above one—it does not follow that E strongly supports $\neg H$. In fact, it rather follows that E weakly disfavors $\neg H$, because $\Pr(E|\neg H)/\Pr(E|H)$ will be slightly below one.

Corroboration When two or more independent witnesses testify to the truth of the same proposition, and their stories are relatively unlikely, the probability of the proposition in question should increase significantly. This phenomenon of ‘confidence boost’ is known as corroboration. In case of circumstantial evidence, the analogous phenomenon is called convergence. Cohen (1977) argues that no probabilistic measure of evidential support captures this phenomenon. He examines different probabilistic proposals—Boole’s formula (Boole, 1857), Ekelöf’s principle (Ekelöf, 1964), a formula due to Lambert and Kruskal (1988)—and finds them inadequate. He argues that the confidence boost that is expected from corroboration is not adequately captured by any of these proposals.

More recently, better probabilistic accounts of corroboration have been developed (Taroni et al., 2014; Robertson et al., 2016; Fenton and Neil, 2018). A recurrent theme in this line of work is that corroboration can be accounted for in probabilistic terms by multiplying the likelihood ratios of the individual pieces of evidence. The thought is that the result of multiplying the likelihood ratios of different pieces of evidence combined exceeds the likelihood ratios associated with the individual pieces of evidence (see general discussion in Bovens and Hartmann, 2004, chp. 5). Cohen, however, insisted that the confidence boost due to corroboration should be large, and the large boost is not reflected as a result of multiplying the likelihood ratios. What needs further exploring is the size of the confidence boost and the features of the evidence that affect the boost. Urbaniak and Janda (2019) offer a detailed discussion of Cohen’s objections and candidates for a solution.

7.3 The problem of priors

Another objection often levelled against legal probabilism is the problem of priors. This problem emerges as one sets out to assess the posterior probability of a hypothesis given the available evidence, $\Pr(H|E)$. Recall the formula of Bayes’ theorem:

$$\Pr(H|E) = \text{BayesianFactor}(H, E) \times \Pr(H).$$

To carry out the calculations, Bayes’ theorem requires as a starting point the prior probability of the hypothesis, $\Pr(H)$, irrespective of the evidence. The correct assessment of this prior probability is by no means obvious. Different strategies have been proposed.

First, the prior probability can be equated to $1/k$ where k is the number of alternative equiprobable hypotheses. If there are k possible hypotheses and no one is more probable than the other, it is natural to assign $1/k$ to each hypothesis. This approach, however, would render prior probabilities quite sensitive to the choice of hypotheses and thus potentially arbitrary. In addition, this approach is particularly unsuitable for criminal cases. If two hypotheses are ‘the defendant is guilty’ and ‘the defendant is innocent’, the prior probability of each would be 50%. Defendants in criminal cases, however, should be presumed innocent until proven guilty. Prior probability of guilt at the level of .5 seems excessive. The presumption of innocence—a procedural protection afforded to all defendants in many countries—should require that the prior probability of guilt be set to a small value (Friedman et al., 1995). But it is unclear how low that value should be. Would .001 be sufficiently low or should it be .000001? All that can be said, perhaps, is that in criminal cases the prior probability of guilt should be extremely low (Friedman, 2000).

Alternatively, the prior probability can be equated to $1/n$ where n is the number of total possible suspects or wrongdoers who could have committed the crime or civil wrong under dispute at trial. This, too, is a plausible proposal. Since someone must have committed the wrong, absent any further evidence anyone could have committed it, so $1/n$ is a reasonable starting point. But this proposal also quickly runs into difficulties. In some cases, whether a wrong was committed by anyone can itself be disputed. Or there might be cases in which an illegal act was certainly committed and the defendant took part in it, but it is not clear what illegal act it was, say murder or manslaughter.

To avoid some of the difficulties described above, more recent models rely on relevant background information, for example, geographical information about people's opportunities to commit crimes (Fenton et al., 2019). But even if these models are successful in giving well-informed assessments of prior probabilities, a deeper difficulty lingers. That is, any assessment of prior probabilities, no matter how it is done, is likely to violate existing normative requirements of the trial system (Dahlman, 2017a; Engel, 2012; Schweizer, 2013). If the assessment of prior probabilities relies on demographic information, people who belong to certain demographic groups will be regarded as having a higher prior probability of committing a wrong than others. But if some people's priors are higher than other people's priors, it will be easier to convict or find liable those who are assigned higher priors, even if the evidence against them is the same as the evidence against those assigned lower priors. This outcome can be seen as unfair, especially in criminal cases, since it exposes to a higher risk of conviction those innocents who are assigned a higher prior probability because of the demographic group they belong to (Di Bello and O'Neil, 2020).

Perhaps, as some have suggested, legal probabilists should do away with prior probabilities and rely on likelihood ratios instead as a guide for trial decisions (Sullivan, 2019). Another strategy for avoiding the problem of priors is to consider an interval of values and see the extent to which different possible priors affect the posterior probability (Finkelstein and Fairley, 1970), as discussed in Section 1.4.

7.4 The reference class problem

Another challenge to legal probabilism is the reference class problem. The reference class problem, originally formulated by Venn (1866), arises because the same event may belong to multiple reference classes in which the frequencies of the event in question are different. A common approach, given among others by Reichenbach (1949, 374), is to rely on 'the narrowest class for which reliable statistics can be compiled.' This may work in some cases provided reliable statistics are available. But what if someone belongs to different classes that are equally narrow?

Consider the case of Charles Shonubi, a Nigerian citizen working in New Jersey, who was arrested on December 10, 1991 at JFK Airport in New York for smuggling heroin into the United States. He was found carrying 103 balloons in his gastrointestinal tract containing 427.4 grams of heroin. During the sentencing proceedings, the prosecutor argued that since Shonubi made seven trips between the United States and Nigeria prior to his arrest, he smuggled a larger total quantity of heroin than 427.4 grams. The prosecution offered data on amounts of heroin seized from 117 Nigerian drug smugglers who were arrested at JFK airport between September 1, 1990 and December 10, 1991. The expert for the prosecutor, Dr. Boyum, calculated that it was 99% likely that Shonubi smuggled at least 2090.2 grams in total before his final trip (US. v. Shonubi 895 F. Supp 460, E.D.N.Y. 1995).

Shonubi was a member of the reference class 'people found carrying heroin while traveling between Nigeria and New York' but also the class 'toll collectors at the George Washington bridge.' Why rely on the former and not the latter to make inferences about how much heroin Shonubi smuggled into the United States? (Colyvan et al., 2001). What follows ex-

amines the specific difficulties that the reference class problem poses for legal probabilism and how legal probablists might respond.

The challenge Allen and Pardo (2007) argue that the reference class problem poses a challenge for legal probabilism and more specifically for probabilistic measures of evidentiary strength such as likelihood ratios (see earlier in Section 2). The problem is that the same piece of evidence may be assigned different likelihood ratios depending on the reference class chosen. For example, the denominator in the likelihood ratio associated with a DNA match is the probability of a match given that a random person, unrelated to the crime, is the source. This probability depends on the frequency of the profile in a select population. But which reference population should one choose? Since nothing in the natural world picks out one reference class over another—the argument goes—the likelihood ratio would be an arbitrary measure of evidentiary strength.

It is tempting to dismiss this challenge by noting that expert witnesses work with multiple reference classes and entertain plausible ranges of values (Nance, 2007). In fact, relying on multiple reference classes is customary in the assessment of DNA evidence. In *Darling v. State*, 808 So. 2d 145 (Fla. 2002), for example, a Polish woman living in Orlando was sexually assaulted and killed. The DNA expert testified about multiple random match probabilities, relying on frequencies about African-Americans, Caucasians and Southeastern Hispanics from the Miami area. Since the perpetrator could have belonged to any of these ethnic groups, the groups considered by the expert were all relevant under different scenarios of what could have happened.

Unlike expert witnesses, appellate courts often prefer that only one reference class be considered. In another case, Michael Pizarro, who matched the crime traces, was convicted of raping and suffocating his 13-year-old-half-sister (*People v. Pizarro*, 110 Cal.App.4th 530, 2003). The FBI analyst testified at trial that the likelihood of finding another unrelated Hispanic individual with the same genetic profile was approximately $1/250,000$. Since the race of the perpetrator was not known, Pizarro appealed arguing that the DNA evidence was inadmissible. The appellate court sided with Pizarro and objected to the presentation of frequency estimates for the Hispanic population as well as frequencies for any other racial or ethnic groups. The court wrote: 'It does not matter how many Hispanics, Caucasians, Blacks, or Native Americans resemble the perpetrator if the perpetrator is actually Asian.'

The uneasiness that appellate courts display when expert witnesses testify about multiple reference classes is understandable. Perhaps, the reference class most favorable to the defendant should be selected, giving the accused the benefit of the doubt. This might be appropriate in some cases. But suppose the random match probability associated with a DNA match is 1 in 10 for people in group A, while it is 1 in 100 million for people in group B. Always going for the reference class that is most favorable to the defendant will in some cases weaken the incriminating force of DNA matches more than necessary.

Relevance and critical questions Legal probabilists have formulated different criteria for identifying the most appropriate reference class. Franklin (2011) is optimistic. On his approach, the most appropriate reference class for drawing an inference about an event or outcome *B* is the class defined by the intersection of all the features that are relevant to *B*. Relevance is measured statistically given the available data as the co-variation of *B* with the feature in question. Co-variation will be measured using appropriate statistical criteria, such as the correlation coefficient between two variables. For instance, in the Shonubi case, features such as being Nigerian, drug courier, travelling toward JFK, were all relevant for making an inference about the total amount of drugs Shonubi carried. Other features for which data were available, such as being a toll collector at the George Washington Bridge,

were not relevant. The optimism of Franklin's approach, however, is dimmed by how deeply the reference class problem cuts (for more details, see Hájek, 2007).

Instead of focusing on relevance only, the choice of the most appropriate reference class may also include a mix of statistical (or epistemic) criteria as well as non-epistemic criteria. Dahlman (2017b) proposes a list of critical questions, such as: Is the reference class heterogeneous or homogeneous? Is it robust? Does it put people in the reference class at an unfair disadvantage? The first two questions are epistemic, but the third is not. If people in certain ethnic or socio-economic groups engage in criminal acts more often than others, relying on ethnic or socio-economic reference classes may heighten society's stigma and prejudice toward members of these groups. Relying on these reference classes should therefore be avoided, not because they are irrelevant but because they are prejudicial.

Model selection A more general way of thinking about the reference class problem is the model selection problem in that the choice of a reference class is a particular case of the problem of how to model a phenomenon (Cheng, 2009). The model should capture the data to some extent, but not overfit the data. Random variation in the data should not be built into the model. In statistics, different criteria for model selection exist, most notably the Akaike's Information Criterion (AIC), which is a popular measure of the trade-off between model's fit with the data and its complexity. Consider the Shonubi case again. In order to predict the total amount of heroin Shonubi transported, a simple linear model could be used, where each reference class comes with an empirically established multiplier β for the expected total amount of drugs carried based on how much drugs one carried on a single trip. Picking too generic a class (say, 'airline passenger') would lower the empirical adequacy of the model, but relying on too narrow a class would incorporate random noise. A model using the class 'toll collector' would clearly perform (in terms of statistical measures such as AIC) worse than, say a model based on the class 'Nigerian drug courier at JFK.'

Colyvan and Regan (2007) argue that the reference class problem is a form of model uncertainty and can be thought of as uncertainty about the particular statistical model that should be employed. Model uncertainty is evident in the Shonubi case as alternative models were discussed. The first model consisted in multiplying the amount which Shonubi was found carrying when he was arrested by the number of trips he made between the Nigeria and New York. Admittedly, this model was too simplistic. The expert witness for the prosecution, Dr. Boyum, developed a second model. It was based on the DEA data and consisted in the simulation of 100,000 possible series of seven trips by re-sampling sets of seven net weights from the 117 known cases.

Boyum's model was criticized by another expert in the case, Dr. Finkelstein, who complained that the model did not take into consideration systematic differences between trips. Presumably, smugglers will tend to carry larger quantities as they become more experienced. When they carry larger quantities, it should be more likely they would be apprehended. If the trip effect holds, the data would mostly be about advanced smugglers who tend to carry more than beginners. No empirical evidence for or against the trip effect theory could be found. Information on the number of previous trips was missing from the data so an appropriate regression model for the trip effect could not be developed. On the other hand, Judge Weinstein argued that beginner smugglers do practice swallowing with grapes, and thus the learning curve should not be excessively steep. In addition, beginners are more likely to be apprehended than advanced smugglers. If so, the data would not be biased by the trip effect. Interestingly, information on Nigerian drug smuggling practices undermines the trip effect theory: the drug cartel did not waste money on sending half-filled drug mules, but rather made them practice swallowing balloons weeks prior to their first trip (Treaser, 1992; Wren, 1999).

Statistical analyses of the evidence in Shonubi were published after the case was decided.

These analyses took into account other potential sources of error, such as outliers and biased data. Gastwirth et al. (2000) showed that, because of sensitivity to outliers, the inference that the total amount was above 3000 grams was unstable. However, even looking at the data in the light most favourable to the defendant, the total amount of drugs should be above the 1000 gram threshold (Izenman, 2000b; Gastwirth et al., 2000; Izenman, 2000a).

Besides Shonubi, many other cases raise the reference class or model selection in their own way. Just to list a few:

- *Vuyanich v. Republic National Bank* involved race and sex discrimination allegation. The case involved nine different expert witnesses deploying various statistical analyses. The case ended with a 127-page opinion.
- *E.E.O.C. v. Federal Reserve Bank of Richmond* was a similar case, where the appropriateness of various methods for group comparisons and the choice of level of aggregation for the analysis of employment data was at play.
- *Gulf South Insulation v. U.S. Consumer Product Safety Commission* was related to banning the use of urea-formaldehyde foam insulation. The difficulty lied in the choice of a risk assessment model to calculate the risk of increased incidence of cancer.

These cases are interesting as well as complicated. They are discussed in some detail in (Fienberg, 1989).

8 Relation to other entries in SEP

This entry is related to others in SEP, such as entries in formal epistemology, e.g., “Bayes’ Theorem”, “Bayesian Epistemology”, “Confirmation”, “Abduction”, and entries in philosophy of law and legal epistemology, such as “The Legal Concept of Evidence” and “Interpretation and Coherence in Legal Reasoning.”

References

- Aitken, C., Roberts, P., and Jackson, G. (2010). Fundamentals of probability and statistical evidence in criminal proceedings (Practitioner Guide No. 1), Guidance for judges, lawyers, forensic scientists and expert witnesses. *Royal Statistical Society’s Working Group on Statistics and the Law*.
- Aitken, C. and Taroni, F. (2004). *Statistics and the evaluation of evidence for forensic scientists*, volume 16. Wiley Online Library.
- Allen, R. and Pardo, M. (2007). The problematic value of mathematical models of evidence. *The Journal of Legal Studies*, 36(1):107–140.
- Allen, R. and Pardo, M. (2019). Relative plausibility and its critics. *The International Journal of Evidence & Proof*, 23(1-2):5–59.
- Allen, R. J. (1986). A reconceptualization of civil trials. *Boston University Law Review*, 66:401–437.
- Allen, R. J. (2014). Burdens of proof. *Law, Probability and Risk*, 13:195–219.
- Allen, R. J. and Leiter, B. (2001). Naturalized epistemology and the law of evidence. *Virginia Law Review*, 87(8):1491–1550.
- Allen, R. J. and Stein, A. (2013). Evidence, probability and the burden of proof. *Arizona Law Journal*, 55:557–602.
- Arkes, H. R., Shoots-Reinhard, B. L., and Mayes, R. S. (2012). Disjunction between probability and verdict in juror decision making. *Journal of Behavioral Decision Making*, 25(3):276–294.

- Balding, D. J. (2002). The DNA Database Search Controversy. *Biometrics*, 58(1):241–244.
- Balding, D. J. (2005). *Weight-of-evidence for forensic DNA profiles*. John Wiley and Sons.
- Balding, D. J. and Donnelly, P. (1996). Evaluating DNA Profile Evidence When the Suspect Is Identified Through a Database Search. *Journal of Forensic Sciences*, 41(4):13961J.
- Barker, M. J. (2017). Connecting applied and theoretical Bayesian epistemology: Data relevance, pragmatics, and the legal case of Sally Clark. *Journal of Applied Philosophy*, 34(2):242–262.
- Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, 76:169–217.
- Bernoulli, J. (1713). *Ars conjectandi*.
- Bex, F. (2015). An integrated theory of causal stories and evidential arguments. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law - ICAIL '15*, pages 13–22, San Diego, California. ACM Press.
- Bex, F. J. (2011). *Arguments, Stories and Criminal Evidence: A Formal Hybrid Theory*. Number v. 92 in Law and Philosophy Library. Springer, Dordrecht ; New York.
- Biedermann, A., Hicks, T., Taroni, F., Champod, C., and Aitken, C. (2014). On the use of the likelihood ratio for forensic evaluation: Response to Fenton et al. *Science & Justice*, 54(4):316–318.
- Blome-Tillmann, M. (2017). ‘More likely than not’ — Knowledge first and the role of bare statistical evidence in courts of law. In Carter, A., Gordon, E., and Jarvi, B., editors, *Knowledge First—Approaches in Epistemology and Mind*, pages 278–292. Oxford University Press.
- Bolinger, R. J. (2020). Explaining the justificatory asymmetry between statistical and individualized evidence. In *The Social Epistemology of Legal Trials*. Routledge. [forthcoming].
- Boole, G. (1857). On the application of the theory of probabilities to the question of the combination of testimonies or judgments. *Transactions of the Royal Society of Edinburgh*, 21(04):597–653.
- Bovens, L. and Hartmann, S. (2004). *Bayesian epistemology*. Oxford University Press.
- Bradley, D. (2015). *A Critical Introduction to Formal Epistemology*. Bloomsbury Publishing.
- Calabresi, G. (1961). Some thoughts on risk distribution and the law of torts. *Yale Law Journal*, 70:499–553.
- Cheng, E. (2012). Reconceptualizing the burden of proof. *Yale LJ*, 122:1254.
- Cheng, E. and Nunn, G. (2016). DNA, Blue Bus, and phase changes. *The International Journal of Evidence and Proof*, 20(2):112–120.
- Cheng, E. and Pardo, M. S. (2015). Accuracy, optimality and the preponderance standard. *Law, Probability and Risk*, 14(3):193–212.
- Cheng, E. K. (2009). A practical solution to the reference class problem. *Columbia Law Review*, 109:2081.
- Childers, T. (2013). *Philosophy and probability*. Oxford University Press.
- Cohen, J. (1977). *The Probable and the Provable*. Oxford University Press.
- Cohen, J. L. (1981). Subjective probability and the paradox of the Gatecrasher. *Arizona State Law Journal*, pages 627–634.
- Colyvan, M. and Regan, H. M. (2007). Legal decisions and the reference class problem. *The International Journal of Evidence & Proof*, 11(4):274–285.
- Colyvan, M., Regan, H. M., and Ferson, S. (2001). Is it a crime to belong to a reference class. *Journal of Political Philosophy*, 9(2):168–181.
- Cook, R., Evett, I., Jackson, G., and Jones, P. (1998). A hierarchy of propositions: Deciding which level to address in casework. *Science & Justice*, 38(4):231–239.
- Cullison, A. D. (1969). Probability analysis of judicial fact-finding: A preliminary outline of the subjective approach. *Toledo Law Review*, 1:538–598.
- Dahlman, C. (2017a). Determining the base rate for guilt. *Law, Probability and Risk*, 17(1):15–28.

- Dahlman, C. (2017b). Unacceptable generalizations in arguments on legal evidence. *Argumentation*, 31(1):83–99.
- Dahlman, C. (2020). Naked statistical evidence and incentives for lawful conduct. *International Journal of Evidence and Proof*, 24(2):162–179.
- Dant, M. (1988). Gambling on the truth: the use of purely statistical evidence as a basis for civil liability. *Columbia Journal of Law and Social Problems*, 22:31–70.
- Davidson, B. and Pargetter, R. (1987). Guilt beyond reasonable doubt. *Australasian Journal of Philosophy*, 65(2):182–187.
- Dawid, A. P. (1987). The difficulty about conjunction. *The Statistician*, pages 91–97.
- Dawid, A. P. (1994). The island problem: coherent use of identification evidence. In Freeman, P. and Smith, A., editors, *Aspects of Uncertainty: a Tribute to D. V. Lindley*, pages 159–170. John Wiley & Sons, New York.
- Dawid, A. P. (2002). Bayes’s theorem and weighing evidence by juries. In *Bayes’s Theorem*, volume 113, pages 71–90. Oxford University Press.
- Dawid, A. P. and Mortera, J. (1996). Coherent Analysis of Forensic Identification Evidence. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(2):425–443.
- Dawid, A. P. and Mortera, J. (2018). Graphical models for forensic analysis. In *Handbook of Graphical Models*, pages 491–514. CRC Press.
- de Condorcet, M. (1785). *Essai sur l’application de l’analyse à la probabilité des décisions rendues à la pluralité des voix*.
- de Zoete, J. C. (2017). Combining Forensic Evidence. *Science & Justice*.
- de Zoete, J. C., Fenton, N., Noguchi, T., and Lagnado, D. (2019). Resolving the so-called “probabilistic paradoxes in legal reasoning” with Bayesian networks. *Science & Justice*, 59(4):367–379.
- de Zoete, J. C. and Sjerps, M. (2018). Combining multiple pieces of evidence using a lower bound for the LR. *Law, Probability and Risk*, 17(2):163–178.
- Dekay, M. L. (1996). The difference between Blackstone-like error ratios and probabilistic standards of proof. *Law and Social Inquiry*, 21:95–132.
- Dempster, A. P. (1968). A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232.
- Di Bello, M. (2019). Trial by Statistics: Is a High Probability of Guilt Enough to Convict? *Mind*, 128(512):1045–1084.
- Di Bello, M. and O’Neil, C. (2020). Profile evidence, fairness and the risk of mistaken convictions. *Ethics*, 130(2):147–178.
- Di Bello, M. and Verheij, B. (2018). Evidential reasoning. In *Handbook of Legal Reasoning and Argumentation*, pages 447–493. Springer.
- Diamond, H. A. (1990). Reasonable doubt: To define, or not to define. *Columbia Law Review*, 90(6):1716–1736.
- Donnelly, P. (1995). Nonindependence of matches at different loci in DNA profiles: Quantifying the effect of close relatives on the match probability. *Heredity*, 75(1):26–34.
- Donnelly, P. and Friedman, R. D. (1999). DNA Database Searches and the Legal Consumption of Scientific Evidence. *Michigan Law Review*, 97(4):931.
- Dror, I. E., Charlton, D., and Peron, A. E. (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International*, 156:74–78.
- Duff, A., Farmer, L., Marshall, S., and Tadros, V. (2007). *The Trial on Trial (vol 3): Towards a Normative Theory of the Criminal Trial*. Hart Publishing.
- Earman, J. (1992). *Bayes or bust? A critical examination of Bayesian confirmation theory*. Cambridge: MIT Press.
- Ebert, P. A., Smith, M., and Durbach, I. (2018). Lottery judgments: A philosophical and experimental study. *Philosophical Psychology*, 31(1):110–138.
- Edwards, W. (1991). Influence diagrams, bayesian imperialism, and the collins case: an

- appeal to reason. *Cardozo Law Review*, 13:1025–1074.
- Eggleston, R. (1978). *Evidence, proof and probability*, volume 2. Weidenfeld and Nicolson London.
- Ekelöf, P. (1964). Free evaluation of evidence. *Scandinavian Studies in Law*, 8:47–66.
- ENFSI (2015). *Guidelines for Evaluative Reporting in Forensic Sciences*.
- Engel, C. (2012). Neglect the Base Rate: It's the Law! *Preprints of the Max Planck Institute for Research on Collective Goods*, 23. Preprints of the Max Planck Institute for Research on Collective Goods.
- Enoch, D. and Fisher, T. (2015). Sense and sensitivity: Epistemic and instrumental approaches to statistical evidence. *Stan. L. Rev.*, 67:557–611.
- Enoch, D., Spectre, L., and Fisher, T. (2012). Statistical evidence, sensitivity, and the legal value of knowledge. *Philosophy and Public Affairs*, 40(3):197–224.
- Evet, I. (1987). On meaningful questions: a two-trace transfer problem. *Journal of the Forensic Science Society*, 27(6):375–381.
- Evet, I., Jackson, G., and Lambert, J. (2000). More on the hierarchy of propositions: Exploring the distinction between explanations and propositions. *Science & Justice*, 40(1):3–10.
- Fenton, N., Berger, D., Lagnado, D., Neil, M., and Hsu, A. (2014). When 'neutral' evidence still has probative value (with implications from the Barry George Case). *Science & Justice*, 54(4):274–287.
- Fenton, N., Lagnado, D., Dahlman, C., and Neil, M. (2019). The opportunity prior: A proof-based prior for criminal cases. *Law, Probability and Risk*, page [online first].
- Fenton, N. and Neil, M. (2018). *Risk Assessment and Decision Analysis with Bayesian Networks*. Chapman and Hall.
- Fenton, N., Neil, M., and Lagnado, D. A. (2013). A General Structure for Legal Arguments About Evidence Using Bayesian Networks. *Cognitive Science*, 37(1):61–102.
- Ferguson, A. G. (2020). Big data prosecutor and brady. *UCLA Law Review*, 67(4).
- Fienberg, S., editor (1989). *Evolving Role of Statistical Assessments as Evidence in the Courts*. Springer.
- Finkelstein, M. (2009). *Basic concepts of probability and statistics in the law*. Springer.
- Finkelstein, M. O. and Fairley, W. B. (1970). A Bayesian approach to identification evidence. *Harvard Law Review*, 83(3):489–517.
- Foreman, L., Champod, C., Evet, I. W., Lambert, J., Pope, S., et al. (2003). Interpreting dna evidence: A review. *International Statistical Review*, 71(3):473–495.
- Franklin, J. (2011). Objective bayesian conceptualisation of proof and reference class problems. *Sydney Law Review*, 33:545–561.
- Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy*, 71:5–19.
- Friedman, O. and Turri, J. (2015). Is probabilistic evidence a source of knowledge? *Cognitive Science*, 39(5):1062–1080.
- Friedman, R. D. (1986). A diagrammatic approach to evidence. *Boston University Law Review*, 66:571–622.
- Friedman, R. D. (1987). Route analysys of credibility and hearsay. *The Yale Law Journal*, 97(4):667–742.
- Friedman, R. D. (1996). Assessing evidence. *Michigan Law Review*, 94:1810–1838.
- Friedman, R. D. (2000). A presumption of innocence, not of even odds. *Stanford Law Review*, 52(4):873–887.
- Friedman, R. D., Allen, R. J., Balding, D. J., Donnelly, P., and Kaye, D. H. (1995). Probability and proof in State v. Skipper: An internet exchange. *Jurimetrics*, 35(3):277–310.
- Garbolino, P. (2014). *Probabilità e logica della prova*. Giuffrè Editore.
- Gardiner, G. (2018). Legal burdens of proof and statistical evidence. In Coady, D. and Chase, J., editors, *Routledge Handbook of Applied Epistemology*. Routledge.

- Gardiner, G. (2019). The reasonable and the relevant: Legal standards of proof. *Philosophy and Public Affairs*, 47(3):288–318.
- Gastwirth, J. L., Freidlin, B., and Miao, W. (2000). The shonubi case as an example of the legal system's failure to appreciate statistical evidence. In *Statistical Science in the Courtroom*, pages 405–413. Springer.
- Gillies, D. (2000). *Philosophical theories of probability*. Psychology Press.
- Gordon, T. F., Prakken, H., and Walton, D. (2007). The Carneades model of argument and burden of proof. *Artificial Intelligence*, 171(10-15):875–896.
- Haack, S. (2014a). *Evidence Matters: Science, Proof, and Truth in the Law*. Cambridge University Press.
- Haack, S. (2014b). Legal probabilism: an epistemological dissent. In *Haack (2014a)*, pages 47–77.
- Hacking, I. (1990). *The Taming of Chance*. Cambridge University Press.
- Hájek, A. (2007). The reference class problem is your problem too. *Synthese*, 156(3):563–585.
- Hájek, A. (2012). Interpretations of probability. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2012 edition.
- Hamer, D. (2004). Probabilistic standards of proof, their complements and the errors that are expected to flow from them. *University of New England Law Journal*, 1(1):71–107.
- Harcourt, B. E. (2006). *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. University Of Chicago Press.
- Harcourt, B. E. (2018). The systems fallacy: A genealogy and critique of public policy and cost-benefit analysis. *Journal of Legal Studies*, 47(2):419–447.
- Harman, G. (1968). Knowledge, inference, and explanation. *American Philosophical Quarterly*, 5(3):164–173.
- Hastie, R. (2019). The case for relative plausibility theory: Promising, but insufficient. *The International Journal of Evidence & Proof*, 23(1-2):134–140.
- Hawthorne, J. (2004). *Knowledge and lotteries*. Oxford University Press.
- Hedden, B. and Colyvan, M. (2019). Legal probabilism: A qualified defence. *Journal of Political Philosophy*, 27(4):448–468.
- Hepler, A. B., Dawid, A. P., and Leucari, V. (2007). Object-oriented graphical representations of complex patterns of evidence. *Law, Probability and Risk*, 6(1-4):275–293.
- Ho, H. L. (2008). *A philosophy of evidence law: Justice in the search for truth*. Oxford University Press.
- Ho, H. L. (2019). How plausible is the relative plausibility theory of proof? *The International Journal of Evidence & Proof*, 23(1-2):191–197.
- Horowitz, I. A. and Kirkpatrick, L. C. (1996). A concept in search of a definition: The effect of reasonable doubt instructions on certainty of guilt standards and jury verdicts. *Law and Human Behaviour*, 20(6):655–670.
- Izenman, A. J. (2000a). Assessing the statistical evidence in the shonubi case. In *Statistical science in the courtroom*, pages 415–443. Springer.
- Izenman, A. J. (2000b). Introduction to two views on the shonubi case. In *Statistical Science in the Courtroom*, pages 393–403. Springer.
- Kadane, J. B. and Schum, D. A. (2011). *A probabilistic analysis of the Sacco and Vanzetti evidence*. John Wiley & Sons.
- Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4):237–251.
- Kaplan, J. (1968). Decision theory and the fact-finding process. *Stanford Law Review*, 20(6):1065–1092.
- Kaplow, L. (2012). Burden of proof. *Yale Law Journal*, 121(4):738–1013.
- Kaplow, L. (2014). Likelihood ratio tests and legal decision rules. *American Law and Economics Review*, 16(1):1–39.

- Kaye, D. H. (1979a). The laws of probability and the law of the land. *The University of Chicago Law Review*, 47(1):34–56.
- Kaye, D. H. (1979b). The paradox of the Gatecrasher and other stories. *The Arizona State Law Journal*, pages 101–110.
- Kaye, D. H. (1979c). Probability theory meets *res ipsa loquitur*. *Michigan Law Review*, 77(6):1456–1484.
- Kaye, D. H. (1980). Mathematical models and legal realities: Some comments on the poisson model of jury behavior. *Connecticut Law Review*, 13(1):1–15.
- Kaye, D. H. (1982). The limits of the preponderance of the evidence standard: Justifiably naked statistical evidence and multiple causation. *Law & Social Inquiry*, 7(2):487–516.
- Kaye, D. H. (1986). Do we need a calculus of weight to understand proof beyond a reasonable doubt? *Boston University Law Review*, 66:657–672.
- Kaye, D. H. (1999). Clarifying the burden of persuasion: What Bayesian rules do and not do. *International Commentary on Evidence*, 3:1–28.
- Kaye, D. H. and Sensabaugh, G. F. (2000). Reference guide on DNA evidence. In *Reference Manual on Scientific Evidence* (2dn ed.), pages 576–585. Federal Judicial Center.
- Koehler, J. J. (1996). The base rate fallacy reconsidered: descriptive, normative, and methodological challanges. *Behavioral and Brain Sciences*, 19(1):1–53.
- Koehler, J. J. and Shaviro, D. N. (1990). Veridical verdicts: Increasing verdict accuracy through the use of overtly probabilistic evidence and methods. *Cornell Law Review*, 75:247–279.
- Kruskal, W. (1988). Miracles and Statistics: The Casual Assumption of Independence. *Journal of the American Statistical Association*, 83(404):929–940.
- Lacave, C. and Díez, F. J. (2002). A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review*, 17(2):107–127.
- Laplace, P. (1814). *Essai philosophique sur les probabilités*.
- Laudan, L. (2006). *Truth, error, and criminal law: an essay in legal epistemology*. Cambridge University Press.
- Laudan, L. (2010). Need verdicts come in pairs? *The International Journal of Evidence & Proof*, 14(1):1–24.
- Laudan, L. (2016). *The Law's Flaws: Rethinking Trials and Errors?* Law and Society. College Publications.
- Lawlor, K. (2013). *Assurance: An Austinian view of Knowledge and Knowledge Claims*. Oxford University Press.
- Leitgeb, H. (2014). The stability theory of belief. *Philosophical Review*, 123(2):131–171.
- Lempert, R. O. (1977). Modeling relevance. *Michigan Law Review*, 75:1021–1057.
- Levanon, L. (2019). Statistical evidence, assertions and responsibility. *Modern Law Review*, 82(2):269–292.
- Littlejohn, C. (2017). Truth, knowledge, and the standard of proof in criminal law. *Synthese*, pages <https://doi.org/10.1007/s11229-017-1608-4>.
- Loftus, E. F. (1996). *Eyewitness Testimony (Revised Edition)*. Harvard University Press.
- Lucy, D. (2013). *Introduction to statistics for forensic scientists*. John Wiley & Sons.
- Lyon, T. D. and Koehler, J. J. (1996). Relevance ratio: Evaluating the probative value of expert testimony in child sexual abuse cases. *Cornell Law Review*, 82:43.
- Malcom, B. G. (2008). Convictions predicated on DNA evidence alone: How reliable evidence became infallible. *Columbia Law Review*, 38(2):313–338.
- Mayo, D. (2018). *Statistical Inference as Severe Testing*. Cambridge University Press.
- Meester, R. and Sjerps, M. (2004). Why the effect of prior odds should accompany the likelihood ratio when reporting DNA evidence. *Law, Probability and Risk*, 3(1):51–62.
- Mellor, D. H. (2004). *Probability: A philosophical introduction*. Routledge.
- Moss, S. (2020). Knowledge and legal proof. *Oxford Studies in Epistemology*. [forthcoming].

- Nance, D. A. (2007). The reference class problem and mathematical models of inference. *International Journal of Evidence and Proof*, 11(4):259–273.
- Nance, D. A. (2016). *The Burdens of Proof: Discriminatory Power, Weight of Evidence, and Tenacity of Belief*. Cambridge University Press.
- Nance, D. A. (2019). The limitations of relative plausibility theory. *The International Journal of Evidence & Proof*, 23(1-2):154–160.
- National Research Council (1996). *The evaluation of forensic DNA evidence* [NRC II]. Committee on DNA technology in Forensic Science, National Research Council.
- Neapolitan, R. (2004). *Learning Bayesian Networks*. Pearson Prentice Hall.
- Neil, M., Fenton, N., Lagnado, D., and Gill, R. D. (2019). Modelling competing legal arguments using bayesian model comparison and averaging. *Artificial Intelligence and Law*.
- Neil, M., Fenton, N., and Nielson, L. (2000). Building large-scale Bayesian Networks. *The Knowledge Engineering Review*, 15(3):257–284.
- Nelkin, D. N. (2000). The lottery paradox, knowledge, and rationality. *The Philosophical Review*, 109(3):373–409.
- Nesson, C. R. (1979). Reasonable doubt and permissive inferences: The value of complexity. *Harvard Law Review*, 92(6):1187–1225.
- Newman, J. O. (1993). Beyond “reasonable doubt”. *New York University Law Review*, 68(5):979–1002.
- Niedermeier, K. E., Kerr, N. L., and Messeé, L. A. (1999). Jurors’ use of naked statistical evidence: Exploring bases and implications of the Wells effect. *Journal of Personality and Social Psychology*, 76(4):533–542.
- Nitzan, S. (2010). *Collective Preference and Choice*. Cambridge University Press.
- Papineau, D. (2019). The disvalue of knowledge. *Synthese*, <https://doi.org/10.1007/s11229-019-02434-z>.
- Pardo, M. S. (2013). The Nature and Purpose of Evidence Theory. *Vanderbilt Law Review*, 66:547–613.
- Pardo, M. S. (2019). The paradoxes of legal proof: A critical guide. *Boston University Law Review*, 99(1):233–290.
- Pardo, M. S. and Allen, R. J. (2008). Judicial proof and the best explanation. *Law and Philosophy*, 27(3):223–268.
- Park, R. C., Tillers, P., Moss, F. C., Risinger, D. M., Kaye, D. H., Allen, R. J., Gross, S. R., Hay, B. L., Pardo, M. S., and Kirgis, P. F. (2010). Bayes Wars Redivivus – An Exchange. *International Commentary on Evidence*, 8(1).
- Pennington, N. and Hastie, R. (1991). A cognitive theory of juror decision making: the story model. *Cardozo Law Review*, 13:519–557.
- Pennington, N. and Hastie, R. (1993). Reasoning in explanation-based decision making. *Cognition*, 49:123–163.
- Picinali, F. (2013). Two meanings of “reasonableness”: Dispelling the “floating” reasonable doubt. *Modern Law Review*, 76(5):845–875.
- Picinali, F. (2016). Base-rates of negative traits: Instructions for use in criminal trials. *Journal of Applied Philosophy*, 33(1):69–87.
- Poisson, S. D. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*.
- Posner, R. (1973). *The Economic Analysis of Law*. Brown & Company.
- Prakken, H. and Sartor, G. (2009). A logical analysis of burdens of proof. In Kaptein, H., Prakken, H., and Verheij, B., editors, *Legal Evidence and Proof: Statistics, Stories, Logic*, chapter 9, pages 223–253. Ashgate.
- Redmayne, M. (2008). Exploring the proof paradoxes. *Legal Theory*, 14(4):281–309.
- Reichenbach, H. (1949). *The Theory of Probability (English edition)*. University of California Press.

- Renooy, S. (2001). Probability elicitation for belief networks: Issues to consider. *The Knowledge Engineering Review*, 16(3):255–269.
- Robertson, B., Vignaux, G., and Berger, C. (2016). *Interpreting evidence: evaluating forensic science in the courtroom*. John Wiley & Sons.
- Ross, L. (2020). Rehabilitating statistical evidence. *Philosophy and Phenomenological Research*. [online first].
- Roth, A. (2010). Safety in numbers? Deciding when DNA alone is enough to convict. *New York University Law Review*, 85(4):1130–1185.
- Royall, R. M. (1997). *Statistical Evidence: A likelihood paradigm*. Chapman and Hall/CRC.
- Saks, M. J. and Kidd, R. F. (1980). Human information processing and adjudication: Trial by heuristics. *Law and Society Review*, 15(1):123–160.
- Schauer, F. (2003). *Profiles, Probabilities, and Stereotypes*. Belknap Press.
- Schmalbeck, R. (1986). The trouble with statistical evidence. *Law and Contemporary Problems*, 49(3):221–236.
- Schoeman, F. (1987). Statistical vs. direct evidence. *Noûs*, 21(2):179–198.
- Schwartz, D. S. and Sober, E. (2019). What is relative plausibility? *The International Journal of Evidence & Proof*, 23(1-2):198–204.
- Schwartz, D. S. and Sober, E. R. (2017). The Conjunction Problem and the Logic of Jury Findings. *William & Mary Law Review*, 59(2):619–692. DOI: 10.2139/ssrn.2927252.
- Schweizer, M. (2013). The Law Doesn't Say Much About Base Rates. *SSRN Electronic Journal*.
- Scutari, M. and Denis, J.-B. (2015). *Bayesian Networks in R*. CRC Press.
- Sesardic, N. (2007). Sudden infant death or murder? A Royal confusion about probabilities. *British Journal for the Philosophy of Science*, 58:199–329.
- Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press.
- Shen, Q., Keppens, J., Aitken, C., Schafer, B., and Lee, M. (2007). A scenario-driven decision support system for serious crime investigation. *Law, Probability and Risk*, 5(2):87–117.
- Simons, D. J. and Chabris, C. F. (1999). Gorillas in our minds: Sustained inattention blindness for dynamic events. *Perception*, 28:1059–1074.
- Skyrms, B. (1968). Choice and chance: An introduction to inductive logic.
- Skyrms, B. (1980). *Causal Necessity*. Yale University Press.
- Smith, M. (2017). When does evidence suffice for conviction? *Mind*.
- Stein, A. (2005). *Foundations of Evidence Law*. Oxford University Press.
- Stein, A. (2008). The right to silence helps the innocent: A response to critics. *Cardozo Law Review*, 30(3):1115–1140.
- Sullivan, S. P. (2019). A Likelihood Story: The Theory of Legal Fact-Finding. *University of Colorado Law Review*, 90(1):1–66. DOI: 10.2139/ssrn.2837155.
- Suzuki, J. (2015). *Constitutional Calculus: The Math of Justice and the Myth of Common Sense*. Johns Hopkins University Press.
- Sykes, D. L. and Johnson, J. T. (1999). Probabilistic evidence versus the representation of an event: The curious case of Mrs. Prob's dog. *Basic and Applied Social Psychology*, 21(3):199–212.
- Taroni, F., Biedermann, A., Bozza, S., Garbolino, P., and Aitken, C. (2014). *Bayesian networks for probabilistic inference and decision analysis in forensic science*. John Wiley & Sons, 2nd edition.
- Thompson, W. C. and Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior*, 11(3):167–187.
- Thomson, J. J. (1986). Liability and individualized evidence. *Law and Contemporary Problems*, 49(3):199–219.
- Tillers, P. (1997). United States v. Shonubi: A statistical oddity? *Cardozo Law Review*.
- Tillers, P. (2005). If wishes were horses: discursive comments on attempts to prevent indi-

- viduals from being unfairly burdened by their reference classes. *Law, Probability and Risk*, 4(1-2):33–49.
- Treasure, J. B. (1992). Nigerian connection floods u.s. airport with asian heroin. *New York Times*, <https://www.nytimes.com/1992/02/15/nyregion/nigerian-connection-floods-us-airports-with-asian-heroin.html>.
- Tribe, L. H. (1971). Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review*, 84(6):1329–1393.
- Triggs, C. M. and Buckleton, J. S. (2004). Comment on: Why the effect of prior odds should accompany the likelihood ratio when reporting DNA evidence. *Law, Probability and Risk*, 3:73–82.
- Urbaniak, R. (2018). Narration in judiciary fact-finding: a probabilistic explication. *Artificial Intelligence and Law*, pages 1–32. [DOI: 10.1007/s10506-018-9219-z].
- Urbaniak, R. (2019). Probabilistic legal decision standards still fail. *Journal of Applied Logics*, 6(5).
- Urbaniak, R. and Janda, P. (2019). Probabilistic models of legal corroboration. *The International Journal of Evidence & Proof*, pages 1–23. [DOI 10.1177/1365712719864608].
- Urbaniak, R., Kowalewska, A., Janda, P., and Dziurosz-Serafinowicz, P. (2020). Decision-theoretic and risk-based approaches to naked statistical evidence: Some consequences and challenges. *Law, Probability and Risk*. [DOI: 10.1093/lpr/mgaa001].
- van der Gaag, L. C., Renooij, S., Witteman, C. L. M., Aleman, B. M. P., and Taal, B. G. (2013). How to elicit many probabilities.
- Venn, J. (1866). *The logic of chance: An essay on the foundations and province of the theory of probability, with especial reference to its application to moral and social science*. Macmillan.
- Verheij, B. (2017). Proof with and without probabilities. correct evidential reasoning with presumptive arguments, coherent hypotheses and degrees of uncertainty. *Artificial Intelligence and Law*, pages 1–28.
- Vlek, C., Prakken, H., Renooij, S., and Verheij, B. (2014). Building bayesian networks for legal evidence with narratives: a case study evaluation. *Artificial Intelligence and Law*, 22:375–421.
- Volokh, A. (1997). N guilty men. *University of Pennsylvania Law Review*, 146(2):173–216.
- Walen, A. (2015). Proof beyond a reasonable doubt: A balanced retributive account. *Louisiana Law Review*, 76(2):355–446.
- Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. Chapman and Hall London.
- Wasserman, D. T. (2008). Forensic DNA typing. In Burley, J. and Harris, J., editors, *Companion to Genetics*. Blackwell.
- Wells, G. (1992). Naked statistical evidence of liability: Is subjective probability enough? *Journal of Personality and Social Psychology*, 62(5):739–752.
- Williams, G. (1979). The mathematics of proof (parts I and II). *Criminal Law Review*, pages 297–312, 340–354.
- Wixted, J. T. and Wells, G. L. (2017). The Relationship Between Eyewitness Confidence and Identification Accuracy: A New Synthesis. *Psychological Science in the Public Interest*, 18(1):10–65.
- Wren, C. S. (1999). A pipeline of the poor feeds the flow of heroin. *New York Times*, <https://www.nytimes.com/1999/02/21/nyregion/pipeline-poor-feeds-flow-heroin-traffickers-field-more-swallowers-evade.html>.
- Wright, R. W. (1988). Causation, responsibility, risk, probability, naked statistics, and proof: Pruning the bramble bush by clarifying the concepts. *Iowa Law Review*, 73:1001–1077.
- Zabell, S. L. (2005). Fingerprint evidence. *Journal of Law and Policy*, 13(1):143–179.

Academic Tools

[Auto-inserted by SEP staff]

Other Internet Resources

- Put URL for website 1 here
- Put URL for website 2 here

Related Entries